ACADEMY OF SCIENCES OF THE ESTONIAN S.S.R.

INSTITUTE OF LANGUAGE AND LITERATURE

CONTENTS

INDEX OF AUTHORS

AUDITORY MODELS FOR SPEECH PROCESSING

Matti Karjalainen

Helsinki University of Technology
Acoustics Laboratory, Otakaari 5 A
SF-02150 Espoo FINLAND

ABSTRACT

Computational modeling of the auditory periphery has become an integral part of hearing and speech research in recent years. This reflects the importance of computers and computational models as a research tool for experimenting flexibly in the domain of complex auditory phenomena. Both our general understanding and the fragmental knowledge of details known from hearing research can be reconstructed and tested in the form of functional models.

This paper approaches the auditory models primarily from another aspect: their applications within speech processing. Although there are almost no existing practical applications where systematic modeling has proven to be superior to traditional methods, the approach as such is seen as promising and necessary. Several approaches to auditory modeling are viewed in the paper with the main emphasis on functional and psychoacoustical properties, including some principles proposed for higher-level processing. Potential areas of applications are discussed with examples taken from our own studies.

INTRODUCTION

The theories, models and applications of speech perception are without any doubt lagging behind the level of knowledge in speech production. The main reasons for this are due to the complexity of the hearing system and the difficulties in experimenting with it, the lack of basic understanding of the higher-level processes and the problems in the implementation of experimental models to simulate the auditory system.

What is (or could be) auditory modeling?

The development of computers and software-based simulation makes it more and more attractive to experiment with principles of hearing. To some extent electronic and even mechanical models have been tried but the computer has become a superior tool for the task. The concept of *auditory model* is used normally to refer to a computational model of the peripheral hearing system. The *physiological* functions of the basilar membrane and other cochlear processes up to the neural levels are considered as the primary subject to be simulated by the models.

Another theoretical and experimental basis for auditory modeling comes from *psychoacoustics*, where the correspondence to the underlying physiology is not direct anymore. Perception thresholds and psychophysical "transfer functions" are more central to the approach. Psychoacoustical concepts like pitch and loudness that are related to the peripheral hearing are well developed and exact to a high degree. They have been verified by subjective listening experiments. More abstract properties exhibit fuzziness and random behaviour but can be included in computational models if they are stable enough.

The third approach to modeling is to hypothesize functional principles that possibly could be found in the hearing system. They may not be verified by direct physiological or psychological experiments. Most auditory models concerning higher levels of hearing will probably be of this type because the physiological basis is too complex and hard to access, and even the psychological approach does not test and validate the models. The borderline between auditory modeling and general information processing principles is not very clear at these levels.

Why auditory models?

Auditory modeling is attractive as a research tool because it presents the possibility to test hypotheses and experiment with new ideas flexibly in a proper context. The hearing system consists of complex subsystems that tend to be nonlinear and contain feedback loops, which makes it practically impossible to apply analytical modeling methods except to small subproblems. Computational models are useful also in conceptualizing the signal and information processing aspects in hearing apart from the underlying physiology.

The basic research of hearing is only one of the motivations for auditory modeling. Major challenges for future work are to be found in potential applications, especially in speech recognition. The human hearing system is the best processor to recognize speech messages; why not to try to duplicate it in technical form. The results so far show that this will not be done easily. In principle, however, this approach is promising and necessary, at least to gain a deeper insight into the many problems of speech recognition.

This paper reflects the point of view of the author towards auditory modeling. Physiological models are not seen as the only, and even not the major subject of research, when it comes to applications. Especially for speech processing we need flexible functional models based on signal processing and artificial intelligence. The rest of the paper will tie together a number of subproblems in auditory modeling along with some applications and experiments performed by our own research group.

MODELS OF THE PERIPHERAL HEARING SYSTEM

External and Middle Ear Models

Computational modeling of the hearing system begins from the acoustics of the external ear. Localization of sound and the frequency sensitivity of the ear are greatly influenced by the acoustical details of the pinna and ear canals. The experimental studies and measurements [1] that have been made have led to successful results in computationally reconstructing an authentic sound environment sensation [2]. In combination with cochlear and neural modeling this could lead to better directional selectivity and sound localization [3] e.g. in speech recognition devices. Otherwise the role of the external and middle ear is a relative simple, almost linear filter as a part of a complex auditory model, contributing to the frequency sensitivity properties.

Cochlear Modeling

The physiology of the inner ear [4] - [8] is a main source of knowledge providing a concrete basis for present auditory modeling. This area of research is fairly rich in results and approaches, see [9] - [14], but no comprehensive and systematic cochlear models exist in computational form.

Modeling of the inner ear can be divided into several subproblems. The mechanics of the basilar membrane has received considerable attention since the studies of von Békésy [7]. The mathematically elegant principle of the nonhomogeneous transmission line must be cnhanced with nonlinear processes and complex interactions with hair cells and neural processes [15] - [17]. Some recent results propose interesting computable models of interaction to improve the sensitivity and selectivity of the inner ear, Zwicker [18] and Lumer [19]. The acoustic emission, see Kemp [20], [21], should also be included into a full-scale model. The mechanical to electrical and neural transduction takes place in the hair cells and can be modeled in physiological detail or functionally including the random nature of single cell firings, e.g. Schröder and Hall [22], Lyon [23].

Towards Higher Levels

As the computational modeling domain moves towards the neural levels, more functional principles must be used instead of physiological facts. Mixing of physiology, psychology and highly hypothetical ideas in the form of computer programs is an important approach. Some models are more oriented towards the study of advanced computational implementations and applications than the hearing process as such, e.g. Lyon [23] - [25]. Reaching higher abstraction levels in relation to physiology by computer programs may prove to be valuable when studying the representation of speech and complex stimuli in the hearing system.

The problem of neural representations of speech signals has become a subject of remarkable research in recent years, [10], [26] - [28]. There are different explanations of how the spectral and temporal information is coded into the neural signal. The saturation effect of a single nerve fiber in sending amplitude data must be taken into account. The computational models try to capture the essentials of this process in different ways: e.g. the synchrony model proposed by Seneff [29], [30] is based on the firing synchronism principle found in the auditory nerve to avoid the spectral structure from being flattened. Seneff also made a generalization of the principle so that it can be applied to pitch detection.

Psychoacoustical Models

Some existing computational models find their theoretical and experimental basis primarily in psychoacoustics. The concepts of Bark scale (critical band scale), loudness and loudness density spectrum, masking curves, temporal time constants etc. [31] cannot be entirely reduced to the physiology of hearing. Examples of auditory models that are closely related to psychoacoustics are given e.g. by Schröder et al. [32] and Zwicker [33]. Both of them were developed with technical applications in mind.

Including Phonetic Aspects

Some research groups have worked by experimenting and modeling the perception of speech and its phonetically relevant features. Peripheral models of hearing tend to be nonspecific in relation to speech. How should the formants and formant transitions be processed by auditory models, and how should the phonetic features and categories be reflected in them? These problems are important from the point of view of applications, especially speech recognition.

Carlson, Granström et al. have discussed these questions and proposed several models for auditory speech analysis [34] - [36]. Klatt has a similar approach and he suggests a phonetic distance measure for comparison and classification of phonemes [37] - [39]. Principles and models relating auditory concepts to higher-level perception of speech are studied by Chistovich et al., [40] - [42]. Among them is the concept of center of gravity.

Auditory Modeling and Traditional Speech Processing

Many technically oriented systems for speech processing contain features that model the human hearing to some extent but some widely used methods do not exploit auditory features at all. It has been shown that linear predictive coding (LPC) in the original form is not optimal because it is based on a linearly weighted frequency scale. With Bark and loudness scaling its performance could in principle be considerably better [43] - [45]. Hermansky et al. have presented novel modifications of LPC analysis to include many important auditory features that can be applied to speech recognition [45] - [46].

AUDITORY SPECTRUM COMPUTATION

Most auditory models analyze audio signals by returning something we could call an *auditory spectrum.* This is natural because the inner ear (basilar membrane, hair cells) also forms some kind of a spectrum analyzer, even if it is different from technical devices and algorithms for the Fourier transform.

The models for auditory spectrum analysis can be divided into two classes according to the processing of temporal dynamics. If we are not especially interested in the detailed time constants of the resulting (short-time) spectrum representation, we can first apply the Fourier transform and then warp the frequency scale to the Bark scale. Otherwise, we need a transmission-line or filter-bank type analyzer to allow more freedom in the design of temporal features.

Auditory Spectrum by Fourier Transform

The human auditory system may be seen as a spectrum analyzer that differs from Fourier analyzers in many ways. The most important differences are:

1. spectral emphasis by the inverse of the equal loudness curves,
2. use of the Bark scale (critical band scale) instead of the Hertz frequency scale,
3. frequency domain resolution of about one Bark,
4. masking effect in the frequency domain and spreading of the spectral components, and
5. time domain dynamics: temporal integration and masking effect in the time domain (forward and backward masking).

All these properties are known from psychoacoustics [31] but there have not been very many attempts to apply them in practical applications. Schröder & al. have used a computational model when evaluating signal-to-noise ratios in speech transmission [32]. We adopted their mathematical formulation with minor modifications as follows:

* Computation of the Fourier transform with a 35 ms Hamming window.

* Emphasis of the spectrum by an approximation of the frequency sensitivity curve of the ear (inverse of the equal loudness curve).

* Transformation of frequency **f** fo Bark variable **x** by:

$$\mathbf{x} = 7\,\mathrm{arsinh}(\mathbf{f}/650\mathrm{Hz})\,.$$

* So called "excitation function" **E(x)** is found by smoothing the Bark-scaled pre-emphasized power spectrum **S(x)** with a "spreading function" **B(x)**:

$$\mathbf{E}(\mathbf{x}) = \mathbf{S}(\mathbf{x}) * \mathbf{B}(\mathbf{x}), \qquad \text{(* indicates convolution)}$$

where **B(x)** in our model is a piecewise linear approximation of the Schröder et al. spreading function

$$10\log(\mathbf{B}(\mathbf{x})) = 15.81+7.5(\mathbf{x}+0.474) - 17.5\sqrt{1+(\mathbf{x}+0.474)^2}$$

by linear slopes (+ 25 dB/Bark, -10 dB/Bark) and power series approximation for the top of the curve (see Fig. 1b).

* dB-scaled **E(x)** is the final auditory spectrum used in the study. Two examples of such spectra of simple signals are shown in Fig.1. The spectrum of an impulse (1a) has a form which is similar to the frequency sensitivity of the ear. The auditory spectrum of a sine wave (1b) gives the masking curve and an approximate form of the spreading function **B(x)**.

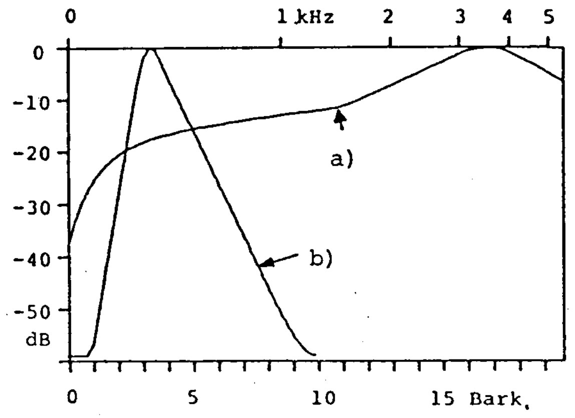


Fig. 1. Auditory spectra of simple test signals: (a) impulse spectrum and (b) sine wave spectrum.

Some examples of auditory spectra with corresponding Fourier spectra for speech sounds are plotted in Figures 2 and 3. The Finnish vowel /a/ shows clearly how the harmonic structure in a Fourier spectrum is smoothed out but the main formants are retained in an auditory spectrum (Fig. 2). In the fricative /s/ the random variation of the Fourier spectrum is also smoothed and the "fricative formant" shows up in the form of a normal vowel formant (Fig. 3).

Auditory Spectrum by Filter-bank Modeling

It was found to be difficult to include proper temporal dynamics when using the Fourier transform techniques. The filter-bank principle is well suited to auditory spectrum analysis because the human auditory system - basilar membrane and hair cells - also consists of a multi-channel analyzer. The bandwidth of the overlapping channels is about one critical band or 1 Bark. Instead of thousands of hair cells in the biological system it is enough to have 1 - 4 channels per one Bark in a computational model. This means 24 - 96 channels covering the 24 Bark audio range. With 0.5 Bark spacing our model has 48 channels, which seems to be a practical compromise between good resolution of spectral representation and a low amount of computation.

Each channel consists of a bandpass filter, a square-law rectifier, a fast linear and a slower nonlinear lowpass filter, and a dB-scaling stage (Fig.4).

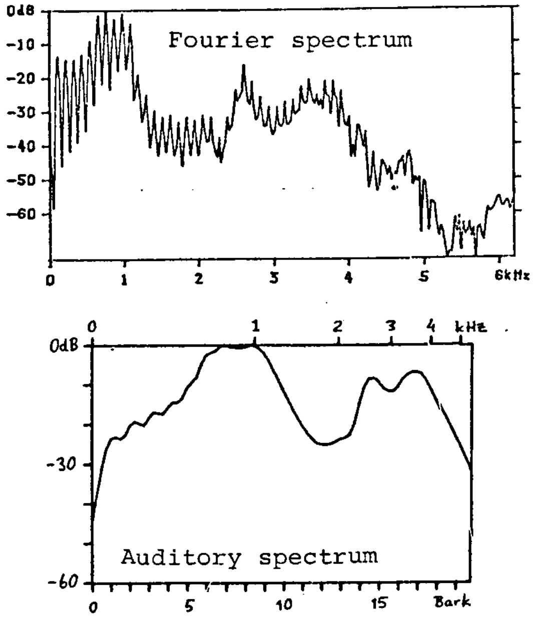


Fig. 2. Fourier spectrum and auditory spectrum for a Finnish vowel /a/.

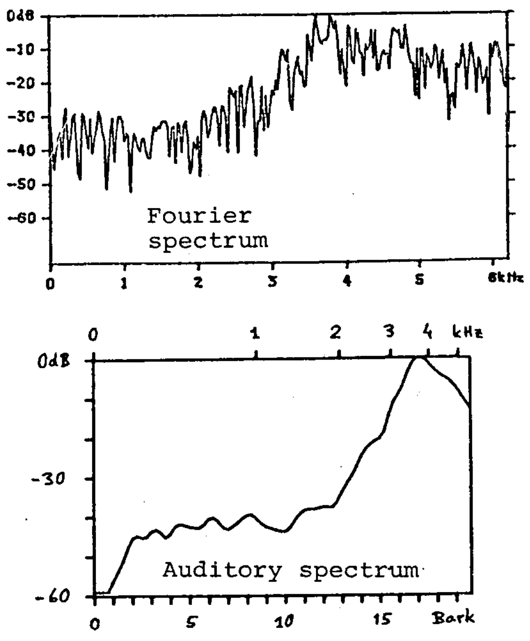


Fig. 3. Fourier spectrum and auditory spectrum for a Finnish fricative /s/ (in context /assa/).

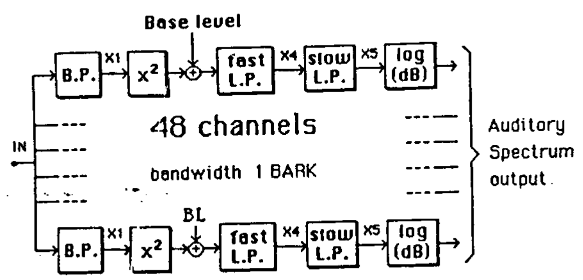


Fig. 4. A 48-channel filter-bank model for auditory spectrum computation. B.P.=bandpass, L.P.=lowpass filter, x^2= square-law detection, LOG = dB-scaling.

Bandpass filter bank

Bandpass filters with 0.5 Bark spacing and about 1.3 Bark bandwidth give the desired frequency selectivity to the model. Each bandpass is a 256th order FIR-filter, carefully designed to have a frequency response which is the mirror image of the spreading function B(x) given by Schröder et al.

This filter bank design gives a good approximation of the desired masking properties in the frequency domain. Computation of the filter bank was implemented as a matrix multiplication in an array processor (Floating Point Systems FPS 100). Even an array processor could not run it in real time. By a proper IIR-filter design the speed of the computation could be more than 10 times faster but accurate design of these filters is a difficult task.

Not only frequency selectivity but also the frequency response (sensitivity) of the ear must be built into the filter bank. The simple way we used is to let the relative gains of the channels be proportional to the inverse of the equal loudness curve (60 dB-level).

Rectification

The rectification effect in the hair cells of the inner ear is primarily of the half-wave type. Since a square-law element was needed for temporal integration in our model, we ended up using it without any half-wave rectifier. We found that in the auditory spectrum analysis of speech this makes no noticeable difference. A constant level is added after the rectification to simulate the threshold of hearing.

Filters of temporal integration and forward masking

The remaining two filters are for smoothing the outputs of the rectified bandpass filters. The faster one is a first-order lowpass with a time constant of about 3 ms. The second one is more important. Its purpose is to implement many effects: temporal integration and pre- and postmasking effects.

Temporal integration is realized by linear first-order lowpass filtering (time constant of about 100 ms) applied to the output of the square-law rectifier. Premasking is not a very important and critical phenomenon. No additional modeling was necessary to match it well enough.

Postmasking was found to be more difficult to implement in sufficient detail. A linear lowpass filter with a 100 ms time constant yielded postmasking effect that was many times too long. We used a nonlinear (logarithmically linear) behaviour of the filter for masking conditions (X4 < X5). The form of the temporal masking pattern is now close to the actual one found in psychoacoustical studies [31] but a delay of about 10 ms present in the real masking effect is lacking in the model. The overall response of the slow nonlinear lowpass can be stated now:

$$X5(n) = K1*X4(n) + (1-K1)*X5(n-1), \qquad \text{if } X4 \geq X5,$$
$$X5(n) = X5(n-1)*\exp(K2*\log(X4(n)/X5(n-1))), \quad \text{if } X4 < X5,$$

where X4 and X5 are the input and output of the filter, K1 and K2 the filter coefficients, and n the discrete time variable. A good value for K1 was found to be 0.0005, and 0.0007 for K2 when the sampling frequency is 20 kHz.

Auditory short-time spectra computed by the model can be displayed in many forms: spectral series, spectrograms, etc. Examples of these are given later in this paper.

AUDITORY FORMANTS AND FORMANT SPECTRA

The auditory spectrum, as was analyzed by the models above, is not a speech-specific representation. Attempts to utilize it or other similar preprocessing methods in speech recognition have shown only moderate results, see e.g. [47]. It is obvious that some further processing of auditory spectra is needed to exhibit speech-specific features and more "phonetic-like" auditory representations of speech.

Some hints and guidelines can be found e.g. from the studies of Klatt [37] - [39], paying special attention to the formant peak regions in the auditory spectra. Global properties such as the slope of the spectrum have only a minor effect on the phonetic quality of a sound. Klatt suggested the use of *phonetic distance measures* [39] based on local properties of the formants in auditory spectra. Another concept that is closely related to auditory formants is the *center of gravity* by Chistovich et al. [40].

Emphasizing and Sharpening the Auditory Formants

Possible conclusions that may be drawn from the results of using short-time auditory spectra in speech recognition could be that the perceptually important formant peaks are excessively smoothed and the local properties of the formants are not prominent enough. Is it possible to compensate for these effects? There are neurophysiological principles that are candidates for the spectral sharpening effect: *lateral inhibition* is one such candidate. A strong excitation at a certain place along the basilar membrane tends to suppress the neighbouring channels.

The formant features can be sharpened or emphasized computationally in many ways. We can perform highpass or bandpass filtering of the auditory spectrum in the Bark domain to supress the global forms (e.g. spectral tilting) and to emphasize the local formant peaks. This can be realized by convolving the loudness-scaled auditory spectrum by a proper spatial (Bark domain) bandpass filter impulse response. Figures 5 and 6 show original auditory spectra for a vowel /ä/ and fricative /s/ along with the resulting *auditory formant spectra*, as we call them.

In both cases the auditory formant spectrum exhibits clearly the formant peaks so that the global spectrum structure does not have a major dominance. Serial displays of auditory spectrum and auditory formant spectrum are shown in Fig. 7 for the vowel combinations /aiai/.

The Concept of Auditory Formant

The perceptual relevance of the peaks in auditory spectra implies the usefulness of the concept *auditory formant*. It must be recognized as different from the acoustic and articulatory aspects of formants even if there is a clear correspondence between them. A useful characterization of the auditory formant is to state it as any *peak or relatively localized high-loudness region*, a kind of landmark in an auditory spectrum.

Several studies have been done on the perceptual behavior of auditory formants and formant groups, see e.g. Chistovich [40]. The integration of closely spaced formants, the concept of center of gravity, etc. are principles that should also be implemented in computational models.

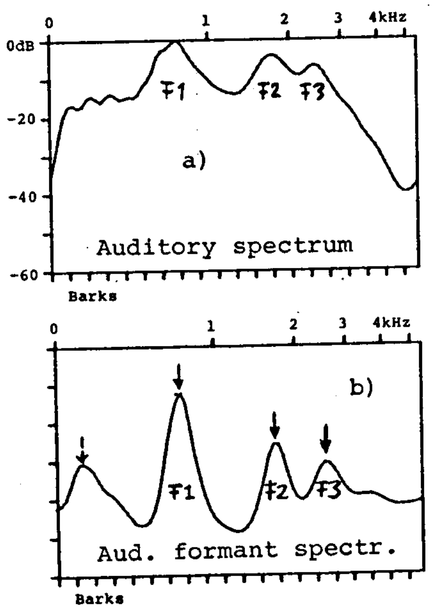


Fig. 5. Finnish vowel /ä/: (a) auditory spectrum, (b) derived auditory formant spectrum

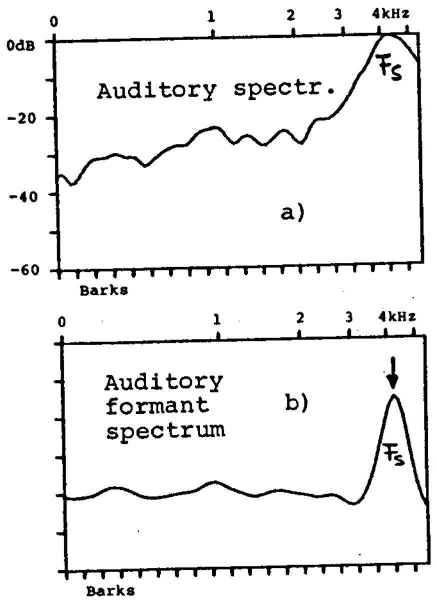


Fig. 6. Fricative /s/: (a) auditory spectrum, (b) auditory formant spectrum.

Is it possible to extract auditory formants and to describe them as discrete units? The auditory formant spectrum above is a good data source for this extraction. In Figures 5, 6 and 7 a spatial bandpass filtering with a one Bark resolution was applied to give a proper pre-emphasis to the spectrum. A peak-picking algorithm easily finds the formants as the local maxima of the curves. Fig. 8 illustrates how the short-time auditory spectrum of the utterance /kuusi/ (Fig. 8(a) dB-scaled, Fig. 8(b) loudness-scaled) is transformed to an auditory formant spectrum, Fig. 8(c). The formant peaks are finally plotted and shown in a spectrogram-like display (Fig. 8(e)).

The one Bark resolution does not always work. Closely spaced formants may give a better respońse e.g. to a 2 Bark resolution filtering, Fig. 8(d). (See also the 2 Bark formant spectrogram in Fig. 8(f)). This finding shows the need for different resolutions in different contexts. According to Chistovich the auditory system can integrate neighbouring formants up to a distance of 3.5 Barks [40]. This corresponds to about a 2-3 Bark resolution in our bandpass filtering.

Since there is no single optimal resolution a better strategy is to use multiple resolutions in parallel. This means that the formant peaks are picked to form several formant lists. Later on it is possible to utilize the data that seems to be the most reliable based on the context.

We can also visualize the multiple resolution auditory formant data in a spectrogram form by using different gray levels or colors for the formant trajectories of different resolutions. Fig. 9 shows the mixed result of 1 and 2 Bark auditory formant analyses for the word /kuusi/. The general principle of multiple resolution analysis is discussed below.

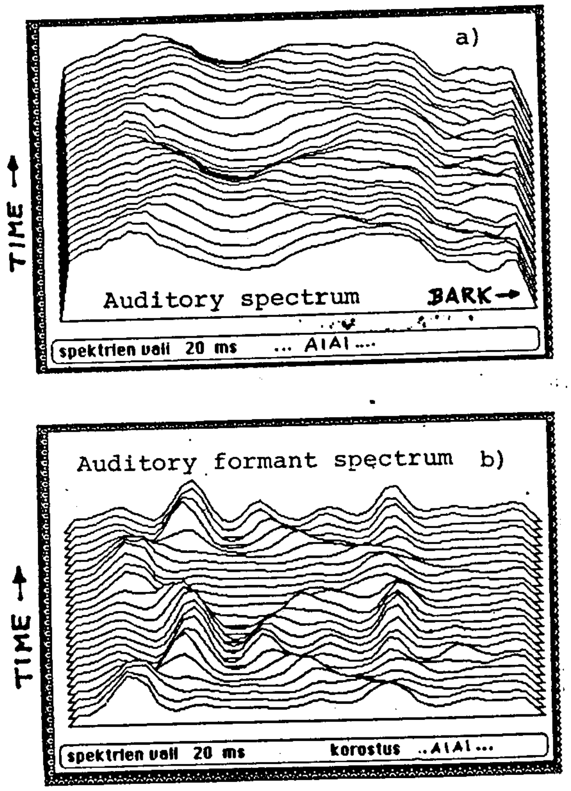


Fig. 7. Spectral series for the vowel combination /aiai/: (a) auditory spectrum, (b) auditory formant spectrum.

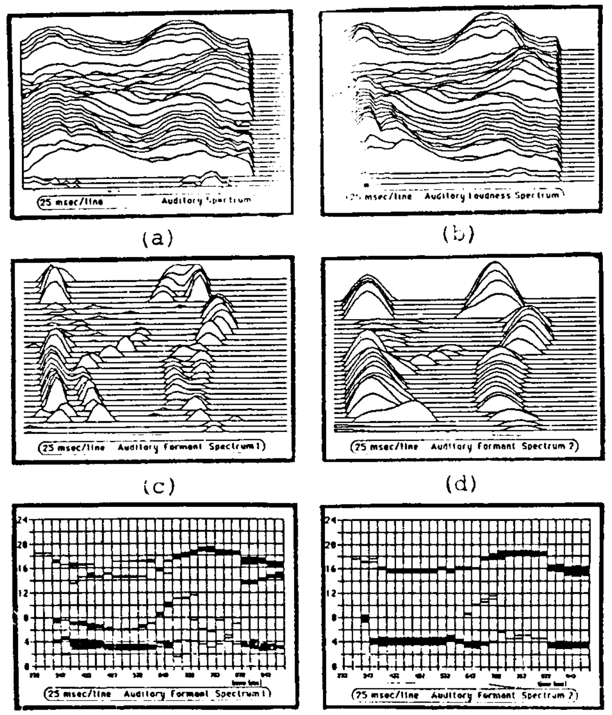


Fig. 8. Auditory spectrum presentations for the word /kuusi/: (a) original dB-scaled spectrum, (b) loudness-scaled spectrum, (c) auditory formant spectrum with 1 Bark resolution, (d) auditory formant spectrum with 2 Bark resolution, (e) formant spectrogram of 1 Bark resolution and (f) formant spectrogram of 2 Bark resolution.

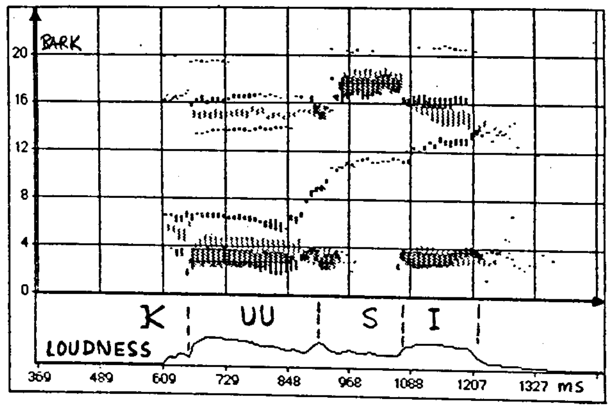


Fig. 9. An auditory formant spectrogram with two overlayed displays for resolutions of 1 Bark (black) and 2 Barks (gray).

Local vs. Global Features of Auditory Spectra

Both the auditory formant spectrum and the detection of discrete formant parameters emphasize the prominent *local* features in auditory spectra. We could also analyze and characterize the *global* properties. An average spectral slope and the center of gravity over the whole audio range are good examples of such global attributes. From the point of view of speech perception the absolute values of these parameters are not as important as the relative changes they exhibit. For instance, large static spectral tilting is allowed with only minor change in the phonetic quality of a vowel [39].

Spectral tilt or center of gravity can also be computed over any limited range in the Bark domain. An interesting special case is to analyze closely the effective movement ranges of the lowest formants, e.g. over 2 to 6 Barks for F1. The values of these parameters describe the average slope or the approximate position of the formant in the defined range. Fig. 10 shows the results of such an analysis for the full audio range, F1, F2, and F3&F4 range, along with the loudness function and a two-resolution auditory formant spectrogram for the Finnish word /viisi/.

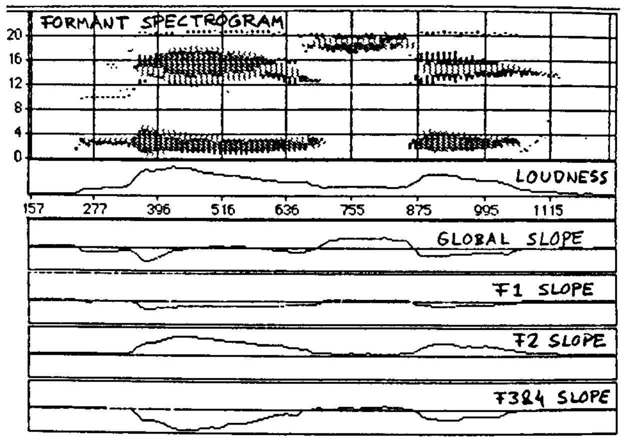


Fig. 10. Several different analyses for the Finnish word /viisi/: (a) formant spectrogram, (b) loudness function, (c) global slope and the local spectral slopes for the formant ranges F1, F2 and F3&F4.

Analysis of Formant Movements

It is known that the hearing system is especially sensitive to *changes* in sound. It is also known that the auditory system contains specialized analyzers for frequency sweeps and formant movements [48]. Such detectors may have an important role in the perception of speech signals and they should be included in computational auditory models. To some degree, the derivatives of the slope funtions above represent this kind of information. The output of an advanced detector could be a series of "formant movement events" similar to the time structure analysis method in the following section.

TIME STRUCTURES AND MULTIPLE RESOLUTION ANALYSIS

Time is one of the most difficult and least understood dimensions in speech signal analysis. The rhythm and timing in real speech varies widely according to the context and therefore straightforward methods of segmentation do not work reliably. The transformation from continuous-time representations to discrete units in time should be studied more carefully so that time resolution is seen as one parametric scale.

Let us consider a set of parametric or feature functions as analyzed from a speech signal. Fig. 11 shows the total loudness function (sum of all the filter-bank channels), nonstationarity (relative change in short-time auditory spectrum) and the global spectral slope as a function of time, along with the formant spectrogram for the utterance /yksi/. What is a flexible and reliable way to do "segmentation" based e.g. on the loudness function?

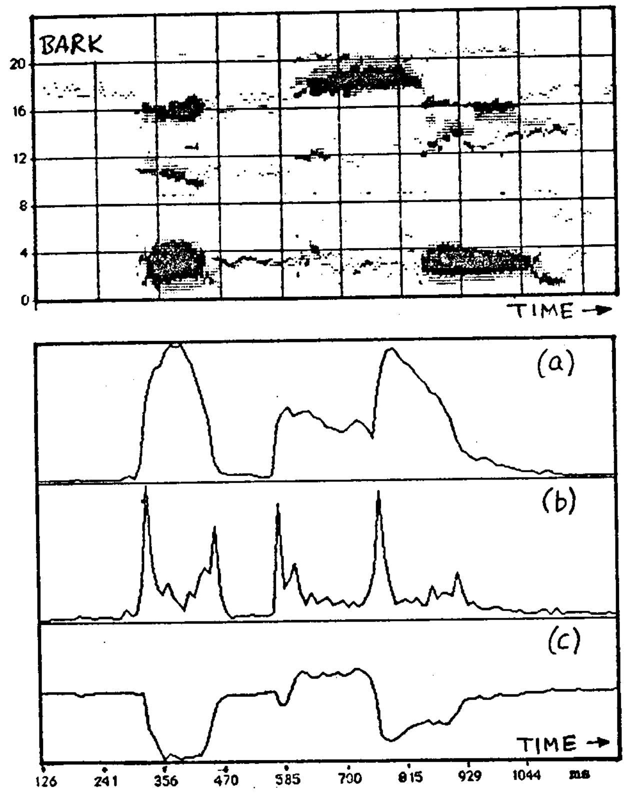


Fig. 11. Auditory formant spectrogram and multiple temporal feature functions for the Finnish word /yksi/: (a) total loudness function, (b) nonstationarity function and (c) global spectral slope.

If proper bandpass filtering is applied to the loudness function the "events" that match best to the impulse response of the filter are emphasized. The same principle is used as was for the filtering of formants in the frequency domain. An example of a useful impulse response for a *resolution filter* is shown in Fig. 12.

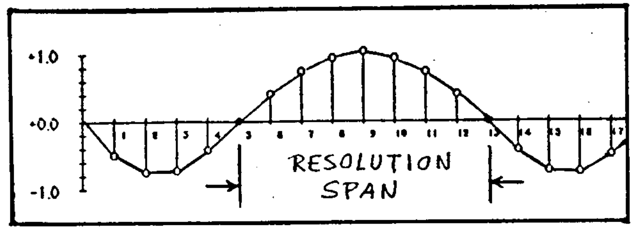


Fig. 12 An example of an impulse response for a resolution filter.

Any single filter emphasizes the events of its corresponding time resolution the most. The extrema (maxima and minima) of the response are easily picked up as prominent *events* in the time structure of the signal. To be more flexible *a set* of resolution filters can be applied to the loudness function in parallel. In this *multiple resolution analysis* each resolution filter channel produces a list of potential *loudness events*. Other parametric functions like the global spectral slope create their own event lists and list structures. The idea of multiple resolution analysis has some resemblance to the scale-space filtering proposed by Witkin [49].

As an example of using the principle, nine filters with time resolutions ranging from 10 to 320 ms were applied to the loudness function of Fig. 11. The convolution results are plotted in Fig. 13. A continuous scale of resolutions is in principle the ideal case but a series of filters with resolution ratios of about 1:√2 was found to be practical. The method of multiple resolution analysis leads to an excessive amount of computation in comparison to single resolution (single window, frame, etc.). This is the cost to be paid for more flexibility. In highly parallel neural networks such computational redundancy is easily achieved but with present digital signal processing hardware it is a problem.

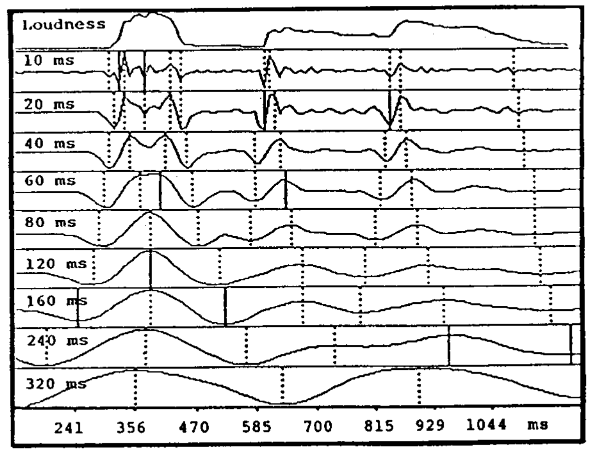


Fig. 13. Resulting curves from the multiple resolution analysis of the loudness function. Vertical lines indicate potential event positions.

Event-based Approach to Auditory Speech Analysis

Each feature to be used in a speech analysis system and each resolution of feature produces a corresponding list of events, containing much redundant information. In Fig. 13 e.g. the maxima of the neighbouring channels are closely interrelated. By a proper method we can discard many of the peaks as masked by more prominent neighbouring peaks. The potential events can be organized into the form of complex event list structures and processed further by rule-based and other artificial intelligence methods. This approach is discussed in more detail by Altosaar and Karjalainen in [50].

The event-based approach may be useful at several levels of auditory modeling. We could apply it at the auditory nerve level by picking up the most prominent peaks from the multiple resolution filtering of a single critical band channel in the model of Fig. 4, point X4. Here the range of interesting time resolutions is within a typical pitch period of speech. At the output level of the model (point X5) the resolution range corresponds to typi- cal speech segments of 10 to 300 ms. The prosodic features reveal still longer event objects. By parallel processing and concurrent programming techniquesa realization of this approach could be undertaken.

TEMPORAL FINE STRUCTURE ANALYSIS OF SPEECH SIGNALS

It is still a common belief that the temporal fine structure of speech signals within a pitch period (below 10 ms span) and the phase properties are irrelevant to speech perception. This belief is mostly due to the interpretations of the studies by Helmholtz [6]. Even if this is true as the first approximation several findings suggest a more important role for these details. For example, the success of the so-called multi-pulse LPC [51] in comparison to the simple impulse source and all-pole modeling shows how much the detailed time structure may affect the sound quality. In our Finnish speech synthesis studies we found that a careful zero-phasing (i.e. setting the phase of all harmonics to zero) of a natural utterance /illi/ changed it to be heard sometimes rather like /inni/.

Concepts Related to Temporal Fine Structure

The auditory spectrum output in our filter-bank type auditory model (Fig. 4) does not represent the temporal fine structure of speech signals at all. It is therefore possible to obtain identical auditory spectra for a voiced and an unvoiced sound from this model. The *degree of voicing* and *pitch for voiced sounds* are certainly concepts with close relationship to the temporal fine structure. This information should be analyzed from the fast response of the auditory filter bank, i.e. the point X4 in Fig.4, corresponding to the first neural stages of the hearing system.

Lyon [23] has presented computational auditory models to analyze the periodicity properties of speech signals by following the principles proposed by Licklider [52]. The models rely on correlation and coincidence functions from neural firings.

The *phase properties* of speech signals are difficult both to analyze and to interpret meaningfully. A traditional way of looking at phase has been by the Fourier transform of the signal. The phase in this sense is, however, very sensitive to noise, reverberation and other disturbances in speech. If any concept of *auditory phase* can be formed, it must be defined in a totally different way.

A step towards auditory phase could be to interpret it from the point of view of the modulation envelope in different critical band channels (Fig. 4). The fast response corresponding to the output of the hair cells in the hearing system carry this information. In the case of voiced speech sound these outputs exhibit the fundamental frequency of the speech. The relative phase shift of these pitch modulations between the neighbouring channels could be useful as the auditory phase function. The same data could be expressed also in the form of auditory group delay.

Sound Separation

The role of auditory analysis of temporal fine structure takes on a new appearance when we set the goal of modeling to be the phenomenon of *sound separation*. People can easily follow a single speaker in a high noise environment (e.g. the cocktail party effect). To make machines recognize speech at or below the level of background noise, a successful modeling of sound separation is needed.

This is a fairly new subject for serious computational modeling. Weintraub [53] has made remarkable contributions by studying some peripheral processes of auditory analysis in sound separation. He has emphasized the need for a multilevel strategy to solve this problem.

Some special cases of sound separation can be experimented with easily. We used the fast response outputs of our filter-bank model (Fig. 4) and computed the cepstrum for each of them. If there was a voiced speech sound that dominated a Bark channel, the corresponding pitch period was easily found as a dominant peak in its cepstrum. By summing the cepstra of all the channels the pitch periods of the individual sounds in a mixed signal were possible to be separated. After this it is feasible to estimate the spectrum of each speech sound in those frequency areas where the signal is not totally masked by other sounds. Such methods tend to be computationally so excessively heavy that it prevents their use in any real-time applications now or in the near future.

APPLICATIONS OF AUDITORY MODELS

There are very few if any practical applications of computational auditory models. This is quite natural because the subject of research is complex and relatively new. Some preliminary results of using them e.g. in speech recognition have shown poor or at best only marginal results. This has been discouraging but it has not stopped either basic research or application-oriented work in the field. If the human hearing system performs as the best as a speech recognizer, why couldn't a good model of it be the best speech recognizer as well.

Speech recognition has been an explicit motivation in the development of several practically oriented auditory models [33], [36], [46], [54] and implicitly in many other cases. One of the earliest works was documented by Zagoruiko and Lebedjev [55]. Only recently has there been signs of obtaining better results than with traditional methodology, see e.g. Cohen [54]. It is premature to draw conclusions about the real status of auditory models in speech recognition. It seems to be evident that no fundamental problems can be solved without models covering many or even all essential levels from acoustic signals to liguistic processing. One area of preprocessing where auditory models could help is in high background noise conditions.

Speech analysis in phonetics and basic speech research can gain from applying computational models of hearing. For example, auditory spectra give a picture of the important spectral features in speech signals from the point of view of perception. Carlson and Granström [35] and Klatt [38] among others have discussed the development of an auditory spectrograph. Traditionally the articulatory and acoustic aspects have been more dominant in speech analysis because of better instrumentation and tools for experimental work. With modern signal processors, personal computers and new programming techniques it was possible to develop a speech research workstation called ISA [56] that utilizes many of the representations described in this paper.

Speech synthesis is a process where auditory models can not be used directly. In the development of speech synthesis, however, we have been succesful in applying them. The microphonemic method by Lukaszewicz and Karjalainen [58] showed how the auditory formant spectrogram exhibited just the information needed for extracting pitch period prototypes from real speech. The use of the Bark scale was essential.

In *speech coding* the auditory models could be applied in the analysis phase, as a design tool, and in performance analysis, see Schröder et al. [32]. It is shown that the compactness of LPC-analysis could be impoved if auditory features could be integrated into the coding [43] - [45]. Unfortunately the auditory analysis tends not to preserve the properties that are needed for easy resynthesis of speech.

Measurement of sound quality (especially nonlinear distortion) in speech transmission was studied by us to find a better correlation between subjective and objective measures, Karjalainen [58], [59] and Helle and Karjalainen [60]. We have shown that an auditory spectrum distance of 2 dB corresponds to the just noticeable level of nonlinear distortion in speech signals. A prototype of a microprocessor-based distortion measurement system was also developed.

Technical audiology and phoniatrics are other areas of potential applications. In the delopment of hearing aids and cochlear implants it is evident that some kind of auditory models will be used in the future. In phoniatrics the properties of pathological voice can be analyzed based not only on articulatory and acoustic measurents but also on advanced auditory models [56].

CONCLUSION

This paper has presented an overview of auditory modeling from the point of view of speech processing research and applications. Both physiological, psychoacoustical and higher-level functional models are needed to gain a deeper understanding of the underlying phenomena and to be able to apply this knowledge to speech technology. Auditory modeling is a difficult area of research where progress is not always rapid. It also takes time to transfer the results into practice. Within the last ten years the interdisciplinary studies of computational auditory modeling have shown trends to expand and grow. Without any doubt this tendency will continue as the computational capabilities of modeling rapidly develop.

ACKNOWLEDGEMENTS

The paper presents some results from a research project on the modeling of auditory and speech communication that was financed by the Academy of Finland. The Posts and Telecommunications of Finland supported the development of some of the applications. I am grateful to all my co-workers who have made a contribution to the results.

REFERENCES

[1] **Blauert J.**, Spatial Hearing. The MIT Press, Cambridge 1983.

[2] **Pösselt C., Schröter J., Opitz M., Divenyi P.L.**, Generation of Binaural Signals for Research and Home Entertainment. Proc. of ICA-86, pp. B1-6, Toronto 1986.

[3] **Lyon R.**, A Computational Model of Binaural Localization and Separation. Proc. IEEE ICASSP-83, Boston 1983.

[4] **Pickles J. O.**, An Introduction to the Physiology of Hearing. Academic Press, London, 1982.

[5] **Möller A.R.**, Auditory Physiology. Academic Press, New York 1983.

[6] **Helmholtz H.**, On the Sensations of Tone. Dover Publications 1954.

[7] **von Békésy G.**, Experiments in Hearing. McGraw-Hill, New York, 1960.

[8] **Zwislocki J.J.**, Five Decades of Research on Cochlear Mechanics. J. Acoust. Soc.of Am, 67 (1980), pp. 1679-1685.

[9] **Allen J.B.**, Cochlear Modelling. IEEE ASSP Magazine, Vol.2 number 1, Jan 1985.

[10] **Carlson R., Granström B.** (ed.), The Representation of Speech in the Peripheral Auditory System. Elsevier Biomedical Press, Amsterdam, 1982.

[11] **Schröder M.R.**, Models of Hearing. Proc. of IEEE 63 (1975), pp. 1332-1350.

[12] **Allen J.B.**, Cochlear Micromechanics -- A Physical Model of Transduction. J. Acoust. Soc. Am. 68 (1980), pp. 1660 - 1670.

[13] **Delgutte B.**, Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve. In ref. 10.

[14] **Dolmazon J.**, Representation of Speech-Like Sounds in the Peripheral Auditory System in the Light of a Model. In ref. 10.

[15] **Neely S.T., Kim D.O.**, An Active Cochlea Model Showing Sharp Tuning and High Sensitivity. Hearing Research 9 (1983), pp. 123-130.

[16] **Khanna S. H., Leonard D.G.B.**, Basilar Membrane Tuning in the Cat Cochlea. Science, vol. 215, pp. 305-306, Jan. 1982.

[17] **Davis H.**, An Active Process in Cochlear Mechanics. Hearing Research 9 (1983), pp. 79-90.

[18] **Zwicker E.**, Peripheral Preprocessing in Hearing and Psychoacoustics as Guidelines for Speech Recognition. Proc. of Montreal Symposium on Speech Recognition, Montreal, 1986, pp. 1-4.

[19] **Lumer G.**, Computer Model of Cochlear Preprocessing (Steady State Condition) I. Basics and Results for one Sinusoidal Input Signal. Acustica Vol. 62 (1987), pp. 282-290.

[20] **Kemp D.T.**, Stimulated Acoustic Emission from within the Human Auditory System. J. Acoust. Soc. Am. 64 (1978), pp. 1386 - 1391.

[21] **Kemp D.T., Anderson S.D.**, Proc. of the Internat. Symposium on Nonlinear and Active Mechanical Processes in the Cochlea. Hearing Science, 2, no. 3 and 4, 1980.

[22] **Schröder M.R., Hall J.L.**, Model for mechanical to Neural Transduction in the Auditory Receptor. J. Acoust. Soc. Am. 55 (1974), pp. 1055-1060.

[23] **Lyon R.**, Computational Models of Neural Auditory Processing. Proc. of IEEE ICASSP-84, Tampa 1984.

[24] **Lyon R.**, A Computational Model of Filtering, Detection and Compression in the Cochlea. Proc. of IEEE ICASSP-82, Paris 1982.

[25] **Lyon R.**, Experiments with a Computational Model of a Cochlea. Proc. IEEE ICASSP-86, Tokyo 1986.

[26] **Evans E.F.**, Representation of Complex Sounds at the Cochlear Nerve and Cochlear Nucleus Levels. In ref. 10, pp. 27-42.

[27] **Möller A.R.**, Neurophysiological Basis for Perception of Complex Sounds. In ref. 10, pp. 43-60.

[28] **Sachs M.B., Young E.D., Miller M.I.**, Encoding of Speech Features in the Auditory Nerve. In ref. 10, pp. 115-130.

[29] **Seneff S.**, Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model. Proc. IEEE ICASSP-84, pp. 36.2.1-4, San Diego 1984.

[30] **Seneff S.**, A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research. Proc. of IEEE ICASSP-86, pp. 37.8.1-4, Tokyo 1986.

[31] **Zwicker E., Feldtkeller R.**, Das Ohr als Nachrichtenempfänger. S. Hirzel Verlag, Stuttgart, 1967.

[32] **Schröder M.R., Atal B.S., Hall J.L.**, Objective Measures of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. In: Frontiers of Speech Communication Research, pp. 217-229, (ed. Lindblom & Öhman). Academic Press 1979.

[33] **Zwicker E, Terhardt E., Paulus E.**, Automatic Speech Recognition Using Psychoacoustic Models. J. Acoust. Soc. Am. 65 (1979), pp. 487-498.

[34] **Carlson R., Granström.**, Model Predictions of Vowel Dissimilarity. STL-QPSR 3-4/1979, pp. 84-104.

[35] **Carlson R., Granström B.**, Towards an Auditory Spectrograph. In ref. 10, pp. 109-114.

[36] **Blomberg M., Carlson R., Elenius K., Granström B.**, Auditory Models and Isolated Word

Recognition. STL-QPSRL 4/1983, 1-15.

[37] **Klatt D.H.**, Speech Perception: A Model of Acoustic - Phonetic Analysis and Lexical Access. J. Phonetics 7 (1979), pp. 279 - 312.

[38] **Klatt D.H.**, Speech Processing Strategies Based on Auditory Models. In ref. 10, pp. 181-196.

[39] **Klatt D.H.**, Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step. Proc. of IEEE ICASSP-82, pp. 1278-1281, Paris 1982.

[40] **Chistovich L.A, Sheikin R.L., Lublinskaja V.V.**, "Centres of gravity" and Spectral Peaks as the Determinants of Vowel Quality. In: Frontiers of Speech Comm. Research, pp. 143-157 (ed. Lindblom & Öhman). Academic Press 1979.

[41] **Chistowicz L.A., Lublinskaya V.V., Malinnikova E.A., Ogorodnikova E.A., Stoljarova E.I., Zhukov S.JA.**, Temporal Processing of Peripheral Auditory Patterns of Speech. In ref. 10, pp. 165-180.

[42] **Chistowicz L.A.**, Central Auditory Processing of Peripheral Vowel Spectra. J. Acoust. Soc. Am. 77 (3), March 1985, pp. 789-805.

[43] **Makhoul J., Cosell L.**, LPCW: an LPC vocoder with linear spectral warping. Proc. IEEE ICASSP-76.

[44] **Koljonen J., Karjalainen M.**, Use of Computational Psychoacoustical Models in Speech Processing: Coding and Objective Performance Evaluation. Proc. of IEEE ICASSP-84, San Diego 1984.

[45] **Hermansky H., Fujisaki H., Sato Y.**, Spectral Envelope Sampling and Interpolation in Linear Predictive Analysis of Speech. Proc. of IEEE ICASSP-84, San Diego 1984.

[46] **Hermansky H., Tsuga K., Makino S., Wakita H.**, Perceptually Based Processing in Automatic Speech Recognition. Proc. of IEEE ICASSP-86, pp. 37.5.1-4, Tokyo 1986.

[47] **Blomberg M., Carlson R., Elenius K., Granström R.**, Experiments with Auditory Models in Speech Recognition. In ref. 10, pp. 197-201.

[48] **Lacerda F., Moreira H.O.**, How Does the Periphera Auditory System Represent Formant Transitions? / Psychophysical Approach. In ref. 10.

[49] **Witkin A.P.**, Scale-Space Filtering: A New Approach t Multi-Scale Description. Proc. of IEEE ICASSP-84, pp 39A.1.1-4, San Diego 1984.

[50] **Altosaar T., Karjalainen M.**, An Event-Basec Approach to Auditory Modeling of Speech Perception. I this proceedings.

[51] **Atal B.S., Remde R.**, A New Model of LPC Excitation for Producing Natural-Sounding Speech at Lov Bit Rates. Proc. of IEEE ICASSP-82, Paris 1982.

[52] **Licklider J.C.R.**, A Duplex Theory of Pitch Perception. Experimentia, 7 (1951), pp. 128-133.

[53] **Weintraub M.**, A Theory and Computational Model of Auditory Monaural Sound Separation. Doctoral Thesis, Stanford University, August 1985.

[54] **Cohen J.R.**, Application of an Adaptive Auditory Model to Speech Recognition. In Proceedings of the Montreal Symposium on Speech Recognition. McGill University, Montreal 1986.

[55] **Zagoruiko N.G., Lebedjev V.G.**, Models of Speech Signal Analysis Taking into Account the Effect of Masking. Acoustica 31 (1975), pp. 346-348.

[56] ISA, Intelligent Speech Analyser. Instruction manual, Vocal Systems, Finland.

[57] **Lukaszewicz K., Karjalainen M.**, Microphonemics - High-Quality Speech Synthesis by Waveform Concatenation. In this proceedings.

[58] **Karjalainen M.**, Sound Quality Measurements of Audio Systems Based on Models of Auditory Perception. Proc. of IEEE ICASSP-84, San Diego 1984.

[59] **Karjalainen M.**, A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of IEEE ICASSP-85, Tampa 1985.

[60] **Helle S., Karjalainen M.**, Perception and Measurement of Distortion in Speech Signals - An Auditory Modelling Approach. In this proceedings.

INTEGRATION AND SEGREGATION IN SPEECH PERCEPTION

BRUNO H. REPP

Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695

In this paper I present an overview of some recent research on speech perception. To reduce my task to manageable size, I have chosen to focus on the topics of perceptual integration and segregation, which have guided, more or less explicitly, a considerable amount of speech perception research and theorizing in recent years. This will be a selective review, therefore, but I hope it will nevertheless convey some of the flavor of contemporary ideas and findings, even though that flavor will be tinged with my own favorite spices.

CONCEPTUAL FOUNDATIONS

Integration and segregation are hypothetical perceptual functions (or processes) that link physical structures in the world with mental structures in the brain. An integrative function maps multiple physical units (trivially, a single physical unit) onto a single mental unit, whereas a segregative function maps multiple physical units (sometimes, paradoxically, a single physical unit) onto different mental units. Though mutually exclusive for any particular physical structure at any given time, these two processes nevertheless cooperate in sorting a complex stream of sensory inputs into an orderly sequence of perceived objects and events.

These definitions seem rather straightforward, but they rest on four important assumptions: (1) The physical and mental worlds are not isomorphic. (2) There are objectively definable units in the physical world. (3) There are units in the mental world that are different from the physical units. (4) There are perceptual functions or processes that accomplish the mapping between the two types of units. I will briefly defend each of these assumptions; at the end of this presentation, I will consider the consequences of abandoning some or all of them.

The first assumption, that the mental world is not isomorphic with the physical world, reflects the facts that physical variables are filtered and transformed by sensory systems, that perception is a function not only of the current sensory input but also of the past history of the organism, and that there is often an element of choice in perception which permits alternative perceptual organizations for the same sensory input. Without this assumption, it would be difficult to say anything meaningful about perception, except that it happens.

The second assumption, concerning the existence of physical units, is necessary in order to be able to talk about perceptual integration. These units or dimensions are what is being integrated. Perceptual segregation, too, ordinarily implies that certain objective lines of division can be found in the sensory input. It is always possible to find a physical description that is more finely grained than our description of the perceptual end product. The fact that the machines we use to assess physical characteristics of speech are mere transducers (or, at best, model only peripheral auditory processes) generally assures a mismatch between physical and perceptual descriptions even when the grain size is comparable (and even though our visual perception is engaged in interpreting the machine outputs). Although there are different ways of characterizing the physical energy pattern, they are all equally valid for descriptive purposes. It is an empirical question whether or not perceivers are sensitive to any observed physical divisions, i.e., whether these divisions can serve as the basis for perceptual segregation or whether they are bridged by integrative processes. Research of this kind may enable us to find a physical description with a simpler mapping onto perceptual units.

The third assumption concerns the existence and nature of perceptual (mental) units. There is no theory of speech perception that does not assume mental units, usually the ones supplied by linguistic theory. The argument has been over the "perceptual reality" of syllables, phonemes, and features, and over their relative primacy in perceptual processing (see, e.g., [69, 83, 95, 102, 146]). However, which level of the linguistic hierarchy is perceptually and behaviorally salient depends very much on the task and the situation a perceiver is in. As McNeill and Lindig ([102], p. 430) have aptly put it, "what is 'perceptually real' is what one pays attention to." The validity of the basic linguistic categories, questions of detail aside, is guaranteed by the success of linguistic analysis. Linguistic units provide us with a vocabulary in which to describe the time course of accumulation and perceptual processing of linguistic information. Even though the perceptual processes themselves may be of an analog nature, we need discrete concepts to theorize and communicate about these processes. From this perspective, it is not an empirical issue but a fact that perceivers process features, phonemes, syllables, words, etc., since they are what speech is made of. Their awareness of these categories is another matter that shall not concern us here. (See [90, 99, 106].) Clearly, speech perception generally proceeds without awareness of all but the highest levels of description (i.e., the meaning of the message).

The fourth assumption is that there are perceptual processes in the brain that map sensory inputs onto internal structures. While such processes have been traditionally assumed in psychology since the demise of radical behaviorism, a new challenge (to the other assumptions as well) comes from the so-called direct realist school of perception, which claims that perceptual systems merely "pick up" the information delivered by the senses [54, 60]. I will return to this issue later. Here I merely note that the same input is not always perceived in the same way. Contextual factors, past experience, expectations, and strategies may alter the perceptual outcome, and this seems to require the assumption of perceptual processes that mediate between the input and the perceiver's interpretation of it. Whether these processes (and indeed, integration and segregation as such) are thought of as neural events with actual time and space coordinates or as abstract functional relationships between physical and mental descriptions is irrelevant to most of the research I will discuss here.

Having attempted to justify the four principal assumptions, it remains for me to mention two issues that are important in much research on perceptual integration and segregation. One is the question of whether the processes inferred are specific to the perception of speech or whether they represent general capacities of the auditory or cognitive system. By a speech-specific function I mean one that operates on properties that are unique to speech. There is no question that general capacities to integrate and segregate are common to all perceptual and cognitive systems. Speech perception presumably results from a combination of general and speech-specific perceptual functions (see, e.g., [39]). Just as speech resembles other sounds in some respects and differs in others. One frequent research strategy, therefore, is to determine whether or not *particular* instances of integration or segregation can be observed in both speech and nonspeech perception. This question can be asked only if the physical characteristics of speech and nonspeech stimuli are comparable--a condition that is notoriously difficult to satisfy (see, e.g., [112]). The mental descriptions of speech and nonspeech are, by definition, different at some higher level; thus the empirical question is whether that level is engaged in a particular integrative or segregative process.

The other issue is whether a particular integrative or segregative function is obligatory or optional. This question is sometimes linked with that of speech-specificity in that a higher-level, speech-specific function might seem easier to disengage than a lower-level auditory one. This is true in so far as adopting the deliberate strategy of listening to speech as if it were nonspeech (which is often difficult to achieve) may have the effect of eliminating certain forms of integration or segregation. It seems to be difficult or impossible to disengage phonetic processes through conscious strategies within the speech mode (e.g., by linguistic parsing [135, 136]). Moreover, it has been suggested [86] that some speech-specific functions do not really represent a "higher" level of perception but rather a mode of operation that, because of its biological significance, takes precedence over nonspeech perception, and if so, these functions may indeed be difficult to manipulate. On the other hand, in the auditory (nonspeech) mode listeners often have a variety of perceptual strategies available, especially when there are few ecological constraints on the stimulation, even though certain functions of peripheral auditory processing are surely obligatory. Thus, although it is useful to gather information about the relative flexibility of a process, this may not bear directly on the question of speech-specificity, as both speech and nonspeech perception are likely to involve levels of varying rigidity.

One final prefatory remark: Although one may legitimately talk about the integration of syllables into words and of words into sentences, or about the segregation of syntactic constituents from each other, I am not going to consider such higher linguistic processes in the present review. By speech perception I mean primarily the perception of phonetic structure without regard to lexical status or meaning, and my review is restricted accordingly.

INTEGRATION

The function of integrative processes is to provide coherence among parts of the input that "belong together" according to some perceptual rule or criterion. Auditory integration occurs within the physical dimensions of time, (spectral) frequency, and even space (in the case of artificially split sources); thus it creates temporal, spectral, and spatial coherence of sound sources. In part this is due to the limited resolution of the auditory system along each of these dimensions, but auditory events will often cohere even when there are discriminable changes within them. The larger these changes are, the more noteworthy the integrative process will seem to us. The perception of phonetic structure involves, in addition, integration of relevant information across all physical dimensions of the speech signal--a function requiring higher-level perceptual or cognitive mechanisms.

Temporal integration

Basic processes of sensory integration and auditory organization ensure the temporal coherence of any relatively homogeneous auditory input, including components of speech. This form of integration is so obvious as to hardly deserve comment. Thus, for example, successive pitch periods of a vowel are perceived as belonging together (i.e., as a single vowel, not two or many) even though their duration and spectral composition may change as a function of intonation, diphthongization, and coarticulation. While there may be a physical basis for subdividing a sound into smaller units such as individual pitch pulses or transition versus steady state, the rate and extent of change from one unit to the next are too small to disrupt sensory integration. Nevertheless, changes occurring within such units (e.g., transitions in a vowel or fricative noise) may have perceptual effects. That is, perception of temporal coherence does not imply insensitivity to changes over time, only that these changes are not large enough to cause perceptual segregation.

Growth of loudness. Temporal integration at this most elementary level has the consequence that, as the duration of a relatively homogeneous sound increases, its perceived loudness or perceptual prominence will also increase, up to a certain limit. In psychoacoustic research, the lowering of the detection threshold and the growth of loudness with increasing stimulus duration are well-established phenomena (see, e.g., [26, 192]). The time constant of the (exponential) integration function is about 200 ms, which encompasses the durations of virtually all relatively homogeneous speech events. While loudness judgments or explicit threshold measurements are uncommon in speech perception research, tne effect of an increase in the duration of a signal portion can be shown to be phonetically equivalent to that of an increase in its intensity, especially when the relevant signal portion is brief.

One example is provided by studies in which the duration and relative intensity of aspiration noise were varied orthogonally as cues to the voicing distinction in synthetic syllable-initial English stop consonants [31, 126]. Although the trading function obtained was much steeper than the typical auditory temporal integration function, it bore some similarity to integration functions obtained in an auditory backward masking situation [189], which is not unreasonable in view of the following vowel. It seems likely that the observed time-intensity reciprocity reflects basic properties of the auditory system, rather than speech-specific processes. Indirect support for this hypothesis comes from a study showing that the trading relation between aspiration duration and intensity holds regardless of whether or not listeners can rely on phonemic distinctions in discriminating speech stimuli [131]. In another recent study, stop consonant release burst duration and intensity were varied in separate experiments as cues to stop consonant manner in /s/-stop clusters [134]. Since both parameters proved to be perceptually relevant, a trading relation between them was implied. An analogous conclusion may be drawn from an older informal study [88], in which the duration and intensity of stop closure voicing were varied as cues to the perceived voicing status of an intervocalic stop consonant.

Auditory short-term adaptation. An effect closely related to temporal integration is that the auditory nerve fibers responsive to a continuous sound become increasingly adapted. Auditory adaptation is a topic of great interest to psychoacousticians and auditory physiologists, who have identified at least three different time constants of adaptation in animals (see, e.g., [45]). So-called auditory short-term adaptation, with a time constant of about 60 ms, seems the most relevant to phonetic perception. Although ongoing adaptation seems to have no direct perceptual consequences, the recovery of auditory nerve fibers following the offset of a relatively homogeneous stimulus results in reduced sensitivity to other, spectrally similar inputs for a short time period. Consequently, the auditory representation of a speech component whose spectrum overlaps that of a preceding segment will be modified. A striking demonstration of such an interaction was provided in recordings from cats' auditory nerves responding to synthetic /ba/ and /ma/ syllables [34, 35]. Even though the two syllables were identical except for the nasal murmur in /ma/, the auditory response at vowel onset was very different. The murmur, having strong spectral components in the low-frequency range, effectively acted as a high-pass filter, reducing the neural response in the low-frequency region at vowel onset. Recent experiments suggest, however, that this particular auditory interaction has no important consequences for perception of nasal consonants under normal listening conditions [138]. In a more artificial situation, Summerfield and colleagues [160, 162] have demonstrated an auditory aftereffect attributed to short-term adaptation: A sound with a uniform spectrum was perceived as a vowel when preceded by a sound whose spectrum was the complement of the perceived vowel's spectrum. Generalizing to natural speech, these authors pointed out that auditory adaptation effectively enhances spectral change and thus may aid phonetic perception in adverse listening conditions.

One general lesson to be learned from psychoacoustic research on temporal integration, adaptation, and other auditory interactions is that adjacent portions of the speech signal should not be thought of as mutually independent in the auditory system. Whenever a particular component is singled out for attention in careful analytic listening (to the extent that this is possible), influences of surrounding context on the perceived sound must be reckoned with. It is important to keep in mind, however, that listeners normally do not listen analytically but rather attend to the continuous pattern of speech. All peripheral auditory transformations are a natural part of the pattern and, because of past learning, are also represented in a listener's long-term memory of phonetic norms, which provide the criteria for phonemic classification in a language. Since auditory input and central reference both incorporate the distortions imposed by the peripheral auditory system, these distortions cannot be said to either help or hinder speech perception [138]. Only a change in auditory transformations, as might be caused by simulated or real hearing impairment, would prove disturbing to listeners; in normal speech perception, peripheral auditory processes probably do not play a very important role.

Spectral integration

Most speech sounds have complex spectra determined by the resonance frequencies of the vocal tract. Formants are usually visible as prominent energy bands in a spectrogram or as peaks in a spectral cross-section. Why are these bands perceived as a single sound with a complex timbre and not as separate sounds with simpler qualities? Why, indeed, are the individual harmonics of periodic speech sounds not heard as so many simultaneous tones? Even though these questions are provoked by our instrumental and visual methods of spectral analysis, they are not unreasonable, since the ear operates essentially as a frequency analyzer. One answer to these questions is that we *do* proeess these spectral components, only we are not conscious of them and find it difficult to focus selectively on them when asked to do so. Multidimensional statistical analyses of vowel similarity judgments have confirmed that the lower formants function as perceptually relevant dimensions, even though they seem to blend into a complex auditory quality [56, 115, 119], and psychoacoustic pitch matching tasks have revealed that listeners can detect a number of lower harmonics in a complex periodic sound (e.g., [110, 114]). Some central integrative function must be responsible for the perceptual coherence and unity of all these spectral components.

Critical bands. Some spectral integration does take place in the peripheral auditory system. A large amount of psychoacoustic research has established the concept of critical bands, i.e., frequency regions over which spectral energy is integrated, and whose width increases with frequency in a roughly logarithmic fashion [105, 190]. It is now quite common to represent speech spectra on a critical-band frequency scale (the Bark scale) to better take account of the resolving power of the auditory system. However, critical bands cannot account for the fact that formants are integrated into a unitary percept, because the lower formants of speech are usually several critical bands apart, and thus potentially separable. Even the lower harmonics, especially of female and child speech, are spaced more than 1 Bark apart. Critical bands may explain why higher harmonics and higher formants are not well resolved auditorily, but these spectral components do not contribute much phonetic information.

It is difficult, therefore, to point to any direct consequences of critical band limitations for speech perception, except in hearing-impaired listeners, whose critical bandwidths are abnormally large. A recent study by Celmer and Bienvenue [21] may serve as an example. These investigators digitized speech materials, degraded their spectra by simulating critical band integration ranging from one-half to seven times the normal widths, converted the manipulated spectra back into sound, and presented them to groups of normal listeners and to hearing-impaired listeners believed to have abnormally wide critical bandwidths according to independent psychoacoustic tests. The results showed that the degree of critical bandwidth filtering required to cause an intelligibility decrement was directly related to the subjects' measured critical bandwidth. Thus, normal subjects were sensitive to filtering at twice the normal bandwidths, while hearing-impaired subjects, though their intelligibility scores were lower to begin with, tolerated up to five times the normal bandwidths before any decrement in intelligibility occurred. Many other studies, too numerous to review here, have examined correlations between measures of critical bandwidth (or frequency resolution) and measures of speech perception in hearing-impaired individuals, with mixed results (see, e.g., [44, 152]). The looseness of the correlation may be accounted for by the facts that speech perception engages higher-level functions that help overcome peripheral limitations, often requires only relatively coarse spectral resolution, and relies on other physical parameters besides spectral structure.

Integration of harmonics. Given that the lower harmonics of a periodic speech sound are not automatically integrated by the peripheral auditory system, not to mention the lower formants themselves, the question of why they are grouped together in perception still needs to be answered. The most general answer is that they share a "common fate": They usually start and end at the same time; they are at integral multiples of the fundamental frequency; they have similar amplitude envelopes; and there is no alternative grouping that suggests itself. Below I will have more to say about the factors that may cause segregation of harmonics. Principles of auditory organization have received much attention in recent years (see, e.g., [10, 28, 184]), and one interesting conclusion from that research is that, even at such a relatively early stage in auditory processing, speech-specific criteria begin to play a role. They are speech-specific in the sense that a listener's tacit knowledge of what makes a good speech pattern influences the perceptual grouping of auditory components, as presumably does knowledge of other familiar auditory patterns. Yet another answer to the question of why harmonics (and formants) are grouped together is, therefore: They make a speech sound--that is, a complex sound that could possibly have emanated from a human vocal tract.

If it is the case that formant frequencies are salient parameters of speech perception (an assumption that is not made by some researchers who favor a whole-spectrum approach; e.g., [7, 154]), then it is of interest to ask how listeners estimate the actual resonance frequencies of the vocal tract from the energy distribution in the relevant spectral region. This question is especially pertinent with respect to the first formant (F1) in periodic speech sounds, for which critical bands are narrow and frequency difference limens are small. This means that the actual F1 frequency often falls between auditorily resolvable harmonics. Early work by Mushnikov and Chistovich [107] suggested that the brain takes the frequency of the single most intense harmonic as the estimate of F1. Later studies [1, 18], however, have indicated that the subjective F1 frequency corresponds to a weighted average of the two most intense harmonics, and one experiment [30] has shown that the perceptual boundary between /I/ and /e/ can be affected by the intensity of as many as five harmonics between 250 and 750 Hz, spaced 125 Hz apart. This indicates that the weighting function applied by the speech perception system in estimating formant frequencies extends over several critical bands (which are 100 Hz or less in this frequency region). The function is also asymmetric, giving more weight to higher than to lower harmonics, which may reflect a speech-specific constraint related to the fact that changes in actual F1 frequency affect primarily the amplitudes of the higher harmonics in the vicinity of the spectral peak [1]. Listeners thus seem to have tacit knowledge of the physical constraints on the shape of the vocal tract transfer function [29].

Integration of formants. This leads us to the more general question of whether the speech perception system integrates over adjacent formants (or any two peaks in the spectrum) when they are close in frequency but not within a critical band. It has been known for a long time that reasonable approximations to virtually all vowels can be achieved in synthesis with just two formants, and even with a single formant in the case of back vowels [33]. Delattre et al. [33] noted that the approximations were best when the two formants replaced by a single formant were close in frequency (F1 and F2 in high back vowels; F2 and F3 in high front vowels), and that the best single-formant substitute tended to be intermediate in frequency, suggesting that closely adjacent vowel formants form a perceptual composite or average. This idea was later elaborated by the Stockholm research group [18, 19] into the concept of F2', a hypothetical effective formant intermediate in frequency between F2 and F3 (except for /i/, where it falls between F3 and F4). These authors developed a formula for calculating F2' from F1, F2, F3, and F4, which gave good approximations to the results of perceptual matching experiments.

More recently, Chistovich and her collaborators have conducted a number of experiments on the "center of gravity" effect--the demonstrable phonetic equivalence of a single formant to two adjacent formants of varying frequency and/or intensity (see [22] for a review). One important question concerned the critical frequency separation of the two formants beyond which no satisfactory single-formant match could be achieved; it turned out to be about 3.5 Bark, i.e., 3.5 critical bands [23]. This finding has received considerable attention. For example, the 3.5 Bark limit has been related to the separation and boundaries between English vowel categories in acoustic space [166], and it has been used, together with the center of gravity concept, to explain perceived shifts in the height of nasalized vowels, which often have two spectral prominences in the F1 region [4].

It is noteworthy, however, that already Delattre et al. [33] were unable to achieve satisfactory single-formant matches to arbitrary two-formant patterns that did not correspond to familiar vowel categories. This finding, which was replicated by Traunmüller [172, 174] suggests that spectral integration over 3.5 Bark is tied to the perception of phonetic (or phonemic) categories. Specifically, it may reflect the resolution of the auditory long-term memory in which phonetic reference patterns are stored [174]. Indeed, it is an open question whether the 3.5 Bark limit explains the acoustic spacing of vowel categories [166], or whether it is the other way around. A recent study by Schwartz and Escudier [151], however, provides evidence that the 3.5 Bark limit is not the consequence of phonemic categorization. Their data suggest that there is indeed a higher level of auditory representation that serves phonetic classification and includes wide-band spectral integration. The cause of this integration is unknown at present.

Redintegration of artificially separated spectral components. Ultimately, it must be a higher-level process that decides whether a spectral array constitutes a single event or several. Integration over the whole spectrum is the natural state of affairs, since most natural sounds have complex spectra and could not easily be recognized if integration were not the default operation. Even an unrelated set of pure tones is perceived as a single complex structure when sounded simultaneously, as long as no alternative organizations suggest themselves [63, 77]. Such integration is disrupted by temporal or spatial separation of signal components, however; for example, the "auditory profiles" studied by Green and his coworkers are not well perceived when the sinusoidal components are divided between the two earphone channels [64]. With familiar natural events such as speech, perceptual coherence of spectral components may be centrally guided and hence greater and more resistant to disruption. One possible example of this is the phenomenon called spectral-temporal fusion [27] or duplex perception [84], which has been studied extensively in recent years.

Precursors of this research are found in experiments where the formants of synthetic syllables were separated and presented to opposite ears (e.g., F1 to one ear and F2 and F3 to the other). It was found early on that this presentation gave rise to an intact speech percept, with little or no awareness of separate stimuli in the two ears [14]. Similar fusion of dichotic stimuli into a single perceived sound is observed with complete synthetic syllables in the two ears [122] and even with harmonically related tones [37]. More surprising is the finding that perceptual integration continues to occur even when listeners are aware of separate stimuli in the two ears. Thus, Cutting [27] presented the dichotically separated formants at different fundamental frequencies and observed that subjects still reported the percept corresponding to the combination of the formants. (For similar effects with diotic presentation, see [28]). In what is now called the duplex perception paradigm, Rand [120] presented the formant transitions distinguishing two synthetic consonant-vowel syllables (such as /da/ and /ga/) to one ear and the remainder common to the two syllables (the "base") to the opposite ear. In this situation, listeners continue to report one or the other syllable depending on which formant transition is presented, even though that transition is also heard simultaneously as a lateralized nonspeech "chirp." The intact syllable (not the base) is heard in the ear receiving the base. Thus, subjectively at least, auditory fusion takes place despite the auditory segregation of the chirp--a paradoxical situation. This fusion continues to operate when the two signal components are presented at different fundamental frequencies [27] or with slight temporal offsets [139]. A very similar phenomenon can be produced diotically by making the critical formant transition audible through temporal offset [139], amplification [187], or different fundamental frequencies (informal observations). None of these manipulations, within certain limits, destroys the fused speech percept.

One interpretation of these findings [86] is that a specialized speech "module" is responsible for the perceptual integration and apparent fusion, whereas the general auditory system is responsible for the separate chirp percept. Bregman [11], on the other hand, has proposed that the paradoxical co-occurrence of fusion and nonfusion arises from conflicting cues for integration and segregation in the general process of "auditory scene analysis." He and other students of auditory organization have stressed the relative independence of What and Where decisions in auditory perception [13, 28, 38, 184]. It seems that auditory components that have been segregated can nevertheless be recombined in the perception and classification of familiar sound structures. That this recombination in the duplex perception paradigm is genuinely perceptual and not cognitive is indicated not only by the subjective impression of an intact syllable but by the fact that the components (chirp and base) presented by themselves generally do not suggest the "correct" phonetic percept [142].

Integration of phonetic information

Speech consists of a sequence of diverse sound segments which, as everyone knows, do not correspond directly to linguistic units. Changes in spectral structure are often very rapid and lead to great spectral heterogeneity over time. Equally striking is the alternation of qualitatively different sound types (periodic vs. aperiodic, as well as silence). Nevertheless, listeners perceive a coherent event, and thus believe speech to be a coherent stream of sounds. Since there is absolutely no reason to assume that very disparate sound structures are automatically integrated by the auditory system, the subjective impression of auditory continuity must be due to higher-level articulatory and linguistic properties of cohesiveness that capture the listener's attention--a kind of categorical perception (see [132]).

How can our brain perform integrative feats in speech perception that exceed the capabilities of the auditory system? One possibility is that there exists a biological specialization in humans, a "speech module," which performs this task [49, 86]. Alternatively, the answer may be mental precompilation as a consequence of perceptual learning [75]--an assembled module, as it were. What distinguishes speech perception from the auditory perception of arbitrary tones and noises (but not necessarily from the perception of other ecologically significant auditory events) is that the input can be mapped onto meaningful units of various sizes. The integration of the auditory components relating to each unit represented in the perceiver's long-term memory has taken place long ago during the process of speech and language acquisition; it may be instantiated neurally as a flexible (context-sensitive) system of interconnections [46, 75]. These precompiled units then enable a perceiver to immediately relate a number of functionally independent auditory features to a common phonetic percept. Some interesting (and arduous) attempts to simulate this process of perceptual learning and unit formation in nonspeech auditory perception have been reviewed by Watson and Foyle [183], who stress the importance of central processes in the identification and discrimination of complex stimuli. Experienced Morse code operators exhibit similar skills of "integrating" the acoustic dots and dashes into larger units [17], and so do probably perceivers of other meaningful acoustic events in our environment [70, 180], although in none of these instances does the auditory stimulus structure recede as much from awareness as it does in speech perception. From this perspective, speech is unique not so much because it requires specialized perceptual and cognitive functions but because it is structurally different, having originated in the articulatory motor system. Our biological specialization may simply lie in the fact that we can mentally represent a system that complex.

"Integrated" auditory properties. The ability to integrate over dynamically changing sound patterns has occasionally been attributed to the auditory system. Thus, Stevens and Blumstein [8, 153, 154] hypothesized that the onset spectrum following the release of stop consonants provides invariant acoustic correlates of place of articulation. Since there are often rapid spectral changes immediately following the release, and since a spectrum cannot be computed instantaneously, the hypothetical auditory onset spectrum must derive from an integrative process. Stevens and Blumstein hypothesized that the human auditory system integrates over about 25 ms and thus extracts the acoustic property relevant to place of articulation.

The work of Stevens and Blumstein has come under criticism in recent years. Kewley-Port [73] has argued that, for all we know, the auditory system tracks spectral changes over time intervals shorter than 25 ms and presumably delivers information about these changes to phonetic decision mechanisms. Perceptual studies [8, 74] have suggested that listeners are indeed sensitive to spectral changes immediately following the release of stop consonants. The onset spectra themselves do not appear to be as invariant as was originally claimed [81, 164]. Blumstein and her students meanwhile have abandoned the search for invariant properties in onset spectra and have instead gone on to define integrated properties based on the relationship between spectra or intensity measures obtained some interval apart [71, 79, 81]. Even though some of these properties are quite complex, their derivation is still attributed to the auditory system by these researchers. However, since it seems highly implausible that there are general auditory functions which yield so specialized a result, the epithet "auditory" should perhaps be understood as referring merely to the input modality. Clearly, out of the infinity of possibilities, particular relational properties are selected on the basis of phonetic relevance. The integrative computational process thus is specific to speech perception.

Integration of silence and other signal components. Even though it seems unlikely that the auditory system integrates over spectral variation in the speech signal lasting tens of milliseconds, this hypothesis has some measure of plausibility, given the basic continuity of the signal changes. There are many more abrupt changes in the speech signal, however, such as changes in source (from voiced to voiceless, or vice versa), in spectrum (such as /z/ followed by /u/), and in intensity (into and out of closures filled with nasal murmur, voicing, or silence), usually in several of these dimensions simultaneously. It would seem absurd to attribute to the auditory system the capability to integrate across such dramatic signal changes, since the task of auditory perception is to detect changes, not to conceal them. Nevertheless, there is ample evidence from perceptual experiments that listeners can integrate phonetic information across such acoustic discontinuities in the signal. Clearly, this integration must be a higher-level function in the service of speech perception.

Perhaps the most striking instance is the perception of silence in speech. (I have in mind brief silent intervals of up to 200 ms duration, not longer pauses.) From an auditory perspective, silence is the absence of energy, a gap, an interruption that separates the signal portions to be perceived. In speech perception, however, silence is bridged by, and participates in, integrative processes. Rather than being the neutral backdrop for the theater of auditory events, silence is informationally equivalent to energy-carrying signal portions. Relative duration of silence has been shown to be a cue for the perception of stop consonant voicing [76, 87, 116], manner [3, 134, 141], and place of articulation [3, 116, 133]. Why does silence function in this way in speech? The answer must be that it is an integral part of the acoustic patterns that a human listener has learned to recognize. Being an acoustic consequence of the oral closure connected with (voiceless) stop consonants, it has become a defining characteristic of that manner class. Lawful variations in its duration as a function of voicing status or place of articulation also have assumed the function of perceptual "cues." A listener's long-term representation of the acoustic pattern corresponding to a stop consonant thus includes the spectro-temporal properties of the signals preceding and following the closure as well as the closure itself. (The precise nature of that mental representation, or rather of our description of it, need not concern us here; it suffices to note that listeners behave as if they knew what acoustic pattern to expect.) The silence thus is not really "actively" integrated with the surrounding signal portions; rather, the integration has already taken place during past perceptual learning and is embodied in the perceiver's long-term knowledge of speech patterns to which the input is referred during perception.

Not only is silence integrated (in the sense just discussed) with surrounding signal portions in phonetic perception, but acoustically rather different components of the signal are integrated with each other. Thus, for example, the spectrum of a fricative noise and the adjacent vocalic formant transitions both contribute to perception of a prevocalic fricative consonant [91, 185], the formant transitions in and out of a closure contribute to stop consonant perception [168], etc. Just as articulation distributes acoustic information about individual phonemes over time, perceptual integrative functions collect that information and relate it to internal criteria for linguistic category membership. An especially interesting demonstration of this was provided quite recently by Tomiak et al. [171]. Using a well-known technique [59] for testing listeners' ability to selectively attend to stimulus dimensions, they showed that the "fricative noise" and "vowel" portions of noise-tone analogs to fricative-vowel syllables were processed separately by subjects who perceived the stimuli as nonspeech sounds, but were processed integrally by subjects who had been told that the stimuli represented syllables. These latter subjects were unable to selectively attend to either of the two stimulus portions, even though coarticulatory interactions were not present in the noise-tone stimuli. Listeners in the "speech mode" thus seem to process auditory components of speech in an integrative manner even some of the information to be integrated is not actually there; they are scanning for it, as it were.

Independent aspects of the speech signal that contribute to the same phonemic decision combine according to a simple decision rule, as demonstrated in many experiments by Massaro (e.g., [36, 98]). It is possible to trade various of these cues, changing the physical parameters of one while changing those of another in the opposite direction, without altering the phonemic percept. This phenomenon, often referred to as "phonetic trading relations," has been demonstrated in a large number of studies (reviewed in [129]). Fitch et al. [47] showed that listeners have great difficulty discriminating two phonemically equivalent stimuli created by playing off two cues against each other, and they argued that this reflects the operation of a special phonetic process that makes auditory differences unavailable to perception. Whether the process of phonetic information integration is speech-specific is debatable [138], even though it is agreed that the information being integrated is speech-specific. Listeners' difficulty in discriminating phonemically equivalent stimuli is familiar from classical categorical perception research (reviewed in [132]). Experiments on phonetic trading relations that include identification and discrimination tests [6, 47] are generalized categorical perception tasks, in which several physical parameters are varied simultaneously. If each parameter variation by itself is difficult to discriminate except when it cues a category distinction, then joint variations in these parameters will be almost as difficult to discriminate unless a phonemic contrast is perceived. This does not mean, however, that auditory discrimination of such variations is impossible. Appropriate training and use of low-uncertainty discrimination paradigms has been shown to reduce or eliminate categorical perception of single dimensions [20, 128], and it is likely that similar training would enable subjects to discriminate simultaneous variations in several cues, thus demonstrating that their integration does not take place in the auditory system (see also [6]). There is also evidence that certain phonetic trading relations occur only when listeners can make phonemic distinctions, but not within phonemic categories [131].

In summary, the various forms of phonetic cue integration seem to represent, for the most part, speech-specific functions in so far as the articulatory processes and the corresponding linguistic categories that cause the integration are specific to speech. This idea is embodied in Massaro's "fuzzy logical model" of phonetic decision making [98], which assumes that, for each phonemic category, listeners have internal criteria for the degree of presence of various acoustic features in the speech signal. Diehl and his colleagues have recently argued that many trading relations may have a general auditory basis [39, 109]. While their research may show that some trading relations (especially those within a physical dimension) indeed rest on auditory interactions, this is unlikely to be true for the many trading relations that cut across physical dimensions. Although phonetic perception is certainly not immune to auditory interactions, cue integration appears to be mainly a function of speech-specific classification criteria.

Phonetic context effects. Perceivers not only integrate cues directly pertaining to a particular phoneme or complex of articulatory gestures, but they adapt their perceptual criteria to the surrounding phonetic context. Examples of such phonetic context effects are the shift in the /s/-/ʃ/ category boundary depending on the following vowel [78, 91] and the shift in the /b/-/p/ voice-onset-time category boundary depending on the speaking rate or duration of the surrounding segments [65, 103, 157]. For reviews, see [103, 129, 140]. As in the case of phonetic trading relations, some of these effects may have general auditory processing explanations; thus, for example, the effect of vowel duration on perception of the /ba/-/wa/ distinction [104] probably is not speech-specific, as a comparable effect has also been obtained with nonspeech stimuli [113]. Many other effects, however, seem to reflect listeners' tacit knowledge of coarticulatory dependencies in speech production. For example, the different /s/-/ʃ/ boundaries in the context of rounded and unrounded vowels may be related to the occurrence of anticipatory liprounding during the constriction phase in utterances such as "soup" but not in "sap." In a series of experiments using cross-spliced fricative noises and vowels, Whalen [186, 188] has shown that even when the fricative noise itself is quite unambiguous, subjects' reaction time in a fricative identification task is influenced by the following vocalic context, being slower when the fricative noise spectrum is not exactly what would be expected in that context (cf. the study by Tomiak et al. [171] reviewed above). In an unpublished series of experiments, Repp [123] demonstrated an effect he dubbed "coperception," which consisted of slower reaction times to decide that the two consonants are the same in the stimulus pair /aba/-/aba/ than in the pair /aba/-/abi/, even though the pre-closure (VC) portions of these synthetic VCV stimuli were identical in both cases. That is, even though subjects could have made their decisions after hearing /ab/ in the second member of a stimulus pair, they somehow had to take the CV portions of the stimuli into account and then were slowed down by the inequality of the vowels. All these studies show that perceivers integrate all information that possibly could bear on phonetic decisions, and this integration often seems obligatory in nature. It requires special instructions, special (nonphonetic) tasks, and usually some amount of training to disengage phonetic integration mechanisms in the laboratory [6, 127, 128, 136].

Cross-modal integration. In natural speech communication, humans make use not only of auditory but also of visual information, if available. Audiovisual integration at the level of phoneme perception has been a research topic of considerable interest since the discovery by McGurk and MacDonald [101] that subjects presented with certain conflicting auditory and visual speech stimuli report that they "hear" what they see. Their findings have been replicated and extended in a number of studies [89, 97, 157] and others). Massaro [96, 97] has shown that a general rule of information integration based on the degree to which signal features match expected feature values can explain audiovisual integration, auditory cue integration, as well as many other forms of perceptual integration outside the domain of speech. This suggests that we may be dealing with a general function following basic laws of decision theory. Liberman and his collaborators [85, 141], on the other hand, have argued that integration of speech cues, within or across modalities, occurs because they represent the multiple, distributed consequences of articulatory acts or gestures. Some internal reference to processes of speech production is thus implied, as in the "motor theory" of speech perception [86]. However, this account is complementary rather than antithetic to Massaro's model: It is a theory of why integration occurs, whereas Massaro is concerned with how integration works. The phonemes of a language are articulatory events which have characteristic acoustic and optic consequences, and perceivers presumably have tacit knowledge incorporating both of these aspects. If a portion of the speech input satisfies certain auditory and visual criteria for phonemic category membership (as in Massaro's model) this also implies that the gestures characterizing a particular phoneme have been recovered (as in the motor theory). Whether the sensory or the articulatory aspect is stressed in a particular theory is largely a matter of philosophy and perhaps of economy. A complete theory must include both.

Audiovisual integration at the more global level of word, sentence, and discourse comprehension has, of course, been of interest for a long time in connection with hearing impairment and communication in noisy environments. Research on this topic has received a boost in recent years with the advent of modern signal processing technology and of cochlear implants. (See [158] for a review.) The information provided by residual hearing or by electrical stimulation of the auditory nerve supplements that obtained from lipreading to yield enhanced comprehension. In many respects, these two sources of information are complementary, with the auditory channel providing information that is difficult to see, and vice versa. What is of special interest in the present context is that audiovisual comprehension performance often seems to exceed what might be expected from a mere combination of independent sources of information. Thus, Rosen et al. [143] demonstrated that speech intelligibility is improved substantially when lipreading in hearing subjects is supplemented with the audible fundamental frequency contour, or even just with a constant buzz representing the occurrence of voicing. (See also [9, 62].) Since this auditory component by itself provides virtually no information about phonetic structure, it must be the temporal relationships between the auditory and visual channels that contribute to intelligibility [100]. Thus audiovisual speech perception is often more than the sum of its parts; in terms of Massaro's [96] model, the separate sources are integrated before central evaluation. The close integration of inputs from the two modalities is witnessed by anecdotal reports that voicing-triggered buzz accompanying lipreading may assume phonetic qualities [159].

The theoretical issues raised by audiovisual integration have been discussed thoroughly by Summerfield [159]. He, too, concludes that auditory and visual cues to linguistic structure are integrated before any categorical decisions are made. There are four ways of conceptualizing how this integration occurs: (1) The two channels make independent contributions to linguistic decisions, but temporal relationships provide a third source of information. (2) The visual information is translated into an auditory metric of vocal tract area functions. (3) The auditory information is translated into a visual metric of articulatory kinematics. (4) Both are translated into an abstract representation of dynamic control parameters of articulation. This last-mentioned approach [15, 72] may ultimatelv provide the most economic description of speech information in both modalities, and thus may yield the most appropriate vocabulary in which to describe intermodal integration.

Higher-level integration. Human listeners not only integrate auditory and visual information about a speaker's articulations, but they also bring phonotactic, lexical, syntactic, semantic, and pragmatic expectations to bear on their linguistic decisions, provided the auditory and/or visual input is sufficiently ambiguous to give room to effects of such expectations. Some well-known demonstrations of effects in this category are the "phoneme restoration" phenomenon discovered by Warren [181] and studied more recently by Samuel [145], in which lexical expectations fill in missing acoustic information, as it were; the lexical bias effect reported by Ganong [58] and replicated by Fox [57], which causes a relative shift in the category boundaries on acoustic word-nonword (e.g., DASH-TASH versus DASK-TASK) continua in favor of word percepts; and the "fluent restorations" in rapid shadowing of semantically anomalous passages [94]. These phenomena, and a host of related ones often referred to as "top-down" effects, may be considered general forms of cognitive information integration in speech perception. Indeed, Massaro [96] has argued that the rules by which such higher-level information is integrated with the "bottom-up" information delivered by the senses are the same by which acoustic (and optic) speech cues are integrated. Others argue that top-down influences should be strictly separated from bottom-up processes--that they represent general cognitive functions that operate outside the autonomous speech module [49, 86]. According to this second view, integration of bottom-up cues to phoneme identity is a fundamentally different process from the integration of bottom-up and top-down information. My own view in this matter is that speech perception at every level requires domain-specific knowledge stored in a perceiver's long-term memory. The processes by which this knowledge is brought to bear upon the sensory input are part of our metaphoric representation of brain function and thus are bound to be general [138]. In the absence of a radically different vocabulary in which to characterize the processes within a module (though such a vocabulary will perhaps emerge from the study of articulatory dynamics and coordination), the postulate of a speech module harks back to the "black box" of behaviorism. It is quite likely, of course, that phonetic perception is modular in the sense that integration of phonetic cues precedes, and is not directly influenced by, higher-level factors. This issue can be addressed empirically [49, 58, 145, 165]. My point here is that integration, whether it occurs inside a module or outside it, is conceptually the same thing: a many-to-one mapping. Indeed, Massaro's (e.g., [96]) extensive research suggests that the rules of information integration are independent of modularity.

SEGREGATION

The preceding section has illustrated the pervasiveness of integrative processes in speech perception. Much of perceptual and cognitive processing is convergent, with multiple sources of information contributing to single decisions, be they explicit or implicit. Nevertheless, we also need hypothetical mechanisms to prevent all information from converging onto every decision "node." Even though a perceiver's internal criteria for linguistic category membership will automatically reject irrelevant information, information that does not belong is nevertheless often potentially relevant. Thus, in the often-cited cocktail party situation, the voices of several speakers must be kept apart to avoid semantic and phonetic confusions. Various environmental sounds could simulate phonetic events and need to be segregated from the true speech stream. In the speech signal itself, information pertaining to speaker identity, emotion, room acoustics, etc., needs to be distinguished from the phonetic structure, and the overlapping consequences of segmental articulation need to be sorted out. These segregative processes have an important complementary role to play in speech perception: They ensure that integration is restricted to those pieces of information that belong together. Logically, segregation precedes integration, even though functionally they may be just the two sides of one coin. The more physically similar and intertwined the aspects to be segregated are, the more remarkable the segregative process will seem to us.

Temporal and spatial segregation

Without any doubt, there are several factors that enable perceivers to distinguish different sound sources or events, regardless of whether they are speech or not. One of these is temporal separation. Sounds occurring a long time apart will usually not be considered as belongilig to the same event, although they may come from the same source. In speech, a few seconds are usually enough to segregate phrases or utterances, and a few hundreds of milliseconds of separation usually prevent integration of acoustic cues into a single phonemic decision. One demonstration of this fact may be found in studies of the distinction between single and geminate stop consonants. In a classic experiment, Pickett and Decker [111] asked English-speaking subjects to distinguish between utterances such as "topic" and "top pick", varying only the duration of the silent /p/ closure. Between 150 and 300 msec were needed to obtain judgments of two /p/s (and two words) rather than just one; the precise duration depended on the overall speaking rate. (See also [108, 124, 125].) If two different stop consonants follow each other, as in the nonsense word /abda/, about 100 ms of silent closure are needed to prevent integration of the two sets of formant transitions into a single stop consonant percept [43, 124]. Dorman et al. [43] cued the perception of /p/ in "split" solely by inserting a silent interval between an /s/ noise and the syllable "lit" (a percept that may be said to be a pure temporal integration illusion), and subsequently investigated how much silence was needed before subjects reported hearing "s" followed by "lit." This duration turned out to be as long as 600 msec. A subsequent replication [136] obtained a shorter but still surprisingly long interval of 300-400 msec. To cite a final example, Tillmann et al. [170] investigated how much temporal offset of optically and acoustically presented syllables was needed to destroy the audiovisual integration effect discovered by McGurk and MacDonald [101]. It turned out to be 250-300 msec. These various situations have little in common, which explains the different results. The precise duration of the critical interval for segregation surely depends on many factors and does not reflect any general limits of temporal integration. Rather, within the auditory modality it may be related to the closure durations normally encountered in natural speech [111, 130].

Temporal asynchrony is a helpful cue in distinguishing speech from other environmental sounds. This was elegantly demonstrated in a series of studies by Darwin [29, 32], who investigated under what conditions a pure tone added to one of the (pure-tone) harmonics of a synthetic vowel was treated by listeners as part of the vowel spectrum or as a separate nonspeech event. Darwin showed that, when the tone coincided with the vowel, it affected the perceived vowel quality. However, when the onset of the tone preceded that of the vowel or, to a lesser extent, when its offset lagged behind that of the vowel, listeners excluded it from the phonetic information. Similar principles of segregation or "auditory stream formation" have been demonstrated in the perception of nonspeech sounds [12].

Another factor that may cause segregation is spatial separation. In real life, the separation of several simultaneous voices or of speech from background noises is often possible because they are perceived as coming from different locations. In the laboratory, presentation over the two channels of earphones has been used to induce segregation. One interesting case in which this form of spatial separation does *not* seem to prevent integration is split-formant or duplex perception, discussed above. Note, however, that in duplex perception one component of the speech signal (the "chirp") *is* segregated and heard as a separate auditory event; the paradox is that this event is still, at the same time, integrated with the speech in the other ear. (See [11].) There are many other instances, however, particularly those in which there is no temporal overlap between the two signals, where spatial separation is sufficient to disrupt perceptual integration. For example, informal observations suggest that, if the artificial "split" created by concatenating "s" and "lit" with some intervening silence is divided between the two ears, so that "s" occurs in one ear and "lit" in the other, this is exactly what listeners report hearing; that is, there is no /p/ percept any more. Similarly, when nasal-consonant-vowel syllables such as /mi/ or /ni/ are divided between the two ears, so that the nasal murmur occurs in one and the vocalic portion containing the formant transitions in the other, listeners have great difficulty identifying the consonant, or in any case do not perform better than if the two components were presented by themselves [137]. Of course, it is always possible to integrate independent sources of information at a cognitive level. These two examples illustrate the role of spatial separation as a segregating factor. Unfortunately, in real life both temporal and spatial separation are often unavailable as segregating agents, and listeners need additional means of sorting out the incoming stream of auditory information.

Spectral segregation

When irrelevant (speech or nonspeech) sounds are superimposed on speech, listeners have basically two means of segregation at their disposal: Segregation according to local spectral disparity, and according to spectro-temporal (and, in part, speech-specific) criteria of pattern coherence. There are, of course, many sounds in the environment, including those produced by most musical instruments, that are sufficiently different from speech to be perceived immediately as different sources. Local spectral segregation is not always effective, however, and for good reason: First, some nonspeech events (e.g., the pops of bottles or the hisses of steam valves) are spectrally similar to speech sounds and thus are difficult to separate from them locally. Second, and more importantly, speech itself is composed of acoustic segments of diverse spectral composition, and it would be counterproductive if listeners were prone to segregate them, because these segments more often than not map onto the same linguistic unit. Indeed, perceptual segregation of spectrally dissimilar natural speech components can usually be demonstrated only under special conditions, which rarely occur outside the laboratory. Thus, Cole and Scott [24] rapidly iterated fricative-vowel syllables and found that listeners sometimes reported two streams of events: a train of fricative noises, and a train of vowels, especially when the vocalic formant transitions were removed. A similar phenomenon was obtained with the repeated syllable /ska/ by Diehl et al. [40] who then used their findings to explain the different effects of /spa/ or /ska/ stimuli as adaptors (or precursors) in selective adaptation and pairwise contrast paradigms [147, 148]. The selective adaptation task requires cyclic repetition of a single stimulus, the adaptor, and thus may produce "streaming" of signal components, so that /spa/ is heard as /s/ and /ba/, with the phonological status of the stop consonant altered. Repp [128] was able to induce listeners through some training to segregate a fricative noise from a following vowel and "hear out" the spectral quality of the noise. Even the individual formants of vowels may segregate under certain conditions. Following earlier studies showing that it was difficult to perceive the correct temporal order of four rapidly cycling steady-state vowels [169, 182], Dorman et al. [42] found that this was because in such artificial sequences individual formants tend to group together and form separate auditory streams. There are anecdotal reports of phoneticians being able to "hear out" individual formants of vowels (e.g., [66, 150), but this ability has remained rare. Still, these various findings underline the fact that spectrally diverse components of the speech signal are *potentially* segregable; fortunately, however, they are perceptually integrated under normal circumstances.

When two different speech streams co-occur, differences in fundamental frequency, intonation pattern, or voice quality may provide cues for separation, in addition to higher-level factors such as syntactic and semantic continuity. Effects of this kind have been found in classical work on selective attention (reviewed in [176]). More recently, Brokx and Nooteboom [16] obtained a beneficial effect of differences in fundamental frequency and intonation on the identification of meaningless sentences presented against a background of a read story. In the much more artificial situation of two simultaneous steady-state vowels, Scheffers [149] and Zwicker [191] found an improvement in recognition performance when a fundamental frequency difference was introduced. Since the magnitude of the difference beyond one semitone did not seem to play a role, the function of F0 differences in this case seems to be to prevent fusion of the two sounds. Similar, though small, effects of F0 on identification scores have also been obtained in dichotic listening studies using synthetic syllables [67, 121, 167] or vowels [191].

The potential of fundamental frequency (F0) and voice quality cues to segregate *successive* portions of speech has also been demonstrated in the laboratory. The mechanisms studied here must be involved in separating different speakers from each other. Several relevant studies have used stimuli in which perception of a stop consonant rested on the duration of a silent closure interval. Dorman et al. [43] found that when the speech on each side of the silence was produced by different voices, the silence lost its perceptual effectiveness; that is, listeners did not integrate across it. On the other hand, it has been shown [118, 177] that silence retains its effectiveness between syllables produced by male and female voices if the general articulatory and intonational pattern is continuous across the two speakers (achieved by cross-splicing two intact utterances). When the second syllable was spliced onto a first syllable originally produced in utterance-final position, however, the phonetic effect of the silence was disrupted. Thus it seems that dynamic spectro-temporal information about articulatory continuity can override differences in F0 or voice quality. A disruptive effect of discontinuities in intonation on stop consonant perception has also been reported [117], but such an effect was absent in a recent study [135] in which a constant fricative noise preceded the critical silence, suggesting that the breaks in the F0 contour are effective only when voiced signal portions immediately abut the silent closure interval.

Segregation of linguistic and paralinguistic information

So far I have discussed segregation of two kinds: One separates speech from other, irrelevant sounds (including competing speech streams), and the other dissociates consecutive parts of the same speech stream--a laboratory-induced phenomenon to be avoided in natural speech communication. These segregative processes are "literal" in that they result in the perception of separate sound sources. Segregative processes are also essential, however, when listening to a single speech source, and for two reasons. First, the speech signal conveys in parallel, and largely over the same time-frequency channels, information about phonetic composition, speaker characteristics (vocal tract size, sex, age, identity, emotion), and room or transmission characteristics (reverberation, distortion, filtering). A listener needs to separate these three kinds of information, which Chistovich [22] has termed "phonetic quality," "personal quality," and "transmission quality," respectively. (See also [175].) Second, the acoustic information for adjacent phonemes is overlapped and merged, a phenomenon commonly referred to as coarticulation or "encoding." If phonemic units are to be recovered, the information pertaining to one phoneme needs to be separated from that for another--or so it seems. Both these kinds of segregation are not literal in the sense that they make a speech stream disintegrate perceptually; rather, they separate different aspects of a coherent perceptual event by relating these aspects to different conceptual categories or dimensions represented in long-term memory. They operate on the information in the signal, not on the signal itself.

Of the various types of information segregation of the first kind, that of separating vocal tract size information from phonetic information has received the most attention under the heading of speaker normalization. An explicit solution to this problem is of vital importance to automatic speech recognition as well as to any theory of speech perception. In fact, the focus has been so exclusively on the speaker-independent recovery of phonetic information that it is sometimes forgotten that listeners extract several kinds of information in parallel. Rather than "normalizing" their internal representation of the speech wave and discarding information in the process, they presumably use all available kinds of information to mutual advantage.

Studies of speaker normalization have, for the most part, been concerned with vowels rather than consonants, and with acoustic analysis and automatic recognition rather than with human perception. Older normalization algorithms often required knowledge of a speaker's whole vowel space or average formant frequencies (see [41]), whereas more recent work has focused on perceptually more relevant transformations based on parameters that are immediately available in the incoming speech signal (e.g., [163, 166, 173]). There have been relatively few perceptual studies on this topic; the general assumption has been that it is sufficient to define acoustic properties that are relatively speaker-invariant and also plausible in the light of what is known about the auditory system. Demonstrations of "perceptual normalization" usually show a performance decrement in a listening situation where speaker characteristics are varied rapidly and unpredictably, compared to one in which the speaker remains constant [80, 161, 178]. Although emphasis is sometimes placed on the perceptual "advantage" resulting from effective normalization, the negative consequences of presenting contrived and misleading stimuli are perhaps the more salient outcome of this research (which is by no means unique in this respect).

Analogous experiments have been conducted on normalization in the temporal domain--that is, on the perceptual separation of speaking rate from phonetic length (reviewed in [103]). An especially interesting question arises in research on tone languages, where the listener must segregate lexical tones from the overall intonation contour [25] and from speaker-dependent variation in F0 [82]. In that connection, it is noteworthy that there is mounting evidence (reviewed in [144]) that tone and intonation perception (and production) are controlled by opposite hemispheres of the brain. At least some forms of linguistic/paralinguistic segregation may thus have a basis in neurophysiological compartmentalization. A general conclusion to be drawn from research on perceptual normalization is that the auditory parameters underlying phonetic classification are not absolute quantities but relationships in the spectral and/or temporal domain, computed over a relatively restricted temporal interval, whereas properties signalling speaker sex or identity, emotion, speaking rate, etc., accumulate over longer stretches of speech and/or are based on more nearly absolute quantities.

Segregation of intertwined linguistic information

The emphasis on linguistic information in the vast majority of speech perception studies makes it difficult to find good examples of research on perceptual segregation of linguistic and (rather than from) nonlinguistic information. Examples of segregation of equivalent information are easier to find when only linguistic information is involved. This leads me to the final topic, one that has been of enormous significance in speech perception research--the problem of segmentation, that is, the perceptual separation of the overlapped acoustic correlates of adjacent phonemic units, particularly of vowels and consonants.

One traditional view of the listener's task has been that it is one of phoneme (or feature) extraction, including "compensation" for contextual influences on a segment's acoustic correlates (see [54]). Numerous studies have shown that listeners perceive segments as if they knew all the contextual modifications their acoustic representations undergo [129, 140]. Thus, for example, a fricative noise ambiguous between /s/ and /ʃ/ in isolation is perceived as /s/ when followed by /u/ but as /ʃ/ when followed by /a/ [91]. One way of describing this finding is that listeners "know" that anticipatory liprounding for /u/ may lower the spectrum of a preceding fricative noise, so they adopt a different criterion for the /s/-/ʃ/ distinction in that context. This view, which emphasizes the role of tacit phonetic knowledge in speech perception, has recently been elaborated by several authors (e.g., [48, 138]). The perceptual accomplishment seems more integrative than segregative from that perspective.

An alternative view, having an equally long history, has a recent proponent in Fowler [53, 54, 55] who has likened the separation of overlapping segmental information to mathematical vector analysis. According to her theory, listeners literally subtract or factor out the influences of one segment on another, so that invariant segments are "heard." Fowler conceives of phonetic segments as articulatory events, not as abstract mental categories (see the exchange on coarticulation between Fowler [50, 52] and Hammarberg [68]), though listeners are assumed to be able to judge their "sound" [53]. Several experiments [51, 53, 55] were intended to demonstrate this. They showed that subjects judge acoustically different representations of a segment to be more similar than acoustically identical ones if the former occur in their original contexts while the latter have been spliced into inappropriate contexts. However, since only the former match what listeners expect to hear in a given context, these results are also compatible with an alternative account based on tacit knowledge of contextual effects in speech production [129, 138]. That is, rather than having access to the sound of segments [53], listeners may have made their judgments on the basis of the discrepancy of the input from context-sensitive mental norms or prototypes.

Other recent experiments in a similar vein have addressed the separation of nasality and vowel height information in nasalized vowels. Kawasaki [71a] showed that English listeners judge vowels in /m_m/ environment as increasingly nasal as the surrounding nasal murmurs are attenuated; that is, when the nasal consonants are intact, the vowel nasality is attributed to (coarticulation with) the nasal consonants, as it were, and is "factored out" from the vowel percept. Building on this result, Beddor et al. [5] first established that there are different category boundaries on synthesized /bɪd/-/bæd/ and /bɪ̃d/-/bæ̃d/ continua. English listeners apparently interpret some of the spectral consequences of nasalization as a change in vowel height. However, when an appropriate "conditioning environment" was added in the form of a postvocalic /n/, the category boundary on the resulting /bɪ̃nd/-/bæ̃nd/ continuum was identical with that on the /bɪd/-/bæd/ continuum, as if listeners attributed the vowel nasality to (coarticulation with) the nasal consonant and "factored it out" in Fowler's sense. The result is equally compatible, however, with a theory that postulates context-sensitive vowel (or syllable) prototypes. Indeed, it may be difficult to come up with any decisive experiments. Mentalism and realism may simply represent different metatheoretical perspectives.

Current efforts at Haskins Laboratories to model articulation as a sequence of overlapping segmental gestures (e.g., [15, 72]) may ultimately provide ways of recovering these gestures from the acoustic signal and thus provide a machine implementation of Fowler's vector-analytic concept. A promising mathematical technique for achieving the same goal, based on principal components analysis of vocal tract area function parameters, has been proposed by Atal [2] and is currently being explored [92, 93]. The recovery of articulatory parameters from the acoustic signal remains a central problem in speech research because phonemes and alphabets surely represent an articulatory, not an acoustic classification. However, while a solution of this problem would bring us a great step forward, processes of integration and segregation would still be needed to translate the articulatory "score" into a sequence of discrete segments.

SPEECH PERCEPTION WITHOUT INTEGRATION AND SEGREGATION?

In the introduction, I discussed four basic assumptions: the separation of the physical and mental worlds, the existence of physical units, the existence of mental units, and the existence of processes relating the two kinds of units. Can a theory of speech perception do without them? The assumptions are not independent, of course: If the physical and mental worlds are distinct, they must receive different descriptions; to be easily communicable in the scientific world, these descriptions must be in terms of discrete concepts or units; and this results in certain functions or relationships between the two descriptive domains. If the physical and mental worlds were isomorphic, there would be no need for a theory of perception. If one or the other description were without units (more likely an error of omission than a deliberate theoretical choice), then perception would seem either entirely integrative or entirely segregative--not an attractive state of affairs. Denial of functions, however abstract, linking the two domains would merely impoverish perceptual theory. Certainly we need these functions in theories of auditory processing and organization. As to the perception of phonetic information, however, an alternative approach has been proposed.

This approach, stated most eloquently by Studdert-Kennedy [155] and Fowler [54], follows the "direct-realist" perspective of ecological psychology [61, 179]. Although it affirms the existence of linguistic units as articulatory events, it essentially abandons the distinction between the physical and mental domains. The segmental structure of speech (as characterized by the linguist or phonetician) is assumed to be ever-present on its way from the speaker's to the listener's brain. There is assumed to be a direct isomorphism between physical and mental descriptions of speech events (such as phonemes), though it is acknowledged that the appropriate physical and motor-dynamic descriptions have not been fully worked out. Thus this school of thought rejects the idea that the input is divided into parts that need to be integrated or segregated by the listener; rather, the input units are taken to be identical with the perceptual units--that is, they are already integrated or segregated with respect to more primitive acoustic or auditory units. The deliberate strategy of this philosophy is to eliminate classical problems in perceptual research (such as segmentation and invariance) by redefining and redescribing physical events. Rather than being attributed to the perceiver's brain, the burdens of information integration and segregation thus fall upon the investigator trying to find an "integral" description of "separate" speech events. However, this effort is equivalent to that of finding a principled explanation of perceptual integration and segregation: If we can show that certain pieces of input are always integrated, we might as well call them integral and treat them as a single piece in our descriptions--if we only had names for them. Behind the rhetoric and the different terminologies of mentalistic and realistic approaches lies a common goal: to arrive at the most economic characterization of linguistic structure in all its physical incarnations. Clearly, even speech research propelled by a mentalistic philosophy (still predominant in the field) must strive to minimize the work attributed to a speaker-listener's mind. But will we be able to relieve it of its entire burden to integrate and segregate? What we take away (in theory) is likely to re-emerge as logical conjunctions, disjunctions, and relational terms in our physical characterization of speech events. As long as we scientists communicate in conventional language, integration and segregation at some stage in our theories will be difficult to avoid.

ACKNOWLEDGMENTS

Preparation of this paper was supported by NICHD Grant HD-01994 to Haskins Laboratories. I am grateful to Ignatius Mattingly, Susan Nittrouer, and Michael Studdert-Kennedy for helpful comments on an earlier version.

REFERENCES

[1] Assmann, P. F., & Nearey, T. M. (1987). Perception of front vowels: The role of harmonics in the first formant region. Journal of the Acoustical Society of America, 81, 520-534.

[2] Atal, B. S. (1983). Efficient coding of LPC parameters by temporal decomposition. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (Boston), 81-84.

[3] Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. Journal of Experimental Psychology: Human Perception and Performance, 6, 536-563.

[4] Beddor, P. S. (1984). Formant integration and the perception of nasal vowel height. Haskins Laboratories Status Report on Speech Research, SR-77/78, 107-120.

[5] Beddor, P. S., Krakow, R. A., & Goldstein, L. M. (1986). Perceptual constraints and phonological change: a study of nasal vowel height. Phonology Yearbook, 3, 197-218.

[6] Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception & Psychophysics, 29, 191-211.

[7] Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granström (Eds.), The representation of speech in the peripheral auditory system (pp. 95-102). Amsterdam: Elsevier Biomedical Press.

[8] Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. Journal of the Acoustical Society of America, 67, 648-662.

[9] Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. Journal of the Acoustical Society of America, 79, 481-499.

[10] Bregman, A. S. (1978). The formation of auditory streams. In J. Requin (Ed.), Attention and performance VII (pp. 63-76). Hillsdale, NJ: Erlbaum.

[11] Bregman, A. S. (1987). The meaning of duplex perception: Sounds as transparent objects. In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[12] Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. Canadian Journal of Psychology, 32, 19-31.

[13] Bregman, A. S., & Steiger, H. (1980). Auditory streaming and vertical localization: Interdependence of "what" and "where" decisions in audition. Perception & Psychophysics, 28, 539-546.

[14] Broadbent, D. E., & Ladefoged, P. (1952). On the fusion of sounds reaching different sense organs. Journal of the Acoustical Society of America, 29, 708-710.

[15] Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. Phonology Yearbook, 3, 219-254.

[16] Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 10, 23-36.

[17] Bryan, W. L., & Harter, N. (1899). Studies in the physiology and psychology of the telegraphic language. The acquisition of a hierarchy of habits. Psychological Review, 6, 345-375.

[18] Carlson, R., Fant, G., & Granström, B. (1975). Two-formant models, pitch, and vowel perception. In G. Fant & M. A. A. Tatham (Eds.), Auditory analysis and perception of speech (pp. 55-82). London: Academic Press.

[19] Carlson, R., Granström, B., & Fant, G. (1970). Some studies concerning perception of isolated vowels. Speech Transmission Laboratory Quarterly Progress and Status Report, 2-3, 19-35 (Stockholm: Royal Technical University).

[20] Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. Journal of the Acoustical Society of America, 62, 961-970.

[21] Celmer, R. D., & Bienvenue, G. (1987). Critical bands in the perception of speech by normal and sensorineural hearing loss listeners. in M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[22] Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. Journal of the Acoustical Society of America, 77, 789-805.

[23] Chistovich, L. A., & Lublinskaja, V. V. (1979). The center of gravity effect in vowel spectra and critical distance between the formants. Hearing Research, 1, 185-195.

[24] Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. Canadian Journal of Psychology, 27, 441-449.

[25] Connell, B. A., Hogan, J. T., & Rozsypal, A. J. (1983). Experimental evidence of interaction between tone and intonation in Mandarin Chinese. Journal of Phonetics, 11, 337-351.

[26] Cowan, N. (in press). Auditory sensory storage in relation to the growth of sensation and acoustic information extraction. Journal of Experimental Psychology: Human Perception and Performance.

[27] Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review, 83, 114-140.

[28] Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. Quarterly Journal of Experimental Psychology, 33A, 185-207.

[29] Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. Journal of the Acoustical Society of America, 76, 1636-1647.

[30] Darwin, C. J., & Gardner, R. B. (1985). Which harmonics contribute to the estimation of first formant frequency? Speech Communication, 4, 231-235.

[31] Darwin, C. J., & Seton, J. (1983). Perceptual cues to the onset of voiced excitation in aspirated initial stops. Journal of the Acoustical Society of America, 74, 1126-1135.

[32] Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? The Quarterly Journal of Experimental Psychology, 36A, 193-208.

[33] Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. Word, 8, 195-210.

[34] Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. Journal of the Acoustical Society of America, 68, 843-857.

[35] Delgutte, B., & Kiang, N. Y. S. (1984). Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics. Journal of the Acoustical Society of America, 75, 897-907.

[36] Derr, M. A., & Massaro, D. W. (1980). The contribution of vowel duration, Fo contour, and frication duration as cues to the /juz/-/jus/ distinction. Perception & Psychophysics, 27, 51-59.

[37] Deutsch, D. (1978). Lateralization by frequency for repeating sequences of dichotic 400- and 800-Hz tones. Journal of the Acoustical Society of America, 63, 184-186.

[38] Deutsch, D., & Roll, P. (1976). Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. Journal of Experimental Psychology: Human Perception and Performance, 2, 23-29.

[39] Diehl, R. (1987). Auditory constraints on the perception of speech. In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[40] Diehl, R. L., Kluender, K. R., & Parker, E. M. (1985). Are selective adaptation and contrast effects really distinct? Journal of Experimental Psychology: Human Perception and Performance, 11, 209-220.

[41] Disner, S. F. (1980). Evaluation of vowel normalization procedures. Journal of the Acoustical Society of America, 67, 253-261.

[42] Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. Journal of Experimental Psychology: Human Perception and Performance, 1, 121-129.

[43] Dorman, M. F., Raphael, L. J., & Liberman, A. M. (1979). Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, 65, 1518-1532.

[44] Dreschler, W. A., & Plomp, R. (1980). Relation between psychophysical data and speech perception for hearing-impaired subjects. I. Journal of the Acoustical Society of America, 68, 1608-1615.

[45] Eggermont, J. J. (1985). Peripheral auditory adaptation and fatigue: A model oriented review. Hearing Research, 18, 57-71.

[46] Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In N. J. Lass (Ed.), Speech and language: Advances in basic research and practice. Vol. 10 (pp. 337-373). New York: Academic Press.

[47] Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception & Psychophysics, 27, 343-350.

[48] Flege, J. E. (in press). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), Human communication and its disorders, Vol. 1. Norwood, NJ: Ablex.

[49] Fodor, J. A. 1983). The modularity of mind. Cambridge, MA: MIT Press.

[50] Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. Journal of Phonetics, 8, 113-133.

[51] Fowler, C. A. (1981). Production and perception of coarticulation among stressed

and unstressed vowels. Journal of Speech and Hearing Research, 46, 127-139.

[52] Fowler, C. A. (1983). Realism and unrealism: a reply. Journal of Phonetics, 11, 303-322.

[53] Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. Perception & Psychophysics, 36, 359-368.

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

[55] Fowler, C. A., & Smith, M. R. (1986). Speech perception as "vector analysis": An approach to the problems of invariance and segmentation. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and variability in speech processes (pp. 123-135). Hillsdale, NJ: Erlbaum.

[56] Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. Language and Speech, 26, 21-60.

[57] Fox, R. A. (1984). Effect of lexical status on phonetic categorization. Journal of Experimental Psychology: Human Perception and Performance, 10, 526-540.

[58] Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.

[59] Garner, W. R. (1974). The processing of information and structure. Potomac, MD: Erlbaum.

[60] Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton Mifflin.

[61] Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

[62] Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. Journal of the Acoustical Society of America, 77, 671-677.

[63] Green, D. M. (1983). Profile analysis. A different view of auditory intensity discrimination. American Psychologist, 38, 133-142.

[64] Green, D. M., & Kidd, G., Jr. (1983). Further studies of auditory profile analysis. Journal of the Acoustical Society of America, 73, 1260-1265.

[65] Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. Perception & Psychophysics, 38, 269-276.

[66] Halle, M., Hughes, G. W., & Radley, J.-P. A. (1957). Acoustic properties of stop consonants. Journal of the Acoustical Society of America, 29, 107-116.

[67] Halwes, T. G. (1969). Effects of dichotic fusion on the perception of speech. Haskins Laboratories Status Report on Speech Research (Supplement).

[68] Hammarberg, R. (1982). On redefining coarticulation. Journal of Phonetics, 10, 123-137.

[69] Jaeger, J. J. (1980). Testing the psychological reality of phonemes. Language and Speech, 23, 233-253.

[70] Jenkins, J. J. (1985). Acoustic information for objects, places, and events. In W. H. Warren & R. E. Shaw (Eds.), Persistence and change. Proceedings of the First International Conference on Event Perception (pp. 115-138). Hillsdale, NJ: Erlbaum.

[71] Jongman, A., Blumstein, S. E., & Lahiri, A. (1985). Acoustic properties for dental and alveolar stop consonants: a cross-language study. Journal of Phonetics, 13, 235-251.

[71a] Kawasaki, H. (1986). Phonetic explanation of phonological universals: The case of distinctive vowel nasalization. In J. J. Ohala & Jaeger, J. J. (Eds.), Experimental phonology (pp.81-104). New York: Academic Press.

[72] Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. Journal of Phonetics, 14, 29-59.

[73] Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.

[74] Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. Journal of the Acoustical Society of America, 73, 1779-1793.

[75] Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 7, 279-312.

[76] Kohler, K. J. (1979). Dimensions in the perception of fortis and lenis plosives. Phonetica, 36, 332-343.

[77] Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In M. Kubovy & J. R. Pomerantz (Eds.), Perceptual organization (pp. 55-98). Hillsdale, NJ: Erlbaum.

[78] Kunisaki, O., & Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. Annual Bulletin of the Research Institute for Logopedics and Phoniatrics, 11, 85-91 (University of Tokyo).

[79] Kurowski, K., & Blumstein, S. E. (in press). Acoustic properties for place of articulation in nasal consonants. Journal of the Acoustical Society of America.

[80] Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29, 98-104.

[81] Lahiri, A., Gewirth, L., & Blumstein, S. E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. Journal of the Acoustical Society of America, 76, 391-404.

[82] Leather, J. (1983). Speaker normalization in perception of lexical tone. Journal of Phonetics, 11, 373-382.

[83] Lehiste, I. (1972). The units of speech perception. In J. H. Gilbert (Ed.), Speech and cortical functioning (pp. 187-236). New York: Academic Press.

[84] Liberman, A. M. (1979). Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), Proceedings of the Ninth International Congress of Phonetic Sciences (pp. 468-473). Copenhagen: Institute of Phonetics, University of Copenhagen.

[85] Liberman, A. M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.

[86] Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.

[87] Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. Language, 33, 42-49.

[88] Lisker, L. (1978). On buzzing the English /b/. Haskins Laboratories Status Report on Speech Research, SR-55/56, 181-188.

[89] MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. Perception & Psychophysics, 24, 253-257.

[90] Mann, V. A. (1986). Phonological awareness: The role of reading experience. Cognition, 24, 65-92.

[91] Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. Perception & Psychophysics, 28, 213-228.

[92] Marcus, S. M., & Atal, B. S. (1986). Decoding the spech code--applications of temporal decomposition. Journal of the Acoustical Society of America, 80 (Suppl. 1), S17.

[93] Marcus, S. M., & Van Lieshout, R. A. J. M. (1984). Temporal decomposition of speech. IPO Annual Progress Report, 25-31 (Nijmegen, The Netherlands).

[94] Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. Speech Communication, 4, 55-73.

[95] Massaro, D. W. (1975). Preperceptual images, processing time, and perceptual units in speech perception. In D. W. Massaro (Ed.), Understanding language. An information-processing analysis of speech perception, reading, and psycholinguistics (pp. 125-150). New York: Academic Press.

[96] Massaro, D. W. (in press). Speech perception by ear and eye: A paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

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

[98] 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.

[99] Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by ear and by eye. The relationships between speech and reading (pp. 133-148). Cambridge, MA: MIT Press.

[100] McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America, 77, 678-685.

[101] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

[102] McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. Journal of Verbal Learning and Verbal Behavior, 12, 419-430.

[103] Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, NJ: Erlbaum.

[104] Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 25, 457-465.

[105] Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidth and excitation patterns. Journal of the Acoustical Society of America, 74, 750-753.

[106] Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.

[107] Mushnikov, V. N., & Chistovich, L. A. (1973). Experimental testing of the band hypothesis of vowel perception. Soviet Physics-Acoustics, 19, 250-254.

[108] Obrecht, D. H. (1965). Three experiments in the perception of geminate consonants in Arabic. Language and Speech, 8, 31-41.

[109] Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. Perception & Psychophysics, 39, 129-142.

[110] Peters, R. W., Moore, B. C. J., & Glasberg, B. R. (1983). Pitch of components of complex tones. Journal of the Acoustical Society of America, 73, 924-929.

[111] Pickett, J. M., & Decker, L. R. (1960). Time factors in perception of a double consonant. Language and Speech, 3, 11-17.

[112] Pisoni, D. B. (1987). Auditory perception of complex sounds: Some comparisons of speech vs. non-speech signals. In W. A. Yost and C. S. Watson (Eds.), Auditory processing of complex sounds (pp. 247-256). Hillsdale, NJ: Erlbaum.

[113] Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Perception & Psychophysics, 34, 314-322.

[114] Plomp, R. (1964). The ear as a frequency analyzer. Journal of the Acoustical Society of America, 36, 1628-1636.

[115] Pols, L. C. W., van der Kamp, L. J. Th., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46, 458-467.

[116] Port, R. F. (1979). The influence of tempo

on stop closure duration as a cue for voicing and place. Journal of Phonetics, 7, 45-56.

[117] Price, P. J., & Levitt, A. G. (1983). The relative roles of syntax and prosody in the perception of the /š/-/č/ distinction. Language and Speech, 26, 291-304.

[118] Rakerd, B., Dechovitz, D. R., & Verbrugge, R. R. (1982). An effect of sentence finality on the phonetic significance of silence. Language and Speech, 25, 267-282.

[119] Rakerd, B., & Verbrugge, R. R. (1985). Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. Journal of the Acoustical Society of America, 77, 296-301.

[120] Rand, T. C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.

[121] Repp, B. H. (1976a). Effects of fundamental frequency contrast on discrimination and identification of dichotic CV syllables at various temporal delays. Memory & Cognition, 4, 75-90.

[122] Repp, B. H. (1976b). Identification of dichotic fusions. Journal of the Acoustical Society of America, 60, 456-469.

[123] Repp, B. H. (1978a). "Coperception": Influence of vocalic context on same-different judgments about intervocalic stop consonants. Unpublished manuscript (available from the author).

[124] Repp, B. H. (1978b). Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Perception & Psychophysics, 24, 471-485.

[125] Repp, B. H. (1979a). Influence of vocalic environment on perception of silence in speech. Haskins Laboratories Status Report on Speech Research, SR-57, 267-290.

[126] Repp, B. H. (1979b). Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 22, 173-189.

[127] Repp, B. H. (1980). Accessing phonetic information during perceptual integration of temporally distributed cues. Journal of Phonetics, 8, 185-194.

[128] Repp, B. H. (1981). Two strategies in fricative discrimination. Perception & Psychophysics, 30, 217-227.

[129] Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

[130] Repp, B. H. (1983a). Bidirectional contrast effects in the perception of VC-CV sequences. Perception & Psychophysics, 33, 147-155.

[131] Repp, B. H. (1983b). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Communication, 2, 341-362.

[132] Repp, B. H. (1984a). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), Speech and language: Advances in basic research and practice. Vol. 10 (pp. 243-335). New York: Academic Press.

[133] Repp. B. H. (1984b). Closure duration an release burst amplitude cues to stop consonan manner and place of articulation. Languag and Speech, 27, 245-254.

[134] Repp, B. H. (1984c). The role of releas bursts in the perception of [s]-stop clusters Journal of the Acoustical Society of America 75, 1219-1230.

[135] Repp, B. H. (1985a). Can linguisti boundaries change the effectiveness of silenc as a phonetic cue? Journal of Phonetics, 13 421-431.

[136] Repp, B. H. (1985b). Perceptual coherence of speech: Stability of silence-cued sto consonants. Journal of Experimenta Psychology: Human Perception and Performance, 11, 799-813.

[137] Repp, B. H. (1987a). On the possible role of auditory short-term adaptation in perceptio of the prevocalic [m]-[n] contrast. Manuscript submitted for publication.

[138] Repp, B. H. (1987b). The role of psychophysics in understanding speech perception. In M.E.H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[139] Repp, B. H., & Bentin, S. (1984). Parameters of spectral/temporal fusion in speech perception. Perception & Psychophysics, 36, 523-530.

[140] Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. N. Harnad (Ed.), Categorical perception. New York: Cambridge University Press.

[141] Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 4, 621-637.

[142] Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. Perception & Psychophysics, 33, 333-337.

[143] Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. Nature, 291, 150-152.

[144] Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. Journal of Phonetics, 14, 283-302.

[145] Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. Journal of Experimental Psychology: General, 110, 474-494.

[146] Savin, H. B., & Bever, T. G. (1970). The nonperceptual reality of the phoneme. Journal of Verbal Learning and Verbal Behavior, 9, 295-302.

[147] Sawusch, J. R., & Jusczyk, P. (1981). Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance, 7, 408-421.

[148] Sawusch, J. R., & Nusbaum, H. C. (1983). Auditory and phonetic processes in place perception for stops. Perception & Psychophysics, 34, 560-568.

[149] Scheffers, M. T. M. (1983). Sifting vowels. Auditory pitch analysis and sound segregation. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

[150] Schubert, E. D. (1982). On hearing your own performance. In V. L. Lawrence (Ed.), Transcripts of the Eleventh Symposium Care of the Professional Voice (pp. 161-185). New York: The Voice Foundation.

[151] Schwartz, J. L., & Escudier, P. (1987). Does the human auditory system include large scale spectral integration? In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[152] Stelmachowicz, P. G., Jesteadt, W., Gorga, M. P., & Mott, J. (1985). Speech perception ability and psychophysical tuning curves in hearing-impaired listeners. Journal of the Acoustical Society of America, 77, 620-627.

[153] Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1368.

[154] Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (eds.), Perspectives in the study of speech (pp. 1-38). Hillsdale, NJ: Erlbaum.

[155] Studdert-Kennedy, M. (1985). Perceiving phonetic events. In W. H. Warren, Jr., & R. E. Shaw (Eds.), Persistence and change: Proceedings of the First International Conference on Event Perception (pp. 139-156). Hillsdale, NJ: Erlbaum.

[156] Summerfield, Q. (1979). Use of visual information for phonetic perception. Phonetica, 36, 314-331.

[157] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 7, 1074-1095.

[158] Summerfield, Q. (1983). Audio-visual speech perception, lipreading, and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), Hearing science and hearing disorders (pp. 131-182). London: Academic Press.

[159] Summerfield, Q. (in press). Preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), Hearing by eye. Hillsdale, NJ: Erlbaum.

[160] Summerfield, Q., & Assmann, P. (1987). Auditory enhancement and speech perception. In M.E.H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[161] Summerfield, Q., & Haggard, M. P. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. A. A. Tatham (Eds.), Auditory analysis and perception of speech (pp. 115-142). London: Academic Press.

[162] Summerfield, Q., Haggard, M., Foster, J., & Gray. S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perception & Psychophysics, 35, 203-213.

[163] Suomi, K. (1984). On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem. Speech Communication, 3, 199-209.

[164] Suomi, K. (1985). The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables. Journal of Phonetics, 13, 267-285.

[165] Swinney, D. A. (1982). The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E. C. T. Walker, & M. Garrett (Eds.), Perspectives on mental representation. Experimental and theoretical studies of cognitive processes and capacities (pp. 151-167). Hillsdale, NJ: Erlbaum.

[166] Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America, 79, 1086-1100.

[167] Tartter, V. C., & Blumstein, S. E. (1981). The effects of pitch and spectral differences on phonetic fusion in dichotic listening. Journal of Phonetics, 9, 251-259.

[168] Tartter, V. C., Kat, D., Samuel, A. G., & Repp, B. H. (1983). Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions. Journal of the Acoustical Society of America, 74, 715-725.

[169] Thomas, I. B., Hill, P. B., Carrol, F. S., & Garcia, D. (1970). Temporal order in the perception of vowels. Journal of the Acoustical Society of America, 48, 1010-1013.

[170] Tillmann, H. G., Pompino-Marschall, B., & Porzig, U. (1984). Zum Einfluss visuell dargebotener Sprechbewegungen auf die Wahrnehmung der akustisch kodierten Artikulation. Forschungsbericht des Instituts fur Phonetik und Sprachliche Kommunikation der Universität München, 19, 318-336.

[171] Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. Journal of the Acoustical Society of America, 81, 755-764.

[172] Traunmüller, H. (1982). Perception of timbre: Evidence for spectral resolution bandwidth different from critical band? In R. Carlson & B. Granström (Eds.), The representation of speech in the peripheral auditory system (pp. 103-108). Amsterdam: Elsevier Biomedical Press.

[173] Traunmüller, H. (1984a). Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels. Speech Communication, 3, 49-61.

[174] Traunmüller, H. (1984b). Die spektrale Auflösung bei der Wahrnehmung der Klangfarbe von Vokalen. Acustica, 54,237-246.

[175] Traunmüller, H. (1987). Some aspects of the sound of speech sounds. In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[176] Treisman, A. M. (1969). Strategies and models of selective attention. Psychological Review, 76, 282-299.

[177] Verbrugge, R. R., & Rakerd, B. (1986). Evidence for talker-independent information for vowels. Language and Speech, 29, 39-57.

[178] Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? Journal of the Acoustical Society of America, 60, 198-212.

[179] Warren, W. H., Jr., & Shaw, R. E. (1985). Persistence and change: Proceedings of the First International Conference on Event Perception. Hillsdale, NJ: Erlbaum.

[180] Warren, W. H., Jr., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. Journal of Experimental Psychology: Human Perception and Performance, 10, 704-712.

[181] Warren, R. M. (1970). Perceptual restoration of missing speech sounds. Science, 167, 392-393.

[182] Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. Scientific American, 233, 30-36.

[183] Watson, C. S., & Foyle, D. C. (1985). Central factors in the discrimination and identification of complex sounds. Journal of the Acoustical Society of America, 78, 375-380.

[184] Weintraub, M. (1987). Sound separation and auditory perceptual organization. In M. E. H. Schouten (Ed.), The psychophysics of speech perception. The Hague: Martinus Nijhoff.

[185] Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. Journal of the Acoustical Society of America, 69, 275-282.

[186] Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. Perception & Psychophysics, 35, 49-64.

[187] Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. Manuscript submitted for publication.

[188] Whalen, D. H., & Samuel, A. G. (1985). Phonetic information is integrated across intervening nonlinguistic sounds. Perception & Psychophysics, 37, 579-587.

[189] Wright, H. N. (1964). Temporal summation and backward masking. Journal of the Acoustical Society of America, 36, 927-932.

[190] Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. Journal of the Acoustical Society of America, 68, 1523-1525.

[191] Zwicker, U. T. (1984). Auditory recognition of diotic and dichotic vowel pairs. Speech Communication, 3, 265-277.

[192] Zwislocki, J. J. (1969). Temporal summation of loudness: An analysis. Journal of the Acoustical Society of America, 46, 431-440.

PI 2.2.20

VOWEL-RELATED LINGUAL ARTICULATION IN /∂CVC/ SYLLABLES AS A FUNCTION OF STOP CONTRAST

Peter J. Alfonso
Univ. of Connecticut and Haskins Laboratories
Storrs (06268) and New Haven (06511) CT, U. S. A.

Satoshi Horiguchi
Univ. of Tokyo
Tokyo 113, Japan

ABSTRACT

Lateral cineradiographic pellet-tracking of tongue blade and tongue body movements along with formant frequency trajectories show that articulation of the vowels /i/ and /u/ in /∂CVC/ syllables vary by as much as 8-10 mm as a function of the stop consonant environment in which they are produced. The magnitude of the variation is related to the identity of the stop and vowel.

INTRODUCTION

This study represents one of a set of experiments completed or underway at Haskins Laboratories that aim to study the dynamics of vowel articulation. What distinguishes the set of studies from previous work is that the dynamics of vowel articulation is studied by examining data representing the four accessible measurement levels of speech production, namely 1)muscle activity (by hooked-wire electromyography), 2)corresponding movements of the speech structures (primarily by tracking the movements of lead pellets glued to the lips, tongue, and jaw by lateral cineradiography or x-ray microbeam), 3)representative speech acoustic signals, and 4)perceptual testing of selected auditory segments to determine whether or not the underlying articulatory movements provide relevant linguistic information. In the first experiment to use these measurement techniques, a single subject's productions of disyllables of the form /∂pVp/, where V represented one of eleven vowels, were analyzed. As an example of the conclusion that can be drawn from multi-level analysis, the results showed that vowel-related tongue horizontal and vertical movements can have different time constraints in labial environments. When fronting and raising occur together they are necessarily time-locked since they are caused primarily by the same muscle, genioglossus. Backing and raising, on the other hand, can and do occur independently, since they are caused by different muscle groups. While vertical movements for all vowels and horizontal movements for front vowels always began about the moment of implosion for the initial stop, horizontal movements for back vowels began much earlier, even before the acoustic onset of the schwa. Acoustic and perceptual analysis indicated that anticipatory tongue movements were linguistically significant since listeners were able to identify the vowels when presented with only a portion of the schwa segment.

More recently, a second subject has been run following the same procedures but increasing the data base in two ways: 1) the same set of vowels were produced in labial, alveolar, and velar stop environments, and 2) more extensive tongue EMG insertions representing the complete set of lingual extrinsic and accessory muscles. Analysis of EMG data from the second subject in general supports the temporal differentiation in vertical-horizontal tongue movements in the first subject in that the muscles of the tongue appear to be distinctly organized for front versus back vowel production [2]. Biomechanical descriptions of tongue dynamics, such as the relationship between genioglossus and fronting-raising in the first subject, will be enhanced significantly by mapping the EMG data for the second subject onto his x-ray data. While the overall purpose of the combined EMG and x-ray runs is to study the dynamics of vowel articulation in a fashion similar to the initial experiment [1], the paper presented here will focus on the x-ray run. The purpose of the paper is to give a quantitative description at the movement and acoustic levels of the variation in vowel-related tongue articulation that occur as a function of producing the vowels /i/ and /u/ in labial, alveolar, and velar stop environments.

METHODS

An adult male native speaker of American English with a New York City dialect produced two repetitions of /∂CVC/ disyllables where /C/ represents /p/, /t/, or /k/ and /V/ represents one of eleven vowels. The initial and final consonants in each utterance were identical. Lateral cineradiographic films were made at a rate of 60 frames per second while the subject produced isolated syllables at a rate of about one every two seconds. Figure 1 represents a schematic diagram of the midsagittal plane of the vocal tract sketched from a single frame of the x-ray film. The figure shows the location of lead pellets glued to the tongue blade, middle and rear dorsal areas of the tongue surface at the midline, and a jaw pellet attached between the lower central incisors. Measurements of pellet movements were made on a frame-by-frame basis with the aid of

*This work was supported by NIH Grant NS-13617 to Haskins Laboratories. The authors gratefully acknowledge the assistance of Drs. Thomas Baer and Hiroshi Muta in data collection and analysis.

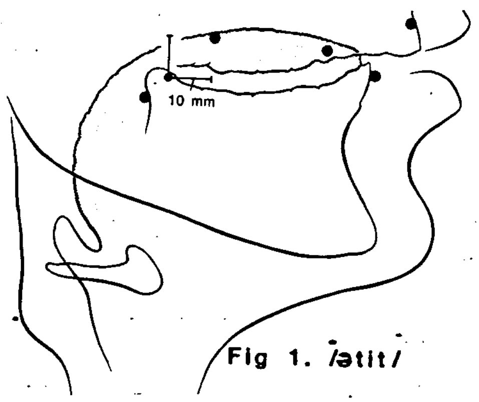


Fig 1. /ətit/

a digitizing tablet. Pellet locations were fixed with respect to two reference positions, the lead pellet attached between the upper central incisors and the point marked by a template reflecting maxillary boundaries. The latter reference location is shown in Figure 1 as the point of origin in the 10 millimeter (mm) grid. Pellet displacement data shown in the following section are displayed in reference to the point of origin. For example, positive vertical and positive horizontal displacement trajectories indicate movements above and to the right of the origin, respectively. Displacement values represent calibrated units in mm.

RESULTS

Considering, first, the dynamics associated with the articulation of /i/, Figure 2 shows vertical movement trajectories for the tongue blade (top panel), middle (mid panel) and rear dorsal (bottom panel) pellets for production of /∂CiC/. Within each panel, the solid line represents trajectories for /∂pip/, the dashed line represents trajectories for /∂tit/, and the dotted line represents trajectories for /∂kik/. Except for infrequent instances when pellet locations were not visible and therefore not tracked, each of the trajectories represent the average displacement of two tokens per syllable. Although trajectories represent combined jaw and tongue displacements, they primarily represent tongue displacements since jaw movements were negligible for this speaker (see Table 1). The ordinate represents calibrated units in mm from the origin (see Figures 1 and 4). The distance between vertical markers along the abcissa represents 100 ms intervals. The solid vertical line at zero time represents the initial consonant release. Thus, the time-span of the trajectories is 800 ms, segmented as a 300 ms and 500 ms interval before and after initial consonant release. The acoustic onset of the schwa and the acoustic duration of the vowel varied with the identity of the stop but, in general, schwa onset occurred from about 150 to 200 ms before consonant release and vowel duration varied from 200 to 250 ms. Implosion of the initial consonant varied from 75 to 125 ms before consonant release. Not unexpectedly, the top panel of Figure 2 shows greater tongue blade vertical displacement during the consonantal segment in the alveolar environment compared to the labial and velar environments. Notice, however, that tongue blade trajectories during the vocalic segment appear similar in all stop consonant environments. On the other hand, the relative vertical displacement trajectories for middle and rear dorsal pellets are clearly different from tongue blade trajectories. The middle and bottom panels of Figure 2 show that the corresponding tongue locations were much higher in the velar environment than in the labial and alveolar environments throughout the syllable. That is, the vertical differentiation begins very early, at about initial consonant closure for the rear pellet and at the earliest point of measurement for the middle pellet, and continues throughout the vowel and final consonant.

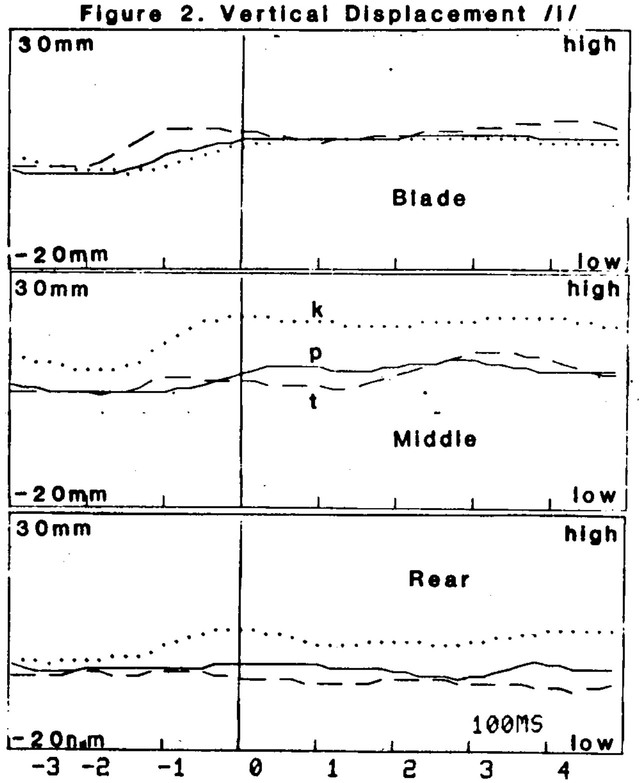
Figure 3 shows horizontal displacement trajectories for the same three pellets. Notice that fronting is clearly differentiated during the vocalic portion of the syllables and that the velar context produces the most front /i/.

Table I lists jaw and tongue displacement measurements in mm from the origin taken from trajectories shown in Figures 2 and 3 at the moment of vocalic peak acoustic amplitude. Acoustic peak amplitude occurred an average of 125 ms after consonant release for the two tokens of /∂pip/ and /∂tit/ and 175 ms after release for /∂kik/. Also shown are average first and second formant values in Hz. The formant values represent the average of five samples taken at five ms intervals beginning 10 ms before through 10 ms after the moment of peak amplitude. Large differences in tongue shape across the three stop environments during /i/ production are clearly indicated in Table I. For example, the middle dorsal area of the tongue is displaced about 10 mm anteriorly and superiorly in the velar stop environment relative to the labial and alveolar environments. The large effect of velar closure on tongue body positioning for /i/ is reflected in the formant frequency values as well; the lowest first formant and the highest second formant values occur for /∂kik/.

Finally, the influence of alveolar and velar stop production on the entire tongue shape during /i/ production is shown in Figures 1 and 4. The figures are schematic diagrams of the midsagittal plane of the vocal tract sketched from x-ray frames corresponding to the temporal interval represented in Table 1. Vocal tract shape at peak vowel amplitude for /i/ during /∂tit/ is shown in Figure 1, and the corresponding interval during /∂kik/ is shown in Figure 4. A comparison of the figures shows that alveolar and velar stop consonant constrictions produce dramatic differences in the tongue shape for /i/ articulation. Taken together, the data indicate that the dynamics of tongue articulation for /i/ is clearly different in each of the three stop contexts, that the velar context yields the most high and front vowel articulation.

Similar analyses have been made for /∂CuC/ syllables, but space limitations rule that they be presented in abbreviated form. For example, Figure 5 shows vertical displacement trajectories of the middle dorsal pellet for /∂CuC/. Note that the velar context produces the highest tongue body movements, and that the vertical differentiation begins before the acoustic onset of the schwa.

Figure 2. Vertical Displacement /i/



A comparison of Figures 2 and 5 shows that the vertical differentiation in vowel articulation is less dramatic in /u/ than in /i/. Figure 6 demonstrates that the alveolar context yields greater tongue blade raising than the labial and velar contexts for /u/ and that the vertical differentiation begins at about acoustic onset of the schwa. Furthermore, a comparison of Figures 2 and 6 shows greater tongue blade vertical differentiation in /u/ than /i/. The tongue blade is lower in labial and velar contexts in /∂CuC/ relative to /∂CiC/.

Horizontal displacement trajectories yield similar trends. First, the velar stop constriction has less influence on vowel related vertical displacement for /u/ than for /i/ and, second, the relative maximum raising usually cooccurs with relative maximum fronting. For example, the middle dorsal pellet in /∂kuk/ is more front and high than in /∂pup/ and /∂tut/. Finally, Table I also shows displacement and formant frequency values for the vocalic segment in /∂CuC/ syllables. The vocalic peak amplitude occurred 140 ms after release for /∂pup/ and 200 ms after release for /∂tut/ and /∂kuk/. Once again, appreciable differences in vowel related tongue displacement as a function of stop constriction location are observed. Although the center frequency of the first and second formant also showed large variation with stop context, the expected relationship between formant frequency and tongue vertical and horizontal displacement is not observed, presumably due to the stop consonant influence on pharyngeal cavity width and the movements of other structures, most likely the lips or larynx.

Figure 5. Vertical displacement /u/

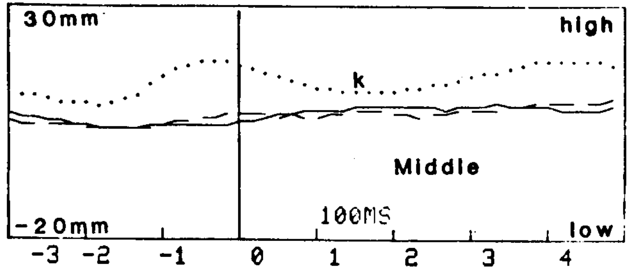
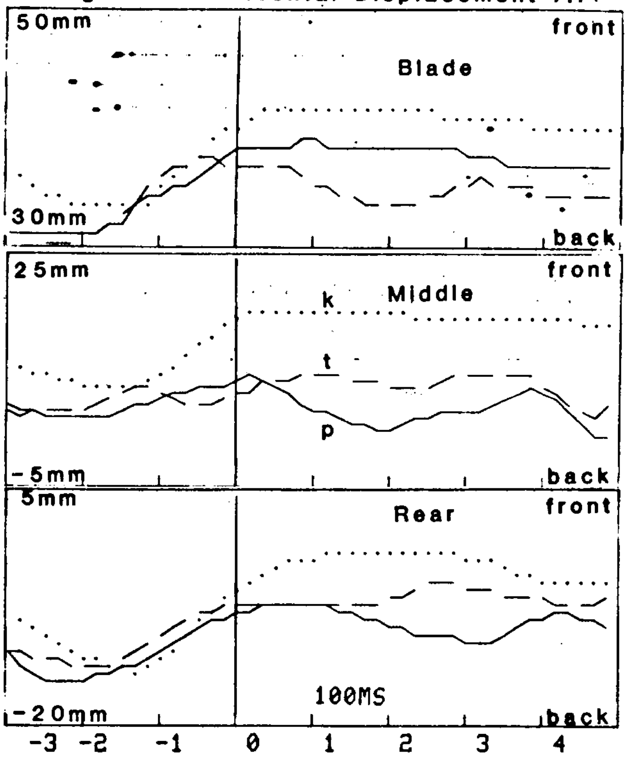


Figure 3. Horizontal Displacement /i/.



DISCUSSION

The results show that for this subject the dynamics of vowel articulation for /i/ and /u/ can be altered significantly as a function of the identity of the stop in which they are produced. Assuming that the middle and rear dorsal pellets yield appropriate estimates of tongue body shape, Table I shows that tongue body configurations for /i/ vary by nearly 10 mm, with the velar context producing the most high and front /i/. For /u/, variation in tongue body configuration is about 6 mm, and again the velar context produces the most extreme lingual displacement.

Figure 6. Vertical Displacement /u/

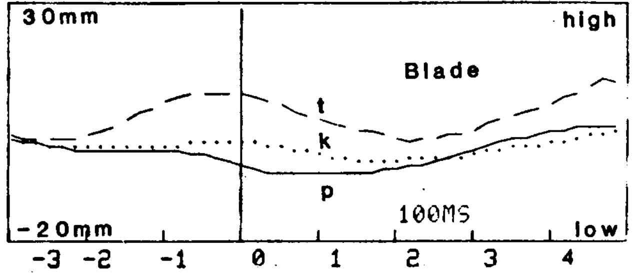


TABLE I.
Average pellet displacement in mm from origin and average first and second formant values in Hz for the vocalic segment of /əCiC/ and /əCuC/.
See text for definition of measurement techniques.

	Vertical Displacement				Horizontal Displacement				Formant	
	Jaw	Blade	Middle	Rear	Jaw	Blade	Middle	Rear	F1	F2
əpip	0	7.2	9.6	-2.4	45.6	38.4	9.6	-8.0	310	1799
ətit	1.6	8.0	9.6	-4.8	46.4	33.6	8.0	-6.4	312	1722
əκik	0.8	7.2	19.2	3.2	48.8	41.6	17.6	-4.8	301	1851
əpup	0.8	-5.6	8.8	-0.8	46.4	28.0	4.0	-16.0	382	941
ətut	1.6	0.8	7.2	-1.6	46.4	30.4	4.8	-11.2	351	1194
əkuk	0.8	-2.4	12.8	0.8	48.0	30.4	9.6	-12.0	389	1007

Tongue blade configurations also vary greatly as a function of the stop. Whereas the magnitude of fronting and raising generally covary in tongue body articulation, they do not seem to covary for tongue blade movements. The variation for tongue blade configurations is about 6 mm in the vertical dimension, the differentiation being larger in /u/ than in /i/ with the alveolar context producing the greatest displacement. In the horizontal dimension, tongue blade displacement varies by as much as 8 mm, the velar context producing the most front tongue blade displacement.

Comparisons between movement data reported here with previously published data are not straight-forward since either the method used to measure tongue displacement or the design of the experiments differ significantly. For example, the patterns reported here can be compared with corresponding data taken from a cinefluorographic experiment that similarly measured the effect of the stop consonant on vowel-related vertical displacement [3]. The two experiments are in agreement in that vowels in velar context are produced with greater vertical displacement relative to labial and alveolar contexts. However, the published displays of the cinefluorographic trajectories appear to indicate that the magnitude of the variation during the vocalic segment is no greater than 5 mm. For a number of reasons, a more appropriate comparison of the magnitude of the variation can be made between experiments that employ pellet-tracking. X-ray microbeam pellet-tracking was used to estimate the variation in a large number of repetitions of the vowels /i/, /a/, and /ae/ occurring in a wide variety of consonantal environments [4,5]. Though not specified in the text, the figures indicate that the variation approximates the 8-10 mm maximum variation reported here. Finally, there are many acoustic based studies on the topic (e.g. [6,7]). The first and second format frequency values reported here are in good agreement with these data.

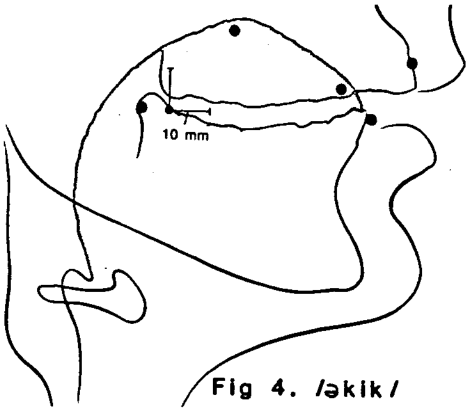


Fig 4. /əkik/

REFERENCES

[1].P.J. Alfonso and T. Baer. Dynamics of vowel articulation. *Language and Speech*, 25:151-173, 1982.

[2].P.J. Alfonso, K. Honda, and T. Baer. Coordinated tongue muscle activity during /∂pVp/ utterances. *Proceedings of the Tenth International Congress of Phonetic Sciences*, IIB:390-394, 1984.

[3].P.F. MacNeilage and J.L. DeClark. On the motor control of coarticulation in CVC monosyllables. *J. Acoust. Soc. Am.*, 45:1217-1233, 1969.

[4].J.S. Perkell and W.L. Nelson. Articulatory targets and speech motor control: A study of vowel production. *Speech Motor Control*, Pergamon, New York, 1982.

[5].J.S. Perkell and W.L. Nelson. Variability in production of the vowels /i/ and /a/. *J. Acoust. Soc. Am.*, 77: 1889-1895, 1985.

[6].K.N. Stevens and A.S. House. Perturbation of vowel articulation by consonantal context: An acoustical study. *J. Speech Hearing Res.*, 6:111-128, 1963.

[7].K.N. Stevens, A.S. House, and A.P. Paul. Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. *J. Acoust. Soc. Am.*, 40: 123-132, 1966.

SOME PROPERTIES OF AN AEROACOUSTICS CHARACTERIZATION OF PHONATION

Richard S. McGowan
Haskins Laboratories
270 Crown Street
New Haven, CT 06511, U. S. A.

ABSTRACT

When the conservation equations governing the motion of air are considered, it is seen that there are sources of sound during phonation other than the volume velocity at the glottis. One possible source is the result of forcing between the solid surfaces near the glottis and the air. Approximations based on a schematic vocal tract show this source to be relatively weak compared with the volume velocity source.

Sources of sound during phonation will be discussed in this paper. The plural is used because there are other acoustic sources than the volume velocity source during phonation. These other sources appear when the three-dimensional character of the fluid velocity field in the vocal tract is considered, showing that a one- dimensional characterization is not enough when considering vocal tract sound production. While one-dimensional models using volume velocity and pressure for dependent variables may be good approximations when considering plane wave propagation of sound, such models do not correctly characterize the production of sound during phonation. These one-dimensional models, or analogs, are based on what is known as a scalar field theory because the dependent variables are scalar.

Both the sound production and sound propagation parts of the one-dimensional analogs should be derivable from the equations of motion for air. The equations of motion of air provide what is known as a vector field theory, because the vector, fluid particle velocity is used instead of the scalar, volume velocity. Here, the inadequacy of the one-dimensional scalar description of the sound source will be discussed by going back to the more primitive notions of the fluid mechanics of air: mass, momentum, and energy conservation. From this point of view, it will be seen that volume velocity supplies only part of the picture in sound production.

Theoretical aeroacoustics is a branch of fluid mechanics providing a means of describing the sources of sound from the conservation equations. The literature in this field is extensive (e.g.[1]) and some of it will be used here. Initially, we will concentrate on the basic differences between the aeroacoustic view and that provided by the current analogs. The real vocal tract will be schematized to illustrate some of the basic principles of the aeroacoustic view. As a result, the sources of sound named here may have to be modified when applied to a real vocal tract.

Because of the extra degrees of freedom provided by the vector nature of fluid mechanics, we can decompose the fluid velocity field, $\underline{u}$, into two separate vector fields, the solenoidal field and the irrotational field [2].

$$\underline{u} = \underline{u}_S + \underline{u}_I \tag{1}$$

The solenoidal field, $\underline{u}_S$, will support rotational motion but not compression and expansion, while the irrotational field, $\underline{u}_I$, can support compression and expansion, but not rotation. Because acoustic motion needs the potential energy of compression and expansion, the irrotational field is a necessary part of such motion. However, the solenoidal field can provide sources of sound, since solenoidal, incompressible fluid motion can provide pressure fluctuations. This field is ignored by the one-dimensional analogs, where all the motion is irrotational, and therefore the one-dimensional analogs ignore possible acoustic sources. Even after the decomposition of the fluid velocity field given by equation (1) has been performed, the three- dimensional character of the solenoidal field also needs to be considered. The full three-dimensional character of the solenoidal field itself needs to be considered because

*This work was supported by NIH Grants NS-13870 and NS-13617 to Haskins Laboratories.

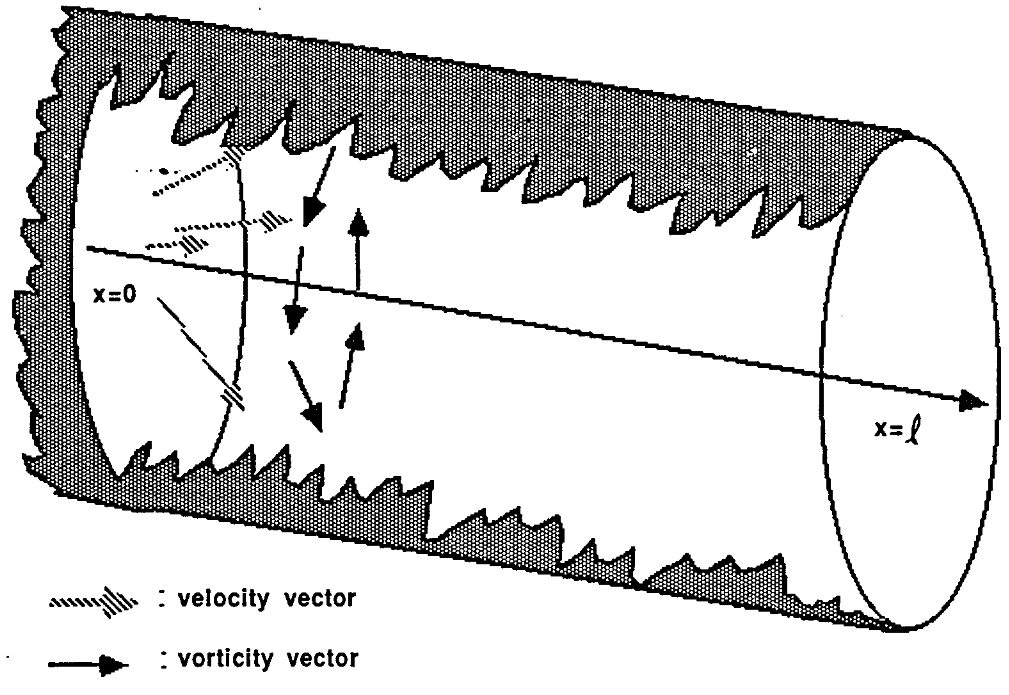


Figure 1: schematic vocal tract

it will be seen that interaction between vectors, derived from this field, provides one example of a sound source. This sound source involves rotational motion of the air.

The sound source involving rotational motion of the air can be derived by considering a particular geometric feature of the vocal tract near the glottis. This feature is the abrupt area change from the glottis into the vocal tract. We will consider the vocal tract to be a straight, semi-infinite, cylindrical tube with an abrupt area change representing the glottis at x = 0 and a mouth at x = l (see Figure 1). The source of sound occurring in such a geometrical configuration will be considered with the understanding that the source could be altered when more features of the geometry of the real vocal tract are considered. To provide a detailed account of sound production would be premature and beyond the scope of this paper. It is well known that there is rotational motion which results when there is flow through a sudden change in area. Rotational motion of the fluid can be inferred from a large drop in pressure head, or stagnation pressure, p_{st}:

$$p_{st} = p + \frac{\rho_0}{2}|\underline{u}|^2 \qquad (2).$$

where p = static pressure and ρ_0 = rest density of air. Pressure head losses have been observed in experiments using static configurations (e.g.[3]). Although investigators have attributed a large portion of the pressure head loss they observed to the formation of turbulence, which is both random and rotational motion of the fluid, it is the rotational motion of the fluid which is the essential factor. In the current analog models this pressure head loss is modeled as a nonlinear resistor without acoustic consequences, other than determining the relationship between transglottal pressure and the volume velocity. Using the theory of aeroacoustics, this head loss is seen as a singularity having consequences in the acoustic far-field.

Note that pressure head loss is known to occur on both sides of model glottis', but the loss above the glottis will have the greatest acoustic consequences in the supraglottal region, so this will be examined more closely. Also, the experiments have been on static configurations, where the results are applied to dynamic phonation with the quasi-steady approximation. (In the dynamic situation it is less likely that the rotational fluid motion just above the glottis is also random.) While the quasi-steady approximation allows for easy estimation of pressure head loss [4], it is not a necessary approximation for asserting the existence of such a head loss in a dynamic situation. It has been shown experimentally that the head loss also occurs in the dynamic situation [5].

Before the acoustic consequences of the head loss are discussed, the cause of the loss from the point of view of the equations of motion need to be discussed. The cause has to do with the formation of vorticity. Vorticity, $\underline{\omega}$, is a function of the velocity vector itself, and provides a measure of the rotational motion of the fluid.

$$\underline{\omega} = \nabla \times \underline{u} \qquad (3)$$

We can imagine an oscillating jet of air exiting from the glottis, and, eventually, expanding into the cylinder. By the definition of the curl and the symmetry of the configuration, the vorticity can be expected to be directed azimuthally about the axis of the cylinder. (There is assumed to be no azimuthal component of velocity, and there is no dependence of the other components on the azimuthal angle.) In fact, the secondary flow just above the glottis may be in the form of a vortex ring.

As evidenced by the loss of pressure head above the glottis, there is a forcing between the solid surfaces and the air. This force is realized in the fluid as what can be called the vorticity-velocity interaction force (per unit volume):

$$\rho_0(\underline{\omega} \times \underline{u}).$$

The radial components for these vectors tend to cancel one another, while the axial components are directed away from the glottis. The net result (in a spatial average) is a vector directed away from the glottis in the axial direction. Because the velocity has a time averaged component, as well as a fundamental and harmonics, the vorticity-velocity interaction force will contain the same frequencies in its spectrum as the velocity, although with different weights. That the loss of pressure head is the result of this forcing can be derived from the momentum conservation equation for an inviscid fluid, Euler's equation, under the quasi-steady approximation [6]. The negative of the gradient of the pressure head is equal to the vorticity-velocity interaction vector:

$$\nabla p_{st} = -\rho_0(\underline{\omega} \times \underline{u}) \qquad (4).$$

From the above discussion, this gradient is directed against the axial direction, which is consistent with what is observed in the static experiments. In the following, the acoustic consequences of such a force will be considered as some basics of sound production are discussed.

To create sound, we can make local density fluctuations by changing the volume of fluid in a region small compared with the wavelength of sound in an oscillatory way. Such a source is called a monopole source, an example of which is provided by the volume velocity entering the vocal tract. If two monopole sources, 180 degrees out of phase, are put near one another (on the scale of wavelength), a dipole source is created. Such a source is inefficient because there is a great amount of partial cancellation [7]. An oscillating force within air causes air to accelerate from one region to another without an overall change in density, and so, resembles the dipole source just mentioned, if the force is active in a region small compared with wavelength. If it is assumed that the pressure head loss above the glottis occurs in a region short compared with wavelength, then the vorticity-velocity interaction force can be said to provide a dipole source of sound (see equation (4)). (In the static experiments the loss appears to occur within 2 cm. of the glottis, which means that this should be a good approximation up to about 3000 Hz.)

What effect does this dipole source have on the sound in the schematic vocal tract ? We follow the aeroacoustic theory of Powell [8], which was later elaborated by Howe [9], to draw the following picture. The oscillating volume velocity at the glottis can be considered a monopole source of irrotational fluid motion, which would be heard as sound in the far-field if nothing else was to intercede. However, because of the abrupt area change, there is a transfer of energy from the irrotational fluid motion to rotational motion. The forcing associated with the transfer of energy is the vorticity-velocity interaction force. Further, this force provides a dipole source of sound. The one-dimensional analogs, while providing a resistance to the volume flow, do not account for the radiation of the pressure fluctuations of the rotational fluid motion. The picture presented here is similar to that described by Howe [10] and Bechert [11] for the attenuation of low frequency sound transmitted from a low Mach number jet.

It is of interest to estimate the ratio of the strengths of the acoustic fields due to these sources. Without filling in the mathematical details, the relative strengths of the two sources can be considered, as well as the efficiency with which they radiate. Suppose the glottal fluid particle velocity is a rectangular wave and the glottal area is a triangular wave. Further, suppose the ratio of the vocal tract area to the glottal area is greater than or equal to five, the duty cycle is greater than one-third, and that the quasi-steady approximation holds. It can be shown that the ratio of the dipole source strength to the monopole source strength goes as the first power of the glottal Mach number, the first power of the ratio of the vocal tract area to the maximum glottal area, and to the first power of the frequency to a good approximation above, say, the second harmonic. As may be expected from the fact that the dipole source strength is a nonlinear function of fluid particle velocity, the importance of this source grows with frequency, after a given frequency.

As far as the efficiency of radiation, the monopole, volume velocity source appears to be efficient because it is located at a high impedance boundary: the glottis. The result is a velocity source at a pressure maximum, which means efficient energy exchange. On the other hand, the dipole source is a pressure type source, located just above the same boundary. This will mean an inefficient radiation, which will improve with frequency, at least as long as the source region is short compared with wavelength (recall the discussion on dipoles). In this frequency regime, the ratio of the the acoustic field due to the dipole source to that due to the monopole source increases as the first power of the frequency.

In the estimates made above, the impedance looking into the glottis is presumed infinite. This allows us to prescribe the input volume velocity into the vocal tract, and it allows us to say that the efficiency of radiation of the volume velocity source is a constant function of frequency. To account for source-tract interaction, a finite, time varying impedance at the glottis needs to be considered. This impedance may be difficult to deduce. The subglottal region needs to be considered as part of the acoustic system and the appropriate Green's, or transfer, function derived for this more complicated geometry. While the imposition of an explicit boundary condition at the glottis is avoided in this manner, other difficulties arise, such as the the fact that different ambient conditions apply in the subglottal and supraglottal regions.

CONCLUSION

Using the three-dimensional character of the fluid velocity field and the geometric property of sudden change in area, a sound source, other than volume velocity, can be identified. This source is a dipole type source, and is a poor radiator at low frequencies. Other modifications to the picture of sound sources during phonation can be expected when a more realistic geometry is considered (e.g. the epiglottis). The difficult question of source-tract interaction needs to be considered by looking at both the subglottal and supraglottal regions simultaneously. A more satisfactory picture of the acoustics of phonation can result by deriving the acoustic field from the conservation equations governing the motion of air.

References

[1] D. G. Crighton. Acoustics as a branch of fluid mechanics. *J. Fluid Mech.*, 106:261–298, 1981.

[2] P. M. Morse and H. Feshbach. *Methods of Theoretical Physics.* McGraw-Hill, New York, 1953.

[3] Jw. van den Berg, J. T. Zantema, and P. Doornenbal, Jr. On the air resistance and the bernoulli effect of the human larynx. *J. Acoust. Soc. Am.*, 29:626–631, 1957.

[4] K. Ishizaka and M. Matsudaira. *Fluid Mechanical Considerations of Vocal Cord Vibration.* Speech Communications Research Laboratory, Santa Barbara, CA, 1972.

[5] U. Ingard and H. Ising. Acoustic nonlinerarity of an orifice. *J. Acoust. Soc. Am.*, 42:6–17, 1967.

[6] L. D. Landau and E. M. Lifshitz. *Fluid Mechanics.* Pergamon Press, New York, 1959.

[7] M. J. Lighthill. *Waves in Fluids.* Cambridge Univ. Press, Cambridge, England, 1978.

[8] A. Powell. Theory of vortex sound. *J. Acoust. Soc. Am.*, 36:177–195, 1964.

[9] M. S. Howe. Contributions to the theory of aerodynamic sound, with applications to excess jet noise and theory of the flute. *J. Fluid Mech.*, 71:625–673, 1975.

[10] M. S. Howe. The dissipation of sound at an edge. *Journal of Sound and Vibration*, 70:407–411, 1980.

[11] D. W. Bechert. Sound absorption caused by vorticity shedding, demonstrated with a jet flow. *Journal of Sound and Vibration*, 70:389–405, 1980.

LABORATORY-INDUCED SPEECH ERRORS IN HINDI

MANJARI OHALA

Linguistics Program
San Jose State University
San Jose, California 95192 (USA)

JOHN J. OHALA

Department of Linguistics
University of California
Berkeley, California 94720 (USA)

ABSTRACT

Although speech errors are claimed to be universal, we have observed no naturally-occurring errors in Hindi which break up words. We therefore tried to induce such errors in Hindi speakers using a laboratory method. Subjects saw printed word pairs which appeared in rapid succession and were randomly required to pronounce some of them out loud, occasionally in reverse order. Some trials resulted in errors. Approximately 3% of all trials yielded errors which involved fragmentation of words, considerably less than the 10 to 40% error rate reported for English. The reason for the lower error rate in comparison to other languages remains to be discovered.

INTRODUCTION

Although there has been scientific interest in speech errors for nearly a century, it is only in the past few decades that there has been a virtual explosion of studies by linguists interested in showing how such errors shed light on issues in linguistic theory [1, 2, 3, 4, 5, 6, 7]. There seems to be an implicit claim in much of this literature that speech errors should be found in all languages; Fromkin has made this claim explicitly (personal communication). To date, errors have been reported primarily for Western Indo-European languages, e.g., German [8], Dutch [6], English [5]. We are aware of a collection of Japanese speech errors (S. Hiki, personal communication). This still leaves the vast majority of languages--even language types--unaccounted for. Relevant to this is the impression of the first author of this paper, a native speaker of Hindi, that she has never encountered in her own Hindi speech or that of others speech errors of the type that break up parts of words, e.g., spoonerisms of the sort '...it is kistomary to cuss the bride' (for '...customary to kiss...').

A SKETCH OF HINDI PHONOLOGY

Hindi has a relatively large number of segments: 20 stops (including affricates), 4 fricatives, 9 sonorant consonants--of these 33 consonant types, 25 can be geminate as well--, 11 oral and 10 nasal vowels [9].

Although medial consonant clusters are abundant and quite complex, initial and final clusters tend to be few, especially in native vocabulary, amounting largely to #C + glide- and -st# or -homorganic nasal + stop#, respectively. Most Hindi words range from one to three syllables; four and more syllables per word are uncommon. The prosodic structure of Hindi is controversial and will be discussed further below.

Hindi words, like those of most Indo-European languages, may be morphologically quite complex, showing affixes (both prefixes and suffixes). In a few cases grammatical categories are marked by vowel ablaut.

EXAMINING THE ANECDOTAL EVIDENCE

What is the observation?

As mentioned above, the observation is that speech errors that involve divisions of words or morphemes into fragments (henceforth WF for 'word fragmentation') with, optionally, their rearrangement into 'erroneous' or unintended strings (whether these strings themselves constitute valid words or not) do not occur naturally in Hindi. The types of errors that seem to be relatively easy to find in other languages, e.g., anticipation: "a [mæt]... < a man's natural inclination", perseveration: "John gave the boy" -> "...gave the goy", transposition: "keep a tape" -> "teep a cape" [5].

Observational error?

What are the possibilities that this observation is faulty--that the errors are there but are overlooked for some reason? We believe this is unlikely for the following reasons.

1. The same observer (that is, the first author) has detected many grammatical errors (e.g., lack of concord) in the speech of Hindi speakers, as in / ləkṛi ko pəkəṛ kər kuṭ:a ko maro / for /... kuṭ:e .../ (literal translation: "stick (postpos.) hold (verb particle) dog (postpos.) beat", free translation: "Take the stick and beat the dog").

Further, this observer has had no trouble observing WF errors made by English-speakers speaking English and even by Hindi-speakers (including the first author herself) when they speak English, e.g., '...crogged freeways < ('clogged freeways').

2. The first author has also asked several other Hindi-speakers, including many trained linguists, if they have observed any speech errors in Hindi (providing them examples from English, if necessary) and their impressions have always coincided with hers: no such errors in Hindi.

It would seem that the anecdotal evidence on the scarcity of Hindi WF speech errors is not marred by observational bias. Nevertheless, as in any issue of this sort, it would be highly desirable to augment our observations with experimental data. Our preliminary attempts to do this are given in the next section.

THE EXPERIMENT

Introduction.
Baars and Motley [1] have introduced a method--with several variants--for obtaining speech errors in abundance in the laboratory (see also [2, 3, 4, 10]). We decided to apply one of these variants to Hindi speakers to see if we could get speech errors in the same way they did. Their method has been applied successfully to speakers of languages other than English [7]. We chose to present stimuli orthographically (using the Devanagari script), requiring the occasional utterance of the stimulus phrases in original or reverse order, and using stimuli which would yield meaningful words if produced with initial consonants reversed.

The method.
A series of two word phrases were presented orthographically one after another to subjects (Ss) for a brief interval. At unpredictable times, Ss were given a signal to pronounce out loud the last phrase that they read (which was then no longer visible). Sometimes the signal required that they pronounce the two words in the same order and at other times in reverse order. Given the pressure of time, etc., Ss were liable to produce some of these spoken trials with speech errors. To present the Hindi words written in Devanagari to our Ss we used a 'memory drum', a device which advances a roll of paper (on which the stimulus words are written) a line at a time, for a controllable interval, such that only one line is visible at any given moment. We presented 145 two word sequences written in black ink with 40 randomly intermixed instructions ('same' and 'reverse') written in red ink to 11 adult male native speakers of Hindi (Indian students at University of California, Berkeley, who could read Devanagari). Ss were paid for their participation. With six Ss the inter-stimulus interval (ISI) was 1.8 second--twice as long as that usually used by Baars and Motley in their studies--but since Devanagari is graphically more complex than the Roman alphabet, this seemed justified. With 4 Ss we used a faster rate of 1.1 ISI. Ss were given 12 stimulus sequences including three instruction words as practice.

The two word sequences occasionally formed what might be construed as a meaningful phrase but generally they did not. The placement of the instruction words and whether they were to repeat the preceding sequence in the same order or the reverse, were randomly placed in the list, except that the instruction words never occurred more than seven trials apart. A portion of the list is given in Table 1, where the items (/sid̠ʰa/) and (/olṭa /) constitute the directions to repeat the last sequence in the same order or reverse, respectively.

Ss were told that this was part of a memory experiment and were instructed that when the words /sid̠ʰa/ and /olṭa / appeared they were to say out loud, in the order indicated, the last two words that they had read. They were told to answer as quickly and as accurately as possible and that they could earn 25% more if their speed and accuracy exceeded an unspecified threshold. In fact, there was no such criteria and all Ss were paid the 'bonus'.

Table 1. Representative sequence of stimulus word and, where relevant, possible errors due to transposition.

Stimulus		Possible Error
sal moči	*(year; cobbler)*	
pʰəl ǰoṛa	*(fruit; collected)*	ǰəl pʰoṛa *(water; sore)*
olṭa	*(REVERSE)*	
ǰel ṯali	*(jail; key)*	
čəṛʰ pəla	*(climb; raised)*	pəṛʰ čəla *(read; went)*
sid̠ʰa	*(SAME)*	

Ss were seated in a sound-treated room, facing the memory drum. The handwritten Devanagari characters were well illuminated and subtended approximately a .45 degree vertical angle in the Ss' visual field. A microphone was placed approximately 10 cm. from the subject's mouth and positioned in a way so as not to obscure the view of the slit showing the stimuli; responses were recorded on a high-quality analog tape recorder for later analysis.

Results.
Table 2 presents the results in terms of number of successful responses and number of errors, the latter broken down (see indented columns) into no response, ordering error (reversing when not required to, failing to reverse when required to), errors attributable to probable misreading (due to graphical similarity of certain Devanagari symbols), errors attributable to probable intrusion of parts of words presented just prior to the target sequence (and thus more a memory error than a speech production error), ambiguous errors (cause unknown), and WF errors. A representative sample of the WF speech errors is given in Table 3.

Table 2. Correct and Erroneous Responses

Condition: Response Type:	1.8 sec ISI	1.1 sec ISI	Total
Correct:	235	103	38 (76.8%)
Errors:	45	57	102 (23.2%)
No Response	8	14	22 (5%)
Failure to follow instructions	8	9	17 (3.8%)
Probable misreading	5	13	18 (4.1%)
Influence of prior stimuli	4	4	8 (1.8%)
Ambiguous	13	9	22 (5%)
Word Fragmentation	7	8	15 (3.4%)

Table 3. Representative speech errors obtained; 'S' = words were to be in same order; 'R' = words were to be in reverse order.

Stimulus	Error
čəṛʰ pəla [S] *(climb; raised)*	čəl pəṛʰa *(go; read)*
d̠er bag [R] *(delay; garden)*	bar d̠eg *(turn; nonsense)*
pʰɛl moṛa [R] *(spread; turned)*	pʰoṛa mɛl *(sore (n); dirt)*
kaṭ kʰil [R] *(cut; puffed rice)*	kʰil kʰaṭ *(puffed rice; cot)*
nila ʃap [S] *(blue; curse)*	nali ʃap *(drain (n); curse)*
ḍal mori [S] *(branch; drain (n.))*	ḍar moli *(nonsense; nonsense)*
ṯəlaʃ pal [S] *(search; raise)*	pəlaʃ pal *(type of tree; raise)*

Discussion.
These results show at least that WF speech errors can be induced in speakers of Hindi in spite of the apparent lack of such in naturalistic situations. The rate at which such errors occurred however, 3.4%, or even the 5% for the shorter ISI, is far less than the 10 to 40% reported by Baars [10] and Baars and MacKay [3]. It is possible that lower ISI's would yield a greater error rate (although Baars [10] suggests that the errors are successfully elicited at ISI rates from about .5 to 3 sec) but we believe that the greater graphical complexity of the Devanagari script requires longer ISI in order to allow the stimuli to be accurately read by the Ss. A smaller ISI would no doubt yield an inefficiently high percentage of uninteresting errors (no responses, misreadings, etc.). This lower error rate, vis-a-vis those obtained for experiments involving English, is compatible with the anecdotal observation that WF speech errors are uncommon in Hindi.

The experiment was not designed to and thus did not give any clues as to why Hindi exhibits so few errors of this sort. Furthermore, as noted by Baars and Motley [2] we have no way of knowing whether speech errors elicited experimentally have all the properties of errors produced under natural situations. However, as in previous work with speech errors, whether gathered naturalistically or in the laboratory, the vast majority of errors resulted in real words [4].

The question arises: how can we be sure that what we counted as WF errors were genuine speech errors, i.e., unintended production errors (like typing mistakes) made after the process of correct planning of the lexical sequence and not failures of memory, etc., i.e., errors made before the planning of the lexical sequence? When the error was a nonsense word we can be fairly sure it was a speech error as this is usually defined. However, in other cases there is, in fact, some ambiguity in the interpretation. Baars and Motley [2] answered this question by operationally defining a slip 'as an error of output that systematically violates the target as presented to the subject'. We follow the same practice here but recognize the desirability of refining the notion of 'speech error' in this type of experiment. It might be advisable in future such studies to allow the subjects to indicate somehow when they detect an error in their own response.

GENERAL DISCUSSION

If it can be accepted that WF speech errors are scarce in Hindi, this immediately raises the question: how is Hindi different from other languages whose speakers exhibit numerous errors? We can examine several possibilities:

Tradition of Word Decomposition.
Could it be the case that the Hindi-speaking community has no tradition which involves analysis of words into parts? The answer would seem to be 'no'. Poetic devices (rhyme, alliteration), certain word games, and many regular phonological processes all require speakers to be able to break words up into syllables and phonemes (for details, see [11]).

The Devanagari script.
Could the Devanagari script somehow account for the scarcity of WF errors? It has been demonstrated in psycholinguistic studies with English speakers, that orthography can have a major influence on native speakers' phonological knowledge [12]. One of the 11 Hindi vowels, /ə/, has no overt symbol when forming part of the CV syllable but is an understood part of each consonantal symbol. If this were the general orthographic practice it might suggest that Hindi speakers (if influenced by the script) would be less able to dissociate C from V in CV sequences and thus would be less likely to break up such sequences. However all the other ten vowels are represented overtly and this would imply that the script presents no bar to the native speaker's analysis of words into their phonemic constituents.

Prosody
There is one aspect of Hindi, however, which may be a good candidate to account for its odd behavior with regard to speech errors, namely its prosodic structure. Although it is disputed whether Hindi has stress or not, even those writers who claim it has stress agree that it is much weaker phonetically than in English and plays little role functionally (differentiates few, if any, minimal pairs; see [13]). Research by the first author [13] seems to indicate that in Hindi stress probably only involves pitch, unlike languages like English, German, and Russian, where stress correlates include pitch, duration, intensity, and vowel reduction. Furthermore, rather than being an immutable property of a word, as generally true in English, stress assignment shows considerable mobility in Hindi since more than one phonetically eligible syllable in polysyllabic words (i.e., strong syllables) can receive stress under different circumstances.

The existence of strong word stress seems necessarily to imply some kind of hierarchical structure

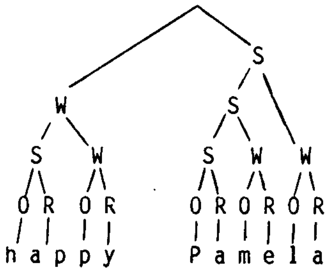


Fig. 1. Hypothetical structure determining stress.

to words and, possibly, phrases, i.e., some structure which clumps syllables into feet, marking one syllable in the clump as dominant (strong) and the others subordinate (weak) [14]. MacKay [15] (and others) have noted that speech errors typically involve segments from the same position in adjacent feet, i.e., syllable initial segments in stressed syllables usually interchange with (or anticipate or perseverate) syllable initial segments in the adjacent stressed syllables, etc. Thus in the phrase 'happy Pamela', with the hierarchical structure indicated in Fig. 1 (where 'S', 'W', 'O' and 'R' stand for 'strong', 'weak', (syllable) 'onset' and (syllable) 'rhyme', respectively), the hypothetical (perhaps improbable) speech error, 'pappy hamela', could occur if the similarly labeled S-nodes /pæ/ and /hæ/ got mixed up but, given the constraints, 'correctly' fitted into eligible positions within the hierarchical structure, i.e., next to low-level W branches.

If stress is not very strong, as is the case in Hindi, such hierarchical structure may either be absent or functionally less important. Then, if speech errors occur primarily at these lower levels of the prosodic hierarchy (one may speculate that this hierarchical structure is cognitively 'costly' and may therefore be more subject to break-down), the lesser salience of this level in Hindi--or its absence--might account for the scarcity of WF speech errors in the language. This, of course, is speculation and needs further investigation.

CONCLUSION

Speech errors which involve breaking up words into parts are scarce in Hindi: they have not yet been observed under naturalistic conditions and occur under laboratory conditions with much less frequency than has been found for comparable studies with English. Hindi, therefore, must be different in some way from those languages exhibiting numerous errors, e.g., English, Dutch, German. A different prosodic structure seems to be a good candidate for the factor giving rise to this difference. This issue is worth pursuing (a) for its typological interest and (b) the light it could shed on the mechanism of speech errors and, in that way, on how speech is produced.

ACKNOWLEDGEMENTS

We thank Mariscela Amador for assisting in the running of the speech errors experiment, Michael Motley for loaning us his memory drum and for giving advice on experimental design. The experiment was supported in part by the Committee on Research, University of California, Berkeley through a grant to the second author.

REFERENCES

[1] Baars, B. J. & Motley, M. T. 1974. Spoonerisms: Experimentally elicitation of human errors. Journal Supplement Abstract Service. Catalog of Selected Documents in Psychology, Fall 1974.

[2] Baars, B. J. & Motley, M. T. 1976. Spoonerisms as sequencer conflicts: Evidence from artificially elicited errors. Am. J. Psychol. 89.467-484.

[3] Baars, B. J. & MacKay, D. G. 1978. Experimentally eliciting phonetic and sentential speech errors: Methods, implications, and work in progress. Language in Society 7.105-109.

[4] Baars, B. J., Motley, M. T., & MacKay, D. G. 1975. Output editing for lexical status in artificially elicited slips of the tongue. J. Verbal Learning & Verbal Behavior 14. 382-391.

[5] Fromkin, V. A. 1971. The non-anomalous nature of anomalous utterances. Language 47.27-52.

[6] Nooteboom, S. G. 1969. The tongue slips into patterns. In A. G. Sciarone et al. (Eds.), Nomen: yden studies in linguistics & phonetics. The Hague: Mouton.

[7] Stemberger, J. P. & Lewis, M. 1986. Reduplication in Ewe: Morphological accomodation to phonological errors. Phonology Yearbook 3.151-160.

[8] Meringer, R. & Mayer, K. 1895. Versprechen und Verlesen: Eine Psychologisch-Linguistische Studie. Stuttgart: Göschensche Verlagsbuchhandlung.

[9] Ohala, M. 1983a. Aspects of Hindi phonology. Delhi: Motilal Banarsidass.

[10] Baars, B. J. 1980. On eliciting predictable speech errors in the laboratory. In V. A. Fromkin (ed.), Errors in linguistic performance. Slips of the tongue, ear, pen, & hand. New York: Academic Press. 307-318.

[11] Ohala, M. & Ohala, J. J. In press. The scarcity of speech errors in Hindi. In L. M. Hyman & C. Li (eds.), Language, Speech & Mind.

[12] Jaeger, J. J. 1984. Assessing the psychological status of the Vowel Shift Rule. J. Psycholinguistic Res. 13.13-36.

[13] Ohala, M. 1986. A search for the phonetic correlates of Hindi stress. In Bh. Krishnamurti (ed.), South Asian Languages. Structure, Convergence & Diglossia. Delhi: Motilal Banarsidass. 81-92.

[14] Liberman, M. & Prince, A. 1977. On stress and linguistic rhythm. Linguistic Inquiry 9.249-336.

[15] MacKay, D. G. 1972. The structure of words and syllables: Evidence from errors in speech. Cognitive psychol. 3.210-227.

THE HISTORY OF THE CLASSICAL VOWEL ARTICULATION MODEL: A REPLY TO CATFORD AND FISCHER-JØRGENSEN

SIDNEY A J WOOD

Department of Linguistics and Phonetics
University of Lund
Sweden

ABSTRACT

This paper is devoted to a discussion of Catford's (1) and Fischer-Jørgensen's (2,3) defence of the classical vowel articulation model. Objections to the model are not directed at Bell's personality, but at the theoretical structure of his model: critical functions of the model are contradicted by empirical data and by acoustical theory. Nor are the objections only relevant for amendments introduced by Jones. That argument is contradicted by the chronology of the debate.

BACKGROUND

The classical vowel model, originally introduced by Bell in 1867 (4) and modified into various versions by other authors, is characterized by the class of central vowels. The model was designed around the single resonance theory, according to which the upper surface of the tongue narrows the mouth channel locally in order to delimit the buccal cavity and tune its natural resonance. Bell postulated a configurative aperture that "may be shifted to any part of the back or front of the palatal arch" (p. 71). He held that the horizontal and vertical position of the tongue arch relative to the roof of the mouth set the size and location of this aperture, so that the natural resonance of the mouth cavity would rise progressively as the tongue moved from low to high at the back, central and front locations in turn.

Much of Bell's terminology was soon changed. Sweet substituted raised for the higher modification of tongue height. Ellis replaced inner and outer by retracted and advanced. The I.P.A. adopted the French tradition of four degrees of opening. Jespersen preferred three degrees, Jones followed the I.P.A. The I.P.A. and Jones retained Bell's term mixed until the 1920s, when central was substituted. The dynamic periods in the evolution of the model and the progress of the debate are the 1880s, and the years around 1915 and 1930, when authors undertook major revisions of their textbooks in response to Bell's original proposal, and to the negative data reported by Meyer in 1910 (5) and by Russell in 1928 (6).

The classical vowel model rapidly superseded the ancient throat-tongue-lip model. It was adopted by the neogrammarians and the I.P.A., and was hypostetized long before it could be tested. The first empirical data on the model, reported by Meyer and by Russell, contradicted some tongue heights postulated by Bell, especially for [ɪ,e] and [ɔ,ɑ]. Phoneticians, already divided between the rival organic and acoustic paradigms, took sides in a bitter feud. Fischer-Jørgensen has given her personal recollections (3) of how the controversy was conducted.

Analysis of 38 sets of midsaggital vowel profiles (7,8,9), collected from the literature, confirm the anomalous heights reported by Meyer and Russell, and gave no evidence of intermediate configurative apertures, i.e., of Bell's class of central vowels. It was concluded that the classical model was based on an oversimplified acoustic theory and that it is contradicted by physiological data, which weakens its validity and explanatory power.

Catford and Fischer-Jørgensen argue that critics of the classical vowel model have tended to exaggerate. They question the value of radiographic data as evidence against Bell's model and they point out that not all sets of x-ray pictures contradict the model. Fischer-Jørgensen also maintains that criticism of the classical model is really directed at amendments introduced by Daniel Jones, and that an alternative articulation model based on resonance phenomena in the entire vocal tract would be less suitable for phonology.

CATFORD'S AND FISCHER-JØRGENSEN'S ARGUMENTS

The Value of Radiographic Data as Evidence

Catford emphasizes that Bell's and Sweet's vowel descriptions were based on perceived muscular sensation, not on objective (radiographic) records of actual tongue position. "There is obviously a close correlation between the objective and proprioceptive data, but one should not expect them to be identical" (p. 23). This recognizes the difficulties faced by Bell and Sweet when they judged tongue positions from muscular sensation. But the argument against the classical model is not just that Bell's kinesthetic vowel judgments are sometimes contradicted by radiographic evidence. That could easily be allowed for and corrected. For example, as Catford suggests, one could recognize [ɪ] as mid or half-open, although Jespersen rejected precisely that solution when he revised (10) in response to Meyer's data. What the x-ray pictures fail to confirm are certain tongue positions predicted by the model itself from the inadequate acoustic theory on which it was based. The criticism is thus directed at the empirical and theoretical structure of Bell's model, and not at Bell's personality or his phonetic skill.

Bell's contemporaries nevertheless disputed several postulated tongue positions. For example, [ɛ] in English let is low according to Bell but mid according to Ellis, Sweet and Storm. The [ɑ] of father is mid according to Bell and Sweet but low according to everyone else. Jespersen found the positions for mid back and low back vowels especially difficult to analyse. This is why phoneticians invented measuring devices like Grandgent's discs, Atkinson s probe, Zünd-Burguet's pneumatic height indicator and Meyer's plastograms, and why they finally turned to radiography. The x-ray profiles show that mid back and low back vowels are difficult because their heights are random. The fact that kinesthetic judgments of tongue position deviate from the true position underlines their unreliability and demonstrates the need for a more reliable method of observation.

Which Sets of X-ray Profiles Support the Model?

Catford and Fischer-Jørgensen point out that some sets of x-ray profiles agree with the traditional description. But any statistical survey of biological events will contain individual items that contradict the trend. The problem is that so many sets of x-ray profiles do not support the classical model on the intermediate apertures of central vowels, on tenseness and laxness and on the heights of mid and low back vowels. There must be other principles governing the articulation of these vowels. The radiographic data can be summarized as follows (7,8):

- So-called central vowels do not have intermediate configurative apertures
- The tongue is usually lower for high [ɪ] than for mid [e]
- Only one third of mid [ɔ] usually come out higher than low [a-ɑ]
- Two thirds of mid [o] usually come out higher than low [a-ɑ]
- Only one third of high [ʊ] usually come out higher than mid [o].

Clearly, this leaves room for many sets of profiles to come out all right. The most disturbing aspect of the x-ray data is the failure to substantiate the class of central vowels, which was Bell's revolutionary innovation. Jespersen (11, pp.18-19) had observed as early as 1889 that there were no low central vowels, since the tongue made a discrete transition between front and back apertures owing to the domed shape of the palate. The elusiveness of the intermediate apertures may well also be the reason why phoneticians started referring to the highest part of the tongue instead.

Daniel Jones and the Cardinal Vowels

Fischer-Jørgensen states that the objections to the classical model are only valid as regards revisions introduced by Daniel Jones in the classical system for the purpose of his cardinal vowel chart (2, p.260; 3, p.82). She particularly mentions (i) use of the highest point of the tongue as a reference point, and (ii) rejection of tense/lax.

These points are not supported by the chronology.

Jones's cardinal vowel scheme, with its articulatory limits and auditory scales of timbre, was a compromise between the organic and acoustic paradigms. But the first editions of Jones's textbooks (12,13) were completely in the organic tradition. The Outline (13) was already in press in 1914, but publication was delayed by the war. Jones's recognition of Meyer's report (13, p.19) survived all the subsequent revisions of (13). The cardinal vowel scheme was introduced in 1917: the x-ray pictures (14) were made in January, the gramophone recording (15) was issued shortly after, and a cardinal diagram was included in Dent's dictionary (16). A brief preface was added to the 1922 reprint of (13), but a full account was not given until 1923 (17, pp.24, 27-41), followed by Jones's contribution to the 1925 Copenhagen conference (18) and the 1932 revision of (13). The cardinal vowel scheme was introduced after Meyer's report, while (13) was not fully revised to include the cardinal scheme until after Russell's report. One can hardly conclude that Meyer's and Russell's reports were directed at Jones. On the contrary, Russell (19) welcomed the timbre scales of the cardinal vowel scheme. But Jones and his associates also assumed that equal increments on the timbre scales corresponded to equal increments of tongue position in Bell terms, so that the cardinal scheme has preserved the classical model within itself. In that sense it is open to the same criticism as the classical model.

The highest point of the tongue, used by Jones as a reference in all his books from (12) onwards, is not a necessary component of the cardinal vowel scheme. For Bell, the size and location of the postulated configurative aperture were congruent with the high/low and front/back position of the upper surface of the tongue. The immediate problem was to identify what part of the tongue was raised and how far. By the 1880s, this had evidently been reduced to determining the highest part of the tongue (Jespersen had occasion to emphasize that this does not coincide with the narrowest part of the mouth channel, see above). The highest part of the tongue was already a well established reference point at a time when Jones was still a child. He did not introduce the concept, he adopted an established practice.

Jones's first stand on tenseness was similar to Sweet's, tensing or relaxing of the tongue with no difference of height (12, p.12). But there were others, like Jespersen (11,20), who believed lax (broad) vowels were slightly lower than the corresponding tense (thin) vowels. He held that this was

achieved by furrowing the tongue, or, especially for lower vowels, by lowering it. The issue was disputed, see Sweet's correspondence with Storm (21). Sweet believed it was a matter of convexing versus flattening of the tongue, all at the same height. The problem was thus whether laxness was distinct from height or whether it could be subsumed with raising/lowering.

Jones gradually amended his initial stand. He reported Meyer's results (13, pp.19-20) and modified his own view of laxness to admit lowering. He finally rejected laxness in favour of lowering in the 1932 revision of (13).

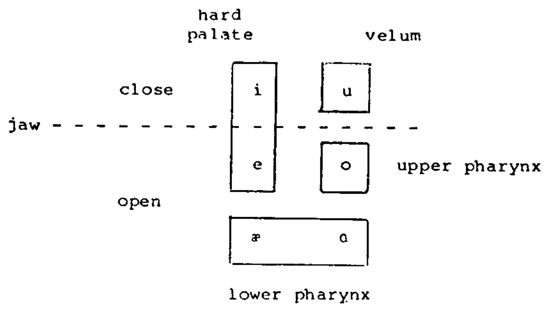
The criticisms of the classical model are clearly not aimed at Jones's usage, which represents his various stands on older issues. His changing views on tense/lax, and especially his innovation of the cardinal vowel scheme, were surely his way of coming to terms with these very same objections to the classical model as the were made by Meyer and Russell. The objections refer to the fundamental structure of the classical model as it was conceived by Bell himself, and concern Jones only in so far as the cardinal scheme reflected and perpetuated the classical model.

The Utility of a Model Based on the Whole Vocal Tract

Fischer-Jørgensen does not expect an alternative model based on the whole vocal tract to be better than the classical model. She doubts whether sound typology supports the four locations analysed from x-ray data in (9) and she claims that this type of model is less useful for phonology.

The model in question recognizes four major classes of vowels depending on the part of the vocal tract that is narrowed - palatal [i-ɛ,y-œ]-like vowels, velar [u-ʊ,ɯ] -like vowels, upper pharyngeal [o-ɔ,ɤ]-like vowels and low pharyngeal [æ-a-ɑ]-like vowels. Within each class, vowels are differentiated by local manoeuvres involving the lips, tongue blade, tongue posture, larynx depression etc. (9,23,24,25).

These manoeuvres are related to the parameters of the classical model as follows:



Fischer-Jørgensen's reference to typology does not take into account the allophonic variation that is typical of "small" vowel systems. A two-phoneme system like Kabardin contrasts a set of low pharyngeal [æ ɑ]-like allophones with a set of palatal, velar and upper pharyngeal (uvular) allophones, phonemically low pharyngeal versus the rest. In three-phoneme systems, there is usually variation between velar [u]-like and upper pharyngeal [o]-like allophones. The same goes for the four-phoneme system quoted by Fischer-Jørgensen. Whatever the language and however simple the vowel system in terms of phonemic contrasts, the speaker utilizes all four locations.

Fischer-Jørgensen cites instances of vowel systems that, she claims, cannot be handled without three or four degrees of height (openness), whereas I proposed just two degrees of jaw position. This was based on radiographic data that showed the jaw opening tended to be narrower than about 9 mm for [i-u]-like vowels (typically 5-7 mm) and wider for [e-o-a]-like vowels (typically 10-12 mm) (23,24,26). The term openness is used in two different senses here, lingual and mandibular respectively. In the view of vowel articulation outlined above, the classical heights are redistributed between the open/close jaw positions and the four locations along the vocal tract. The categories for which Fischer-Jørgensen requires four heights are still available, but now defined in terms that more closely reflect the manoeuvres used in speech.

For example, using more tongue heights enables her to make more generalizations, such as recognizing that mid [e,ø,o] diphthongize to [iə,yə,uə], while low [ɑ] does not, a case that would be impossible to express without more heights. Let us say instead that [ɑ] is low pharyngeal, characterized by hyoglossal and glossopharyngeal activity, while [e,ø,o] and [i,y,u] share genioglossal or styloglossal activity:

$$\begin{bmatrix} \left\{\begin{matrix}+\text{pal}\\+\text{vel}\end{matrix}\right\} \\ +\text{open} \end{bmatrix} \longrightarrow [-\text{open}]$$

This solution makes predictions about the motor reorganization underlying this change.

I prefer to ask if it is possible to handle vowel systems in terms of resonator shaping by observable manoeuvres with known neuromotor activity. Such a model is a more effective instrument of prediction and explanation and should yield more plausible phonetic explanations for phonological problems. A test case is vowel harmony, which Fischer-Jørgensen believes can only be formulated very clumsily in terms of four places of articulation. However, Svantesson (22) has found that precisely this type of model provides the key to a solution of the problem of harmony and vowel shift in Mongolian, by focusing and capturing the variations in pharyngeal width that characterize this phenomenon. Formulating the problem in this way, Svantesson demonstrates that harmony in East Mongolian and its ancestor languages Ancient and Classical Mongolian are related. The shift from fronting harmony to pharyngeal harmony turns out to be a simplification, which offers an explanation for why there are no known examples of a shift in the opposite direction.

CONCLUSION

Catford and Fischer-Jørgensen have defended the classical model by questioning the data and by suggesting that the objections were really aimed at Daniel Jones. I have argued that hypotheses about articulation must be tested with the best available data, and I have shown that Jones's various amendments in fact represent his personal stands on older issues and that they were introduced in response to the data that contradicted the classical model.

The evidence against the classical model continues to recur and the same data consistently support an alternative solution. I do not see it as an exaggeration to report that the same data simultaneously provide confusing evidence for one interpretation and consistent evidence for an alternative interpretation.

REFERENCES

(1) J. C. Catford. "Observations on the recent history of vowel classification". In R. E. Asher and J. A. Henderson (eds), Towards a History of Phonetics, 19-32. Edinburgh University Press, Edinburgh (1981).

(2) E. Fischer-Jørgensen. "Vowel features and their explanatory power in phonology". In Abstracts 10th Int. Congr. Phon. Sc., 259-265. Foris Publications, Dordrecht (1983).

(3) E. Fischer-Jørgensen. "Some basic vowel features, their articulatory correlates, and their explanatory power in phonology". In V. Fromkin (ed), Phonetic Linguistics, 79-99. Academic Press, Orlando (1985).

(4) A. M. Bell. Visible Speech. Simpkin, Marshall and Co., London (1867).

(5) E. Meyer. "Untersuchungen über Lautbildung". Festschrift Wilh. Vietor, 166-249. Special number of Neueren Sprachen (1910).

(6) G. O. Russell. The Vowel. Ohio State University Press, Columbus (1928).

(7) S. Wood. "X-ray and model studies of vowel articulation". Working Papers 23:1-49, Dept. of Linguistics, Lund University (1982).

(8) S. Wood. "The tongue arching model of vowel articulation - its origin, evolution and validity". (Forthcoming).

(9) S. Wood. "A radiographic analysis of constriction locations for vowels". J. Phon. 7:25-43

(10) O. Jespersen. Lehrbuch der Phonetik, 2nd rev. edn. Leipzig: B. G. Teubner (1913).

(11) O. Jespersen. The Articulation of Speech Sounds Represented by Means of Analphabetic Symbols. N. G. Elwert, Marburg (1889).

(12) D. Jones. The Pronunciation of English. Cambridge University Press, London (1909).

(13) D. Jones. An Outline of English Phonetics. B. G. Teubner, Leipzig (1918).

(14) D. Jones. "Experimental phonetics and its uses to the linguist". Proc. Royal Inst. 22:8-21 (1917-19).

(15) D. Jones. Cardinal Vowels. The Gramophone Co., London, Record B804 (1917).

(16) D. Jones (Ed). An English Pronouncing Dictionary. Dent, London (1917).

(17) M. V. Trofimov and D. Jones. The Pronunciation of Russian. Cambridge University Press, London (1923).

(18)"D. Jones. "Das System der Association Phonétique Internationale". In M. Heepe (ed), Lautzeichen und ihre Anwendung in erschiedenen Sprachgebieten, pp. 18-27. Reichsdruckerei, Berlin (1928).

(19) G. O. Russell. "Synchronized x-ray, oscillograph, sound and movie experiments, showing the fallacy of vowel triangle and open-closed theories". In D. Jones and D. B. Fry (eds), Proc. Sec. Int. Congr. Phon. Sc., pp. 198-205. The University Press, Cambridge (1936).

(20) O. Jespersen. Fonetik. Schubotheske Forlag, Copenhagen (1897-99).

(21) K. Foldvik. "Sweet to Storm on narrow and wide". J. Int. Phon. Assoc. 7:4-9 (1977).

(22) J-O. Svantesson. "Vowel harmony shift in Mongolian". Lingua 67:283-327 (1985).

(23) S. Wood. "Radiographic and model studies of the palatal vowels". Working Papers 23:119-155, Dept. of Linguistics, Lund University (1982).

(24) S. Wood. "Tense and lax vowels - degree of constriction or pharyngeal volume?". Working Papers 11:110-131, Dept. of Linguistics, Lund University (1975).

(25) S. Wood. The acoustical significance of tongue, lip and larynx maneuvers in rounded palatal vowels. J. Ac. Soc. Am. 80:391-401.

(26) S.Wood. Jaw opening variation in vowels. Publ. 10, p. 107. Dept. of Phonetics, University of Umeå (1976).

MODEL EXPERIMENTS ON VOWEL REDUCTION IN BULGARIAN

THORE PETTERSSON SIDNEY A J WOOD

Department of Linguistics and Phonetics
University of Lund
Sweden

ABSTRACT

This paper reports the results of model experiments designed to test hypotheses concerning the articulatory correlates of vowel reduction in Bulgarian. It is concluded that reduction can be explained in terms of neutralization of mandibular depression, of lingual and labial compensation for mandibular variation, and of labial activity.

INTRODUCTION

Vowel reduction in Bulgarian is characterized, in traditional terms, by raising and not by centralization. There is a reducing set /e,o,a/, whose reflexes merge with those of a non-reducing set /i,u,ă/ in unstressed syllables in informal speech (/e/⟶[i], /o/⟶[u], /a/⟶[ă]). The character "ă" denotes an [ɐ] or [ɜ]-like timbre.

The actual occurrence of vowel reduction in everyday Bulgarian speech is subject to normative, social and dialect constraints (1). Unstressed /a/ is reduced in all dialects. Complete reduction of both /e/ and /o/ is limited to eastern dialects, but they are reduced to a varying extent elsewhere, depending on the formality of the situation. The reduction of /o/ is common in the Sophia dialect, but /e/ is frequently not reduced.

Figure 1 shows the F1 and F2 frequencies of the vowels in fully stressed syllables, recorded by an informant from the Sophia region. The examples of /u/ with the highest F2 frequencies are preceded by dental consonants. The position of /ă/ is central and midway between /e/ and /o/ (F1 about 350-450 Hz and F2 about 1300-1600 Hz), agreeing with the traditional analysis of the Bulgarian vowel system with /ă/ as an "indeterminate" mid central vowel.

However, the well known correlation between traditional tongue feature usage and formant frequencies does not express a causal relationship, since all parts of the vocal tract contribute to each resonance. One cannot deduce articulation by translating F1 into "height" and F2 into "backness". Further, it was formerly believed that a central tongue position narrowed the mouth channel at some point between the hard and soft palates, but this is not substantiated by midsagittal x-ray profiles (2,3). Instead, the vocal tract is narrowed at one of four locations: along the hard palate for [i-ɛ] and [y-œ]-like vowels, along the soft palate for

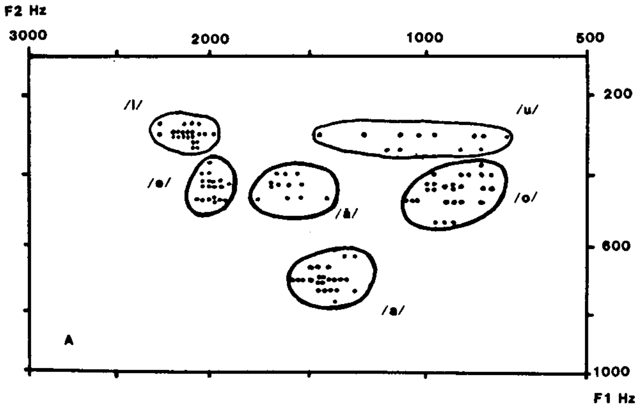


Figure 1
The frequencies of F1 and F2 for stressed vowels in isolated words and in focused words in sentences. The large variation in F2 for /u/ is related to consonant environment (high F2 with preceding dentals).

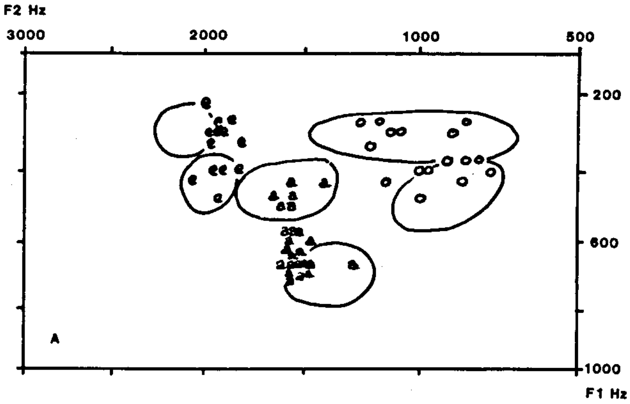


Figure 2
The frequencies of F1 and F2 for unstressed /e,a,o/ in isolated words, compared with the stressed vowel areas (Fig. 1). The variation of F2 for /o/ is related to consonant environment.

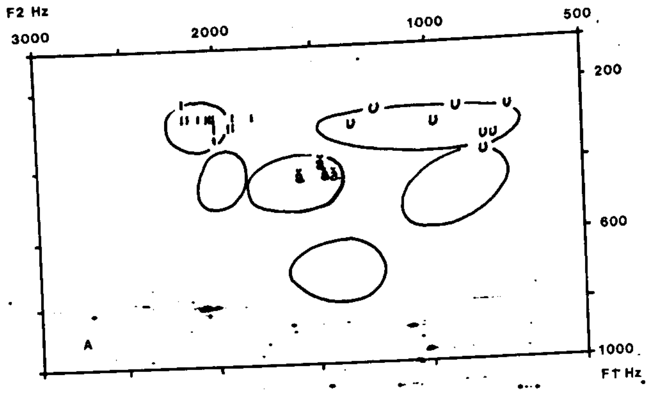


Figure 3
The frequencies of F1 and F2 for unstressed /i,u,ă/ in isolated words, compared with the stressed vowel areas (Fig. 1). The variation of F2 for /u/ is related to consonant environment.

[u-ʊ] and [ɯ]-like vowels, in the upper pharynx for [o-ɔ] and [ɤ]-like vowels, and in the lower pharynx for [æ-ɑ]-like vowels.

Examination of x-ray profiles by Tilkov (4,5) indicates that the tongue articulation for /ă/ and /a/ is similar, with a pharyngeal narrowing. The main difference is that the jaw opening is less open for /ă/, as in the similar difference between [ɛ] and [i] and between [o] and [u]. The spectral effects of these manoeuvres will be tested in the model experiments.

Figure 2 shows how reduction of /e,o,a/ shifted F1 and F2 well beyond the contrastive spectra of the accented vowels, in many cases causing coalescence with /i,u,ă/ respectively. The reduction recorded in Fig. 2 reflects the extent to which this informant has respected the norm and avoided reduction. This was most obvious in the recordings of the word lists. When the words were placed in sentences, there was more reduction, but /e/ typically remained most resistant. Figure 3 shows that the spectra of weak /i,u,ă/ are similar to those of the respective stressed forms.

Possible phonetic explanations for this pattern of reduction include (i) relaxed articulatory control, neutralizing articulatory components of the nonreduced vowel and leading to a continuous transition between unreduced and fully reduced forms. Alternatively, (ii) switching between integral articulations by substituting different manoeuvres, would lead to discrete jumps between reduced and nonreduced forms. Explanation (ii) is, by its nature, categorical. Phonological accounts are also, for convenience, categorical. The spectral regressions between nonreduced and fully reduced /e,o,a/ in Fig. 2 support explanation (i).

MODEL EXPERIMENTS

The spectral consequences of selected articulatory manoeuvres are studied by manipulating the contours

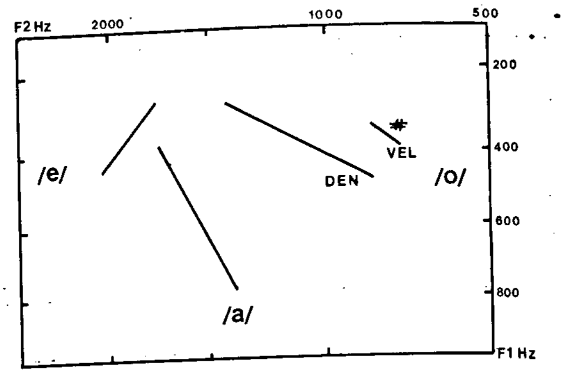


Figure 4
Regressions of F1 and F2 for all stressed and unstressed /e,o,a/. The /o/ renderings are divided into subclasses according to consonant environment: word initial /o/ (✱), preceding velars (VEL), preceding dentals (DEN).

of vocal tract profiles and then computing the resonance frequencies for each new configuration. The rationale for selecting manoeuvres to experiment with is outlined in detail at each example. The general principle is to deduce appropriate manoeuvres in each case from knowledge of speech physiology and the structure of each vowel configuration.

The spectrographic data in Figs. 2, 3 were obtained from 56 words recorded in isolation and in sentences, both in focal and nonfocal positions (6). The recordings were analysed using an LPC method. The trends of the gradual spectral transitions from the nonreduced renderings to the fully reduced renderings are captured by linear regressions in Fig. 4. The aim of the model experiments is to reproduce the regressions by subtracting selected manoeuvres from complete [ɛ], [a] and [o]-like profiles.

Analysis of published radiographic data for some 15 languages (7) indicates that, for nonhigh vowels, the mandible is usually depressed beyond about 8 mm, typically to about 10 mm. Larger openings, to say 15 mm, would occur for example in public speaking. Smaller openings would occur in mumbling. Usually the tongue compensates for mandibular variation by maintaining a suitable degree of narrowing of the palatal passage, of the velar passage or of the pharynx, as the case may be.

Pharyngeal /a/

The first experiment examines the effect of varying the jaw opening from 14 to 6 mm with full lingual compensation in order to maintain the optimum degree of pharyngeal narrowing. This modification is illustrated at A in Fig. 5.

The next experiment assumed that lingual compensation is turned off during reduction (the tongue would not be drawn back into the pharynx at closer jaw positions). Another likely characteristic of reduced /a/ is that the typical pharyngeal activity is weakened. Both of these features will result in a wider pharynx and narrower palatal passage. This

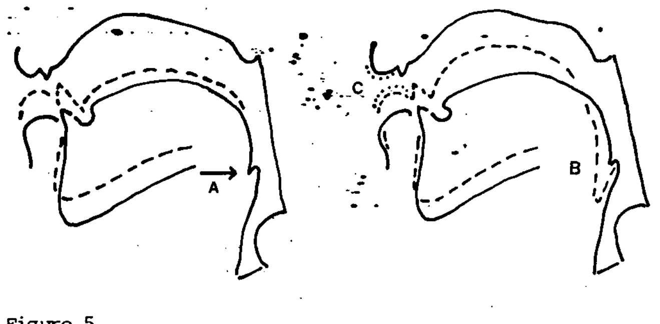


Figure 5
Modelled jaw opening variation 14-6 mm in an /a/-like configuration: with full lingual compensation (A), no lingual compensation (B), and weakened lip spreading (C).

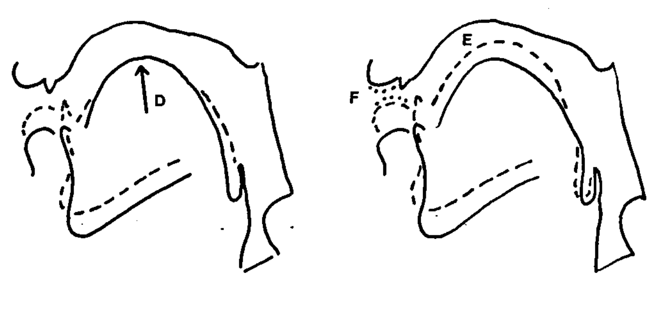


Figure 6
Modelled jaw opening variation 14-6 mm in an /e/-like configuration: with full lingual compensation (D), no lingual compensation (E), and weakened lip spreading (F).

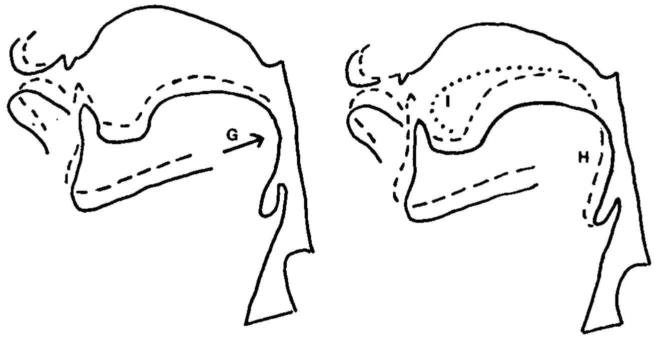


Figure 7
Modelled jaw opening variation 14-6 mm in an /o/-like configuration: with full lingual (G) and labial compensation, no lingual compensation (H), and elevated tongue blade (I).

combined modification is illustrated at B in Fig. 5.

A further possible reduction effect is neutralization of lip spreading (C in Fig. 5).

Palatal /e/

The same variation of jaw opening between 14 and 6 mm with perfect lingual compensation for /e/ is illustrated at D in Fig. 6. This maintains an ideal degree of palatal constriction. E in Fig. 6 illustrates the same mandibular reduction, but without lingual compensation (the narrower palatal passage due to the higher mandible position is not corrected by lowering the the tongue). Again, a further possible feature of vowel reduction is neutralization of lip spreading (F in Fig. 6).

Pharyngovelar /o/

The profile modifications for raising the mandible from 14 mm to 6 mm with perfect labial and lingual compensation in /o/ (to maintain an ideal lip opening and degree of pharyngeal narrowing) are illustrated at G in Fig. 7. Mandibular reduction without lingual or labial compensation, with reduced pharyngeal activity and weakened lip rounding is shown at H in Fig. 7.

Figure 4 shows that reduced /o/ preceded by dental consonants in the informant's speech had a higher F2, which is probably due to the tongue blade remaining elevated in /o/. This modification is illustrated at I in Fig. 7.

RESULTS

The results of the model experiments are recorded in Fig. 8.

Full compensation

With full lingual compensation for mandibular variation 14-6 mm (and labial compensation in /o/), the vowel spectra remained in the respective contrastive areas of the nonreduced vowels (A, D, G; compare /a,e,o/ in Fig. 1). This illustrates the extent to which compensated mandibular variation contributes to the normal spectral variation of stressed vowels in speech.

No lingual compensation

The simulated reduction of /a/ (B in Fig. 8) compares well with the /a/ regression in Fig. 4. The wider pharynx, narrower palatal passage and narrower mouth opening resulting from an uncompensated smaller jaw opening shifted F1 and F2 in the direction of the observed regression.

The simulated reduction of /e/ (E in Fig. 8) shifted the spectrum towards an [i]-like vowel, but with a high F2 rather than with the lower F2 exhibited by the informant (Figs. 2, 4).

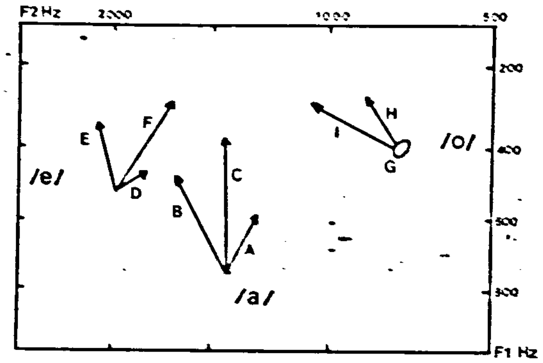


Figure 8
The results of the modelled articulatory modifications (Figs. 5-7), for comparison with the regressions observed in the informant's speech (Fig. 4).

The simulated reduction of /o/ (H in Fig. 8) compares well with the regression for /o/ preceded by velar consonants (Fig. 4).

Weakened lip spreading

Two experiments also assumed weakened lip spreading for the spread-lip vowels /e,a/. The results are recorded at F and C respectively in Fig. 8. Compared with E and B, F2 is lower. Both B and C shift the /a/ spectra towards the /ǎ/ area. For /e/, the weakened lip spreading (F) matches the /e/ regression observed in the informant's speech (Figs. 2,4).

Elevated tongue blade

The final experiment added an elevated tongue blade to the simulated reductions of /o/. The result, illustrated at I in Fig. 8, is a higher F2 compared with H. This compares well with the regression for /o/ preceded by dental consonants (Fig. 4).

DISCUSSION AND CONCLUSIONS

The articulatory behaviour modelled in the experiments successfully reproduced the spectral reduction recorded from the informant. The results demonstrate that the reduction of Bulgarian /e,o,a/ can be explained in terms of neutralization of mandibular depression, of lip spreading or rounding, and of lingual and labial compensation for mandibular variation. This assumes that reduction is a process of subtraction, and not one of substituting a new set of manoeuvres for the reduced vowel. The gradual spectral transition from nonreduced to fully reduced depends on which components happen to be turned down and how far. If there is a hierarchy for neutralization, we suggest the following order: mandibular depression, compensation, lip activity.

The formant frequencies yielded in the experiments do not reproduce those of the informant exactly. To do that, it would be necessary to model the informant's own vocal tract. The results have general interest precisely because the observed spectral tendencies were reproduced by manipulating a set of profiles that were chosen at random.

We propose that the feature that distinguishes the reducing set /e,o,a/ from the nonreducing set /i,ʉ,ǎ/ is the degree of jaw opening, and that the other phonetic correlates of reduction are subsumed. The reducing set are [open] and the nonreducing set [nonopen]. Vowel reduction in Bulgarian will then be captured by the following rule:

$$[+\text{open}] \longrightarrow [-\text{open}] \;/\; \frac{\qquad}{[-\text{stress}]}$$

REFERENCES

(1) T. Pettersson and S. Wood. "Vowel reduction in Bulgarian and its implications for theories of vowel production: a review of the problem". Folia Linguistica, in press.

(2) S. A. J. Wood. "The history of the classical vowel articulation model: a reply to Catford and Fischer-Jørgensen". In this volume.

(3) S. Wood. "A radiographic analysis of constriction locations for vowels". Journal of Phonetics 7:25-43 (1979).

(4) S. A. J. Wood and T. Pettersson. "Vowel reduction in Bulgarian: the phonetic data and model experiments". Forthcoming.

(5) D. Tilkov. Le Vocalisme Bulgare. Publications de la Société de Linguistique de Paris No. 65. Klincksiek, Paris (1970).

(6) T. Pettersson and S. Wood. "A spectrographic study of vowel reduction in Bulgarian". In C. Davidsson et al. (eds), Tionde nordiska slavistmötet, 13-17 augusti 1984. Meddelanden från Stiftelsens för Åbo Akademi forskningsinstitut nr. 102, Åbo (1985).

(7) S. Wood. "Radiographic and model studies of the palatal vowels". Working Papers 23:119-155, Dept. of Linguistics, Lund University (1982). The data for the velar and pharyngeal vowels has not yet been published.

THE AUDITORY FEATURES OF VOWEL AND FRICATIVE PHONEMES

SHIHAB A. SHAMMA

Electrical Engineering Department & Systems Research Center, University of Maryland, College Park, Maryland 20742

Mathematical Research Branch (NIDDK), National Institutes of Health, Bethesda, Maryland 20982

ABSTRACT

The acoustic features of vowels and fricatives are examined in the response patterns of a model of auditory processing. For the vowels, a few harmonics dominate the peaks of the internal representation, reflecting the *formant structure* by their spatial locations, and the *front cavity resonance* by their relative amplitudes. For the fricatives, the most prominant feature extracted is the location of the cut-off frequency in their highpass-like spectra.

Understanding the nature of sound processing in the auditory system is an essential step in determining the acoustic elements of speech sounds and their relevance to perception and articuation. In recent years, important discoveries in peripheral auditory function (both at the basilar membrane/hair cell level [1–3], and from auditory-nerve recordings [4], have facilitated the construction of cochlear models that can adequately replicate the primary response features in the auditory nerve [5]. With such models, it is relatively easy to analyze the response patterns associated with a wide variety of speech sounds, and under many signal conditions. Specifically, it is now possible to generate internal auditory representations of the acoustic spectra, and hence to examine closely the expression of such acoustic features as vowel formant locations, amplitudes, and transitions, and fricative spectral shapes. It is important to note, however, that beyond the peripheral auditory stages, little is known about the central neural networks and the processing they perform on the cochlear outputs. This adds an element of uncertainty to the analysis since apparently useful cues and response features at the auditory nerve level may be irrelevant for phonemic perception and classification if the central nervous system ignores, or is incapable of processing them. We shall address this point further after first illustrating the response patterns to the stimulus /position/ (Fig.1), as generated by a cochlear model.

The peripheral auditory model consists of a linear formulation of basilar membrane mechanics, a fluid-cilia coupling stage which transforms membrane vibrations into hair cell cilia displacements, and a simplified description of the inner hair cell nonlinear transduction of cilia displacements into intracellular electrical potentials. The potentials at each hair cell along the cochlear partition is then taken as a measure of the probability of firing of the nerve fiber innervating it. Many more details of cochlear function can be incorporated in such models, e.g. adaptation at the hair cell/nerve synapse [6], active mechanisms of basilar membrane motion [7], the effects of the middle ear muscles and of the efferent system [8]. This simplified model reproduces the major response properties observed experimentally, especially with relatively steady and broad-band stimuli like vowels and fricatives. The outputs of the cochlear model are computed at 128 equally spaced locations along the cochlear partition, and are all displayed together as a 2-dimensional spatiotemporal pattern representing the ensamble activity of the tonotopically organized array of auditory nerve fibers [9]. The spatial axis is labeled by the characteristic frequency (CF) of each output channel, i.e. the frequency of the tone which produces its maximum final output at that location (see below for further details of the central processing of the cochlear outputs).

The responses to the vowel portions of the stimulus (/I/,/u/) posses a typical structure that is observed in all experimental data [10,11] - that is the dominance of the entire pattern by a few stimulus harmonics. These harmonics correspond to the largest components located near the formants of the stimulus spectrum. They excite travelling waves along the basilar membrane which are evident in the fine temporal structure and spatial spread of the responses. Because of the unique asymmetrical shape of the cochlear filters, the waves decay in amplitude, and begin to acumulate phase rapidly at locations along the array, depending on the frequency of the underlying harmonic [9]. The expression of these features progressively deteriorates as the harmonics become spatially less segregated (less resolved) and begin to interfere (e.g. the responses at the CF's of the higher harmonics). For each of the vowel responses, the identity of the underlying dominant harmonics can be deduced from two sources: (1) The temporal course of the response (e.g. by measuring the frequency of the synchronized response), or (2) by the location of the above described features along the spatial axis (i.e. a tonotopic axis). Thus for the vowel /I/ (Fig.1), there are two response domains, the first is the apical region CF ≤ 2 kHz, corresponding to the F_1 harmonics (2-3), each of which decaying and experiencing phase shifts at its appropriate CF location. An abrupt transition in the response patterns occurs at approximately 2 kHz as the harmonics associated with the higher formants become dominant. These trends are seen again in the /u/ vowel responses, where F_1 is at a lower CF location (≈ 250 Hz) and F_2 is considerably weaker.

The auditory responses of the fricative portions /z/,/š/ differ considerably from those of voiced vowels. To start with, there is a random component in the excitation that is quite evident in the cochlear responses as randomly initiated travelling waves. Another distinctive aspect is the predominance of the high frequency components and their sudden decay at a different location for each of the fricatives. For the voiced fricative /z/, there is an additional voiced component in the excitation waveform.

The CNS derives its auditory percepts from the cues available in the spatiotemporal outputs of the cochlea. The identity of these cues and the way they may be extracted and processed are two issues that are essentially inseperable. In the cochlear patterns, there is an abundance of spatial and temporal cues to the physical parameters of the stimulus [12]. However, given the complexity of the extraction algorithms involved, only a subset of of these cues are probably relevant in that the CNS is actually capable of utilizing them. Since little is known about the anatomical and functional role of the neural networks of the central auditory system, little can be said in support of any processing algorithm aside from the general plausibility arguments regarding its biological implementation and the degree to which the isolated parameters explain the psychophysical measurements [13].

In viewing the cochlear outputs as spatiotemporal images, a set of cues emerge that are robust and particularly easy to extract. These are the spatial edges due to the asymmetrical shape of the cochlear filters [9]. As noted earlier, such edges occur at the regions separating the responses to the strong, resolved components of the stimulus. While the expression of these edges is dependent on the integrity of the phase locked reponses (i.e. the ability to visualize regions of different temporal character), they can also appear in the high frequency regions (where phase locking is minimal) as the peaks and valleys of the spatial average rate profiles. In all cases, the *location* of these edges along the tonotopically organized spatial axis, and their saliency, are reliable indicators of the stimulus frequency and amplitude [13]. As with normal visual images, such spatial discontinuities can be detected and highlited by relatively common and simple neural networks as the lateral inhibitory networks (LIN) [14].

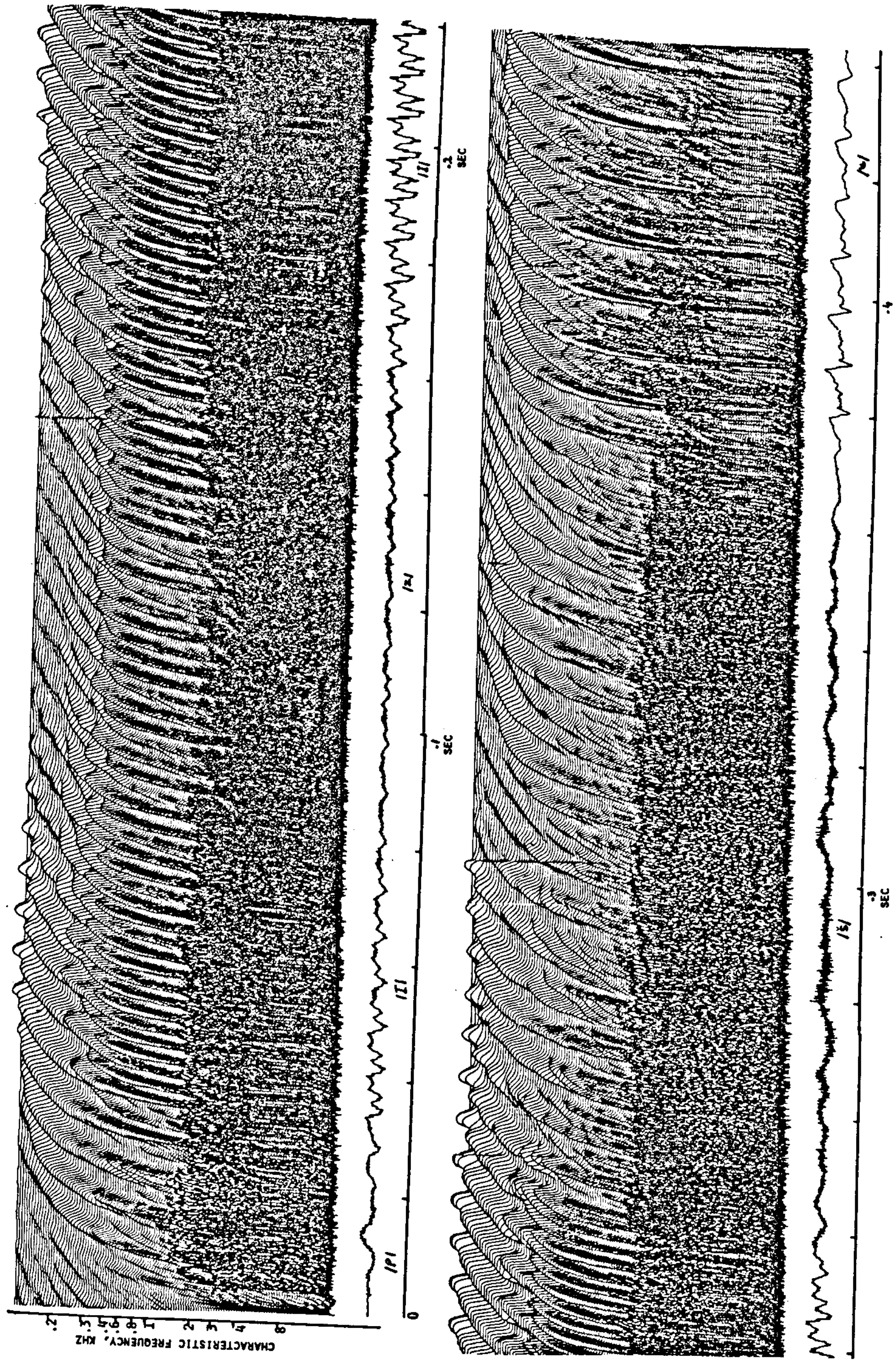
We have processed the cochlear patterns of a wide variety of sounds with models of recurrent and nonrecurrent LIN's [13,15]. The results shown in Fig.2 for part of the word /position/ and in Fig.3 for a vowel series, are generated with a two layer LIN: the first nonrecurrent and performs the initial edge detection and extraction, the second is a recurrent version which further sharpens the outputs of the first layer and preserves only locally large peaks [15]. For the vowel portions of the stimulus, the LIN's typically extract two or three peaks corresponding to the components near the nominal formant frequencies of the vowels; an additional low frequency peak sometimes appears corresponding to the fundamental or second harmonic components of voiced sounds (especially for females). The variability in the locations of these peaks for different speakers and sexes seems to be similar to that observed in traditional spectrogram outputs [15], though this remains to be confirmed with much larger data samples. An interesting aspect of these and other vowel outputs [15] is the systematic change in the *relative amplitudes* of the high-CF and low-CF peaks (or equivilantly, the location of *center of gravity* of the pattern) for different vowels (Fig.3). Thus, for high vowels (e.g. /i,u/), the high-CF peak is always relatively large when the constriction is fronted (as in /i/), and vice versa in the back vowel /u/. In all close vowels (e.g. the frontal /i,y/ and back /u/), the place of the constriction seems to be the primary factor in determining the overall weight distribution of their outputs. Lip rounding seems to have only a secondary effect, increasing slightly the relative size of the higher CF peaks. The open vowels /æ/ and /ɔ/ occupy an intermediate position in that the two peaks are comparable.

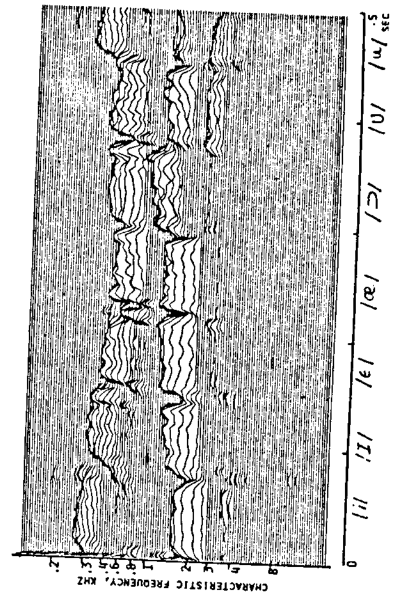
These relations are summarized schematicaly in Fig.4. On the left, the vowels are organized along a continuum in the plane of A_1,A_2 - the relative amplitudes of the low and high-CF peaks respectively. The small arrows indicate the effects of lip-rounding. The figure on the right illustrates the organization of the same vowels on the plane of two articulatory features: The open-close axis reflecting tongue height, and the front-back axis indicating the position of the constriction. These two figures are closely related, in that the vowel continuum in the A_1,A_2 plane (left) can be thought of as the continuum that would result if we project the vowels in the articulatory plane unto the front-back axis. Since movement along the latter axis correlates well with the length of the front cavity, the organization of the vowels in the A_1,A_2 plane (i.e. the relative height of the LIN peaks) may also reflect the effects of the position (frequency) of the 'front cavity resonance' and the so-called F_2' [16], which also move in the same direction for this sequence of vowels [17]. Finally, the effects of lip-rounding in this schematic are viewed only as local modulations (in the direction of the arrows) of the parameters already established by the articulatory features. Therefore, it is possible to reach the same point along the vowel continuum of the left figure with different combinations of lip-rounding and front-back articulations [17].

Correlates of the pitch percept associated with voiced

Fig.1

The spatiotemporal response patterns of a cochlear model to the word /position/. The spatial axis represents the basal-to-apical (bottom-to-top) spread of the cochlear partition; It is labeled by the Characteristic Frequency (CF) of each output channel (see text). The scale marks on the time axis = 20 msec.





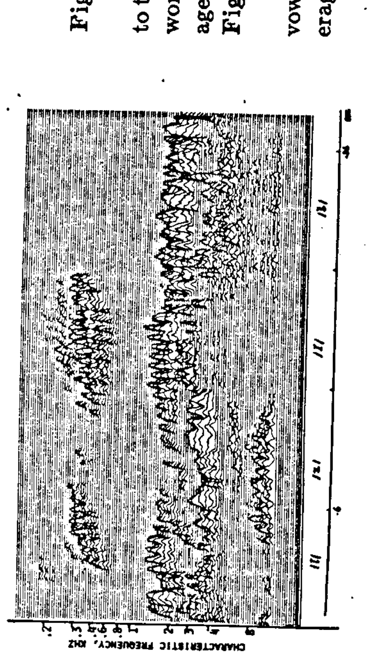


Fig.2
The LIN outputs corresponding to the spatiotemporal patterns of the word /position/. The moving average window is 2 msec wide.

Fig.3
The LIN outputs of a series of vowels as indicated. The moving average window is 10 msec wide.

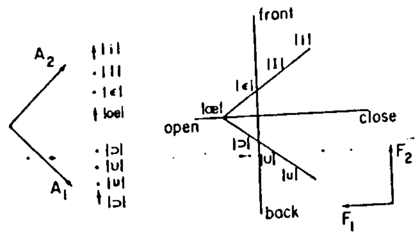


Fig.4
A schematic of the relationship among vowel model parameters (left) and articulatory features (right).

vowels can also be descerned in the LIN responses. In Fig.2 outputs, this is seen in the *beating* of the LIN peaks at the voiced portions of the stimulus[1]. The origin of this temporal character is the combining by the LIN of locally dissimilar waveforms at the regions of discontinuities in the cochlear patterns. These responses are due to different *resolved* harmonics of the same fundamental, and hence beat at this frequancy [15]. As expected, in the LIN outputs of unvoiced fricated speech (e.g. /š/, and the high CF region of /z/) the regular beating is absent.

The LIN outputs (Fig.2) of the fricatives show major peaks that correspond to the most important discontinuity in the spatiotemporal patterns, i.e. the edge created by the rapid cut-off of their high frequencies (Fig.1). The downward CF shift of this peak from that of /z/ to /š/ reflects the lengthening of the frontal cavity which largely determines the high frequency extent and overall spectral shape of the fricative [18].

In summary, auditory processing of speech phonemes isolates specific features that may play an important role in the perception and recognition of these sounds. These auditory features can be related to articulatory aspects such as the formant resonances of vowels and the front cavity resonances of fricatives and vowels. They also contain cues to other attributes of the speech signal, e.g. pitch.

Acknowledgment: This work is supported in part by an NSF initiation award NSFD ECE-85-05581 and NSF Grant CDR-85-00108.

REFERENCES

[1] A. J. Hudspeth & D. P. Corey, "Sensitivity, polarity, and conductance change in the response of vertebrate hair cells to controlled mechanical stimuli," *Proc. Nat. Acad. Sci. U.S.A.* 74(6) (1977), 2407–2411.

[2] I. J. Russell, "Origin of the receptor potential in inner hair cells of the mammalian cochlea - evidence for Davis's theory," *Nature* 301(27) (1983), 334–336.

[3] R. Patuzzi & P. M. Sellick, "Basilar membrane motion and inner hair cell output," *J. Acoust. Soc. Am.* 74(6) (1983), 1734–1741.

[4] M. B. Sachs & E. D. Young, "Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate," *J. Acoust. Soc. Am.* 66 (1979), 470–479.

[5] S. A. Shamma, "Encoding the acoustic spectrum in the spatio-temporal responses of the auditory-nerve," in *Auditory Frequency Selectivity*, B. C. J. Moore & R. Patterson, eds., Plenum Press, Cambridge, 1986, 289–298.

[6] L. A. Westerman & R. L. Smith, "Rapid and short term adaptation in auditory nerve responses," *Hear. Res.* 15 (1984), 249–260.

[7] S. T. Neely & D. O. Kim, "An active cochlear model shows sharp tuning and high sensitivity," *Hearing Res.* 9 (1982), 123–130.

[8] R. Winslow, "A quantitative analysis of rate-coding in the auditory nerve," Ph.D. Dissertation, Johns Hopkins University , 1985.

[9] S. A. Shamma, "Speech processing in the auditory system. I: Representation of speech sounds in the responses of the auditory-nerve," *J. Acoust. Soc. Am.* 78 (1985), 1612–1621.

[10] E. D. Young & M. B. Sachs, "Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.* 66 (1979), 1381–1403.

[11] D. G. Sinex & C. D. Geisler, "Responses of auditory-nerve fibers to consonent-vowel syllables," *J. Acoust. Soc. Am.* 73 (1983), 602–615.

[12] J. O. Pickles, "The neurophysiological basis of frequency selectivity," in *Frequency Selectivity in Hearing*, B. C. J. Moore, ed., Acadamic Press, London, 1986, 51–122.

[13] S. A. Shamma, "Speech processing in the auditory system. II: Lateral inhibition and the processing of speech evoked activity in the auditory-nerve ," *J. Acoust. Soc. Am.* 78 (1985), 1622–1632.

[14] H. K. Hartline, *Studies on excitation and inhibition in the retina*, Rockefeller University Press, New York, 1974.

[15] S. A. Shamma, "The acoustic features of speech phonemes in a model of auditory processing: Vowels and unvoiced fricatives," *J. Phonetics* (1987 (in press)).

[16] R. Carlson, B. Grantstorm & G. Fant, "Some studies concerning perception of isolated vowels," STL-QPSR, 1970.

[17] G. M. Kuhn, "On the front cavity resonance and its possible role in speech perception," *J. Acoust. Soc. Am.* 58(2) (1975), 428–433.

[18] C. G. Fant, "Acoustic description and classification of phonetic units," in *Speech Sounds and Features*, MIT, Cambridge, MA, 1973.

[1] The voicing in the /z/ segment is clearly visible in the Fig.1 responses to the 3rd harmonic component of the fundamental (same as the F_1 harmonic of the preceeding and succeeding vowel) and will be clearer in the LIN output of a slightly louder stimulus. In Fig.3 the voiced vowel outputs also beat, but the LIN moving average window is set at 10 msec in order to clarify the display of the relative amplitudes; This in turn averages out the beating.

EVALUATION OF DISTANCE METRICS USING SWEDISH STOP CONSONANTS

Diana Krull

Institute of Linguistics Department of Phonetics
University of Stockholm S-106 91 Stockholm Sweden

ABSTRACT

In recognition algorithms and certain theories of speech perception the process of signal interpretation is modeled in terms of distance metrics comparing the signal with stored references. In order to evaluate such metrics, listening tests were performed. The stimuli were short (about 26ms) fragments derived from the consonantal release of Swedish $\grave{V}_1C{:}V_2$ "words". A stop (b,d,ḍ,g) appeared in a systematically varied context of phonologically short vowels (i,e,a,o,u). The test yielded confusions which appeared to make qualitative sense in terms of the acoustic properties of the stimuli.

The spectrum level of the stimuli was measured at two time points after the stop release. Euclidean distances were calculated using spectra derived by means of 1/4 octave filter analyses. Two kinds of distances were calculated: static, based on spectra sampled at the first time point, and dynamic, based on the differences in spectral change between the two samoling points. Linear regression analyses performed on symmetrized percent confusions versus stimulus-reference distance produced correlation coefficients of -.85 (static), -.83 (dynamic), and -.92 (static and dynamic combined.)

INTRODUCTION

This investigation is based on the conception of a perceptual space for speech sounds where the distance between different sounds reflects the degree of their perceptual similarity. The greater the similarity between two sounds, the smaller the distance between them. Similar sounds tend to be confused with each other, therefore the number of confusions between sounds can be used as a measure of their perceptual distance. A further assumption is that correct identification of a sound indicates minimal distance from a stored reference.

For both theoretical and practical reasons, it is often desirable to be able to predict perceptual similarity from acoustic data. Such predictions are important especially in automatic speech recognition. To implement such a model, it is necessary, on the one hand, to find a realistic transformation of the speech signal, e.g. in terms of a realistic auditory model, and, on the other hand, an empirically calibrated distance metric.

ELICITATION OF PERCEPTUAL CONFUSIONS

The aim of this study is the evaluation of such a prediction model for Swedish voiced stops. It has been shown for Swedish /1/ that there are considerable coarticulation effects for such stops in intervocalic position. To make use of these effects, stimuli of the form $\grave{V}_1C{:}\acute{V}_2$ were prepared, where the consonant was [b,d,ɖ,g] and the vowel [ɪ,ɨ,a,ɔ,ʊ]. The resulting one hundred nonsense words were read in random order by a male speaker of the Central Swedish dialect. The Swedish grave accent was used in order to give both syllables about equal prominence.

From these "words" shorter stimuli were prepared by cutting out ca 26ms long segments beginning at consonant release. For simplicity, these stimuli will henceforth be referred to as "Burst" although they can contain also the beginning of the vocalic transitions. Notwithstanding the fact that the duration of the noise burst varies with place of articulation, all stimuli were given the same length in order to avoid letting stimulus length constitute an extra place cue.

A tape was prepared where each "Burst" stimulus appeared three times. The order of the stimuli was randomized. 20 native speakers of the Central Swedish dialect listened to the tape, their task being to identify the consonant.

The results of the perception test are shown in 25 confusion matrices, one for each vowel context (Fig.1). In each row of matrices the preceding vowel changes from front to back while for each column of matrices it is the following vowel that changes in the same manner. Comparing the results by vowel contexts and consonants,

PERCENT ANSWERS

STIMULI

	I-I				ɛ-I				a-I				ɔ-I				ʊ-I			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	93	3	2	2	98		2		82	3	7	8	67	8	2	23	86	10	2	2
d		97	3		2	88	8	2	13	70	17		3	67	28	2	7	87	3	3
ɖ		58	40	2	2	33	63	2		32	68		3	33	62	2	·	42	58	
g		25	28	47	8	32	32	28	5	28	23	44	10	32	15	43	2	53	22	23

	I-ɛ				ɛ-ɛ				a-ɛ				ɔ-ɛ				ʊ-ɛ			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	98	2			90	2	5	3	90	2	8		91	2	5	2	89	3	5	3
d		87	10	3		78	20	2	5	57	28	8	2	76	20	2		87	13	
ɖ		32	68			20	80			28	70	2	2	28	65	5		25	73	2
g	8	38	22	32	17	35	13	35	2	27	18	53	7	23	27	43	10	20	15	55

	I-a				ɛ-a				a-a				ɔ-a				ʊ-a			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	97			3	97			3	98			2	91	2		7	97		3	
d	2	75	18	5	20	61	17	2	2	75	23		8	66	23	3	10	63	27	
ɖ		40	52	8	2	28	68	2		15	83	2	15	7	75	3		13	85	2
g	3	15	12	70	3	18	23	56	5	15	2	78	10	17	13	60	28	23	13	36

	I-ɔ				ɛ-ɔ				a-ɔ				ɔ-ɔ				ʊ-ɔ			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	90	2	5	3	-	-	-	-	92		5	3	91	2	5	2	85	2	5	8
d	3	48	37	12	3	55	35	7	5	40	50	5	2	55	40	3		48	35	17
ɖ	7	15	75	3		10	88	2	3	8	89		2	18	80		13	15	62	10
g	15			85	13		3	84	25		7	68	37	2	3	58	42		2	56

	I-ʊ				ɛ-ʊ				a-ʊ				ɔ-ʊ				ʊ-ʊ			
	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g	b	d	ɖ	g
b	92		3	5	92		5	3	98			2	96		2	2	85	5	5	5
d		63	32	5	8	57	17	18	5	76	17	2	10	42	25	23	3	37	30	30
ɖ	2	20	78			18	79	3	33	13	32	22	13	8	74	5	5	23	59	13
g	18	2	2	78	32			68	27		5	68	12		2	86	63			37

Fig.1. Confusion matrices for "Burst" stimuli in 25 vowel contexts.

it can be seen that the confusions form a regular pattern. For example, [g] in front vowel context was often confused with the dental and the retroflex, but seldom with the labial. In back vowel context, on the other hand, the velar was often confused with the labial but almost never with the dental or the retroflex. The consonants seem to have been easiest to identify in the context of /a/. The influence of the preceding vowel was less pronounced than that of the following one. (For more details see /2/). Perceptually, the distance between the velar and the dental is thus small in front vowel context and large in back vowel context, while the reverse is true for the pair labial-velar.'

USING PHYSICAL DISTANCE MEASURES TO PREDICT THE PERCEPTUAL CONFUSIONS

A qualitative comparison of stimulus spectra showed that there are pronounced coarticulation effects and, also, that these can have influenced the direction and number of the confusions. With such effects in mind, three models were chosen for defining the acoustic distances to be correlated to the perceptual confusions. The first model was based on formant frequencies at the moment of consonant release, and the second on sone-Bark spectra /3/.

The third model was based on bandpass filtered spectra sampled at two points in time: t_1, integrated over the first 10ms after consonant release, and t_2, 10ms later. The measurements were carried out with 14 digital 1/4 octave filters, covering a frequency range from about .4kHz to about 4.5kHz. The measured sound pressure levels were plotted as a function of frequency. The resulting spectra showed similarities and differences not only according to the place of articulation of the consonant but also according to the following vowel, thus forming 12 groups: labial, dental, retroflex, and velar stops read in in the context of a following front vowel, /a/, and back vowel. Differences within groups being small, mean values were calculated for each group, both at t_1 and at t_2. The t_1 spectra were normalized with respect to their mean SPL in order to avoid including differences in overall intensity into the distance measure. Two examples of the resulting spectra are shown in Fig.2.

Distances between spectra were then calculated for t_1, the result was called "static" distance, using the Euclidean metric:

$$\mathrm{Dstat}_{i,j} = \sqrt{\sum_{n=1}^{14} \left|L_{i,n} - L_{j,n}\right|^2} \qquad \mathrm{Eq(1)}$$

$\mathrm{Dstat}_{i,j}$ = the distance between stimuli i and j at time t_1
$L_{i,n}$= the level in band n

The changes in spectrum level that occur after stop release show characteristic differences with place of articulation. These dynamic differences have in recent years been investigated especially in connection with the question of acoustic invariance for stop consonants /4/.

Comparing the change in spectrum level of the twelve spectra, it could be seen that at low frequencies the spectrum level rises during the interval between t_1 and t_2 for all spectra, and is comparatively steady at 1.5kHz. It is at frequencies above 1.5kHz that the amount and direction of the change varies in a systematic way: before front vowels the level goes up for the labial, remains unchanged for the dental, and drops for the retroflex and the velar. Before /a/ the level also rises for the labial, but in contrast to the front vowel context, the level drops for both the dental and the retroflex but is stable for the velar. In back vowel context the level remains stable for the labial, while dropping with all other places of articulation, although the drop is comparatively small for the velar. It thus seems that although the change in level is dependent on place of articulation, the following vowel must be taken into account too. There tends to be less change if the spectra of the consonant and the following vowel are relatively similar as is the case for the dental and front vowels, for the velar and /a/ and for the labial and back vowels.

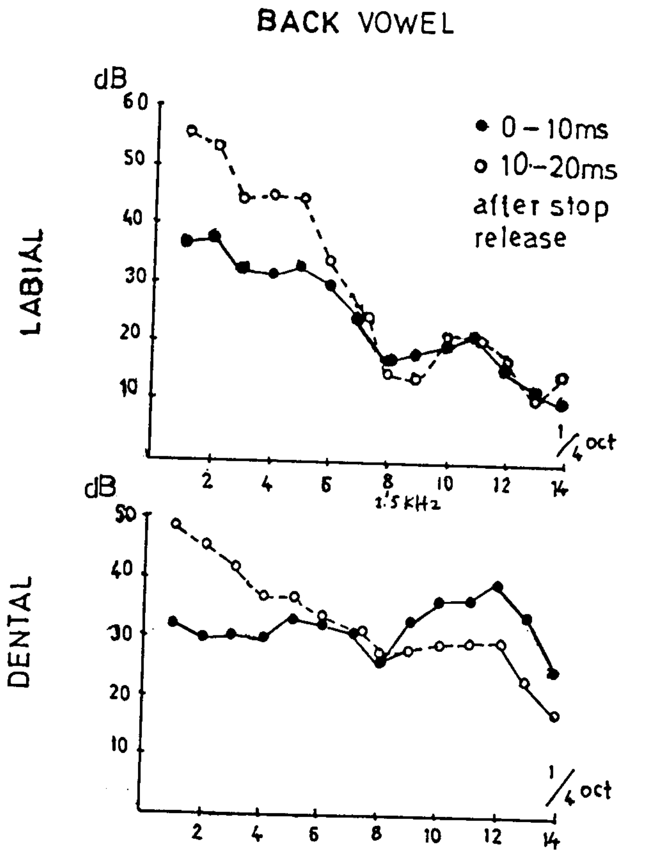


Fig.2. Examples of spectra sampled at two time points after stop release. 1/4 octave band-pass filters were used.

The dynamic distances were calculated in a similar way as the static ones with the help of the Euclidean metric, but on non-normalized spectra and only for the six filter bands above 1.5kHz, that is in the frequency range where there were systematic differences between groups:

$$\mathrm{Ddyn}_{i,j} = \sqrt{\sum_{n=1}^{6} \left|C_{i,n} - C_{j,n}\right|^2} \qquad \mathrm{Eq(2)}$$

where $\mathrm{Ddyn}_{i,j}$= difference in level change in dB from t_1 to t_2 between stimuli i and j
$C_{i,n}$= level change for stimulus i, band n

Before performing regression analyses correlating acoustic distances and perceptual confusions, the results of the perception test were manipulated in two ways: first, the answers were divided into 12 groups in the same way as the spectra and mean values were calculated for each group; second, the answers were symmetrized according to a method described by Klein, Plomp, and Pols /5/. The regression analyses were then calculated between the symmetrized confusion data and three kinds of acoustic measures: (1) static, i.e. difference between spectra at t_1; (2) dynamic, i.e. difference in the amount and direction of change in two spectra; (3) static and dynamic distances combined according to the equation

$$D_{i,j} = \sqrt{(\mathrm{Dstat}_{i,j})^2 + (\mathrm{Ddyn}_{i,j})^2} \qquad \mathrm{Eq(3)}$$

where $D_{i,j}$= combined static and dynamic distance between stimuli i and j

The resulting correlation coefficients are shown in the table below.

	Front vowel	/a/	Back vowel
r(tl) -- static:	-.78	-.93	-.55
r(t2-tl) -- dynamic	-.78	-.94	-.14
static + dynamic	-.80	-.98	-.58

It can be seen that good predictions can be made only for consonants before the vowel /a/. The results were especially negative for the back vowel context. What could be the reason for this? A possible answer could be that the listeners, if they could not recognize the following vowel, used a strategy somewhat different from that assumed here. Even if we are correct in assuming that a comparison of the stimulus with a stored reference does indeed take place in the listeners' processing, we might be wrong in supposing that the stored reference is the spectrum actually associated with the specific VCV word from which the stimulus had been derived. Conceivably, a given stimulus might lead the listener to postulate a "neutral" vowel reference spectrum from a "neutral" vowel context in cases where cues for V_2 were weak or absent. In order to obtain information on these questions, an additional test was carried out with the "Burst" stimuli using eight subjects, their task now being to identify the vowel. The results showed, first, that a back vowel could be identified only after a labial or velar consonant. After a dental or a retroflex listeners heard either a front or a neutral vowel. When the original vowel was a front vowel or /a/, listeners either made few errors or heard a neutral vowel.

With the above considerations and the preceding results in mind, acoustic distances for all stimuli (except labials and velars before back vowels) were calculated using consonants read before /a/ as references. The new correlation coefficients are shown below.

r(1) -- static

Front vowel	/a/	Back vowel	Contexts pooled
-.89	-.93	-.96	-.85

r(t2-t1)--dynamic

Front vowel	/a/	Back vowel	Contexts pooled
-.94	-.94	-.72	-.83

static+dynamic

Front vowel	/a/	Back vowel	Contexts pooled
-.96	-.98	-.89	-.92

References

/1/ Öhman, S. (1966): "Coarticulation in VCV utterances: Spectrographic measurements", JASA 39(1), 151-168

/2/ Krull, D. (1984): "The role of vowel context on the perception of place of articulation of stops", PERILUS, Report III, University of Stockholm

/3/ Krull, D. (1985) "On the relation between the acoustic properties of Swedish voiced stops and their perceptual processing", PERILUS, Report IV, University of Stockholm

/4/ Kewley-Port, D. (1983): "Invariant cues for place of articulation in stop consonants", JASA, 73(1), 322-335

/5/ Klein, W., Plomp, R., and Pols, L.W.C. (1970): "Vowel spectra, vowel spaces and vowel identification", JASA 48(8), 999-1009

A CASE FOR GLOBAL LISTENING STRATEGIES

J.C.T. RINGELING

Dept. of English
Utrecht University
Oudenoord 6
3513 ER Utrecht, The Netherlands

W. EEFTING

Dept. of Phonetics
Utrecht University
Trans 14
3512 JK Utrecht, The Netherlands

ABSTRACT

A case is made for global perceptual strategies. In poor listening conditions subjects appear to be able to perceive and comprehend elliptic speech, albeit with some difficulty. If sufficient semantic information is available, they seem capable of basing themselves on global characteristics in speech sounds, particularly on information related to place of articulation. The study pleads for the formulation of perceptual features to obtain a better insight into the processes operative in speech perception.

1. INTRODUCTION

When Zue (1) showed that a trained spectrogram reader can recognize a substantial number of words from spectral information, the discussion on invariant features in speech perception gained new ground. In the seventies many linguists did not take invariance very seriously, although some invariant features were generally accepted (see e.g. 2 and 3). Naturally, Zue´s success in reading spectrograms was partly caused by extensive use of linguistic expectancy to solve ambiguities, but it made clear that some sort of invariance must be present in speech, although perhaps of a different nature than had traditionally been accepted in terms of linguistic features.

Carlson, Elenius, Granström and Hunnicut (4) and more recently Veenhof and Bloothooft (5) have shown that it may theoretically be possible in many cases to come a long way in arriving at word identification by specifying the acoustic information on the basis of broad phonetic categories. They showed that a classification of phonemes into global categories such as plosives, nasals, fricatives, remaining sonorant or vowel, often provides sufficient information to limit the number of words in a cohort for recognition to take place. This insight that word recognition may be feasible on the basis of a broad phonetic classification has proved helpful in automatic speech recognition (6,7).

However, it is by no means certain if human perception can adequately use a broad phonetic classification in the process of listening to connected speech, and if so, it remains questionable whether listeners base themselves on the same phonetic categories as are frequently adopted in theoretical studies. If we wish to find out what phonetic underlying features can be used in human perception, it is imperative that listening tasks are constructed which vary the amount of acoustic information along global phonetic parameters. An attempt at such a task was an informal study by Ringeling (8), who demonstrated that Dutch listeners could fairly successfully identify sentences in which all consonants had been replaced for consonants that were similar with regard to place of articulation, in such a way that the phonotactic constraints of Dutch were not violated. The use of elliptic speech (see e.g. 9) thus served to manipulate the amount of acoustic information in the speech signal. An English example in ordinary orthography would be the conversion of the saying: ´no place like home´ to ´mow crafe wipone´. The resulting sentences sounded Dutch, but could not readily be understood. However, when redundancy of the acoustic signal was reduced by adding noise to the sentences, it turned out that listeners produced much better recognition scores on the same material. One of the most interesting findings of this study was that subjects were rarely aware of the manipulations that had been carried out. This suggests that a global phonetic analysis had taken place on the basis of similarity of place of articulation. In view of the task at hand, which drew heavily on an intensive use of linguistic expectancy, sentences with constraining context were understood much better than those with relatively neutral content.

Van der Woude (10) based a study on this idea. He investigated the theoretical possibility of arriving at unique identifiability of words by grouping consonants together, either on the basis of manner of articulation, or on the basis of place of articulation. On the basis of a random sample of 100 words from 68,000 word tokens (12,000 word types), he found that specification of Dutch words in terms of broad phonetic classes thus defined, did not yield a clear theoretical advantage to either classification. In his definition of patterns, leaving vowel-quality intact, he found

72 % unique identifiability for grouping consonants on the basis of changes in place of articulation and 78 % for grouping them together on the basis of changes in manner of articulation. Moreover, for those words that could theoretically not be identified uniquely, the remaining cohort of word candidates from a 12,000 wordtype lexicon never exceeded four and was only two in 80 % of the instances.

Yet, even if theoretically both types of classification would seem to qualify as potential approaches for listening strategies, it is evident that actual speech perception need not avail itself of these theoretical possibilities. In fact, it would seem highly unlikely that both strategies are equally effective, since it is well-known from the literature that perceptual confusions on the basis of changes in place of articulation are much more frequent than those on the basis of manner of articulation (see e.g. 11). It was therefore decided to undertake a preliminary study into the perceptual relevance of global phonetic listening strategies on the basis of place-changed and manner-changed consonants.

Miller and Isard as early as 1963 (12), showed that listeners can extract the linguistic content of a message if they have access to normal syntactic and semantic information when speech is presented under high levels of noise. If this linguistic information is also deteriorated, the listeners identification will suffer accordingly. We therefore expect that in the experiment reported on here, sentences with high semantic constraints will be identified correctly more often than neutrally constrained sentences. Moreover, on the basis of what was stated above, we will expect to find a discrepancy in recognition scores based on the amount of phonetic information. If the place-changed and manner-changed consonants lead to unique word patterns, better recognition scores are expected than if the resulting word patterns leave room for ambiguities.

2. METHOD

2.1 Stimuli

21 Sentences were synthesized using the diphone synthesis system, developed by Elsendoorn (13). By using diphones it was possible to preserve a natural flow of speech while changing the consonants at will. Each sentence was synthesized in three conditions:
place-changes: all consonants were systematically replaced for other consonants differing in place of articulation, in conformity with the phonotactic rules of Dutch. The feature voiced/unvoiced in these elliptic sentences remained unaffected.
manner-changes: idem, but differing in manner of articulation.
control: these were stimuli syntesized without manipulation of consonant features.

Three types of sentences were constructed:
sentences consisting of short words (non-unique pattern in terms of global perceptual categories) and neutrally constrained,
sentences consisting of long words (unique pattern in terms of global perceptual categories) and neutrally constrained,
proverbs/sayings, semantically highly constrained.

In corresponding sentences in the three conditions, overall intensity and intonation were kept identical. All sentences were masked with noise at an S/N-ratio of -6 dB, which had resulted in a 90 % correct recognition score of the ´control´ sentences in a pilot experiment. Noise was turned on 1 second before the signal started and turned off .5 s after the speech signal had ended.

2.2 Subjects

21 native speakers of Dutch, aged 20 to 30, served as unpaid participants. No subject reported hearing defects. They were members of staff or students at Utrecht University. Some were phonetically trained, but none were familiar with the stimuli or the aims of the experiment.

2.3 Procedure

In a sound treated room, subjects listened to 3 trial sentences and 18 target sentences. The stimuli were presented binaurally over headphones at a comfortable listening level, using a Revox tape recorder. Each sentence was repeated after a 1 second interval. Items were preceded by a short 200 Hz tone. After each pair of sentences there was an interval of circa 13 seconds to give subjects time to write down their responses. Subjects were encouraged to write down partial responses as well, even if those consisted of separate sounds, fragments of sentences that seemed anomalous etc. Each subject heard each sentence only in one condition to prevent learning effects.

3. RESULTS

Reactions from the subjects and the amount of missing data (63 % of the sentences, 35 % of the content words) indicated that the task was considered quite difficult. In some instances subjects were aware that the material had been manipulated.

In table I the number of correctly identified content words is presented. The condition MANNER-CHANGED was by far the most unintelligible. On average only 3 % of the words were reported correctly. For neutrally constrained sentences in the PLACE-CHANGED condition circa 10 % of the words were identified correctly. It is in this condition that the powerful influence of linguistic constraints can most clearly be observed. In highly constrained sentences over 50 % of the words were recognized. No differences are found within conditions with respect to the type of words presented. Apparently greater word-length, and consequently a higher degree of uniqueness of the word pattern, did not facilitate recognition.

<u>Table I</u>: Number of word responses, subdivided into words reported correctly, incorrectly and failure to respond, for neutrally and highly constrained sentences, in the experimental conditions CONTROL, PLACE-CHANGED and MANNER-CHANGED

HIGHLY CONSTRAINED

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	141	152	141
Correct	127 (90%)	79 (52%)	5 (3%)
Incorrect	3 (2%)	28 (18%)	64 (45%)
Missing	11 (8%)	45 (30%)	72 (52%)

NEUTRALLY CONSTRAINED

UNIQUE WORD-PATTERN

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	160	146	148
Correct	119 (74%)	12 (8%)	4 (2%)
Incorrect	14 (9%)	35 (24%)	27 (19%)
Missing	27 (17%)	99 (68%)	117 (79%)

NON-UNIQUE WORD-PATTERN

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	121	120	119
Correct	90 (74%)	12 (10%)	5 (4%)
Incorrect	17 (14%)	43 (35%)	43 (35%)
Missing	14 (12%)	65 (55%)	71 (61%)

From the data it appeared that correct sentence recognition in the PLACE-CHANGED and MANNER-CHANGED conditions was rare. 90 % of the control sentences were reported correctly when the context was highly constrained. In neutrally constrained sentences this percentage was circa 50 %. For the manipulated versions correct sentence recognition was always below 5 %, except for highly constrained sentences in the PLACE-CHANGED condition, which obtained a 34 % correct recognition score.

4. DISCUSSION AND CONCLUSIONS

In this experiment we hoped to learn something about the type of acoustic and non-sensory information listeners may employ when poor listening circumstances force them to use a global perceptual analysis. Because of the preliminary nature of the experiment, our conclusions can only establish promising areas for further research:

a. Changing manner of articulation does not appear to be a salient characteristic in the identification of spoken sentences.

b. Changing place of articulation appeared to yield satisfactory results in case sufficient linguistic constraints were available. Subjects´ comments indicated that the message can be reconstructed properly and phonetic distortions mostly go unnoticed.

c. Word structure did not turn out to help the listeners in identifying the words correctly, although subjects did attempt to respond to uniquely patterned words more frequently than to non-uniquely patterned words, as can be seen from the percentages of incorrectly identified words. It may well be that uniqueness of word-pattern plays a more salient part if stimulus material is presented in which word boundaries are better available to the listener.

Although the outcome of the experiment clearly shows that PLACE-CHANGED manipulation plays a more important part than MANNER-CHANGED manipulation, the actual recognition scores remain disappointingly low if linguistic constraints are weak. It should be kept in mind, however, that our quest for global perceptual features was hampered by the choice of synthesized material. We did not synthesize plosive-like sounds or nasal-like sounds, but used substitutions of existing phonemes. This means that the listeners were purposely deluded. In view of this, the outcome of the experiment is quite promising . It may well be possible to arrive at core-features underlying perception in the future.

These features may be rather different from what we have traditionally used in articulatory or linguistic terminology. It is, for instance, noteworthy that in multidimensional scaling techniques, when applied to perceptual studies, the dimensions often do not correspond to traditional feature classifications. Similarly, in studies on broad phonetic classifications (such as 4 and 5) non-traditional as well as traditional ´features are used.

We find it important that research should be carried out into the perceptually salient features so as to arrive at a set of variables that are of primary importance to speech perception. The variables in use now, are sometimes haphazard and only used ´because they appear to work´. If we obtain a better understanding of the fundamentally important variables in speech perception, many issues may become more accessible. Notice in this respect that Van der Woude (6) found no theoretical reason to prefer a classification based on PLACE-CHANGED consonants to one that was based on MANNER-CHANGED consonants. But actual perceptual strategies evidently favour a PLACE-CHANGED approach. Nevertheless, we are by no means certain yet, if a PLACE-CHANGED categorization is the best possible approach in global listening strategies. It would be highly counterintuitive if this was not the case, but we will need to lay bare the fundamental features of speech perception first.

REFERENCES:

1. V.W. Zue (1983) Proposal for an Isolated Word Recognition System Based on Phonetic Knowledge and Structural Constraints, in A. Cohen and M. Van den Broecke (eds.): *Abstracts of the Tenth International Congress of Phonetic Sciences*, Foris, Dordrecht, The Netherlands: pp. 299 - 305.
2. K.N. Stevens (1975) Potential Role of Property Detectors in the Perception of Consonants, in G. Fant and M.A.A. Tatham (eds.): *Auditory Analysis and Perception of Speech*, Academic Press, New York, London.
3. M. Umeda (1977) Consonant Duration in American English, *JASA* 61: pp. 846 - 858.
4. R. Carlson, K. Elenius, B. Granström and S. Hunnicut (1985) Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages, *STL-QPSR* 1: pp. 63 - 93.
5. T. Veenhof and G. Bloothooft (1987) Statistics of Sequences of Broad Phonetic Classes in Newspaper Dutch, *PRIPU* 12.1: pp. 39 - 56.
6. D.W. Shipman and V.W. Zue (1982) Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems, *Conference Record, IEEE 82, Int. Conf. on Acoustics, Speech and Signal Processing*: pp. 546 - 549.
7. D.P. Huttenlocher (1986) A Broad Phonetic Classifier, *Proc. of the ICASSP-Tokyo*: pp. 2259 - 2262.
8. J.C.T. Ringeling (1986) Luisteren is Gokken, *Toegepaste Taalwetenschap in Artikelen* 25.2: pp. 28 - 36.
9. Z.S. Bond (1981) Listening to Elliptic Speech: Pay Attention to Stressed Vowels, *Journal of Phonetics* 9: pp. 89 - 96.
10. C. Van der Woude (1987) A Theoretical Look at Global Perception, unpublished M.A. Thesis, English Dept., Utrecht University, The Netherlands.
11. G.A. Miller and P.E. Nicely (1954) An Analysis of Perceptual Confusions Among Some English Consonants, *JASA* 27.2: pp. 338 - 352.
12. G.A. Miller and S. Isard (1963) Some Perceptual Consequences of Linguistic Rules, *Journal of Verbal Learning and Verbal Behavior*, 8: pp. 217 - 228.
13. B.A.G. Elsendoorn (1984) Heading for a Diphone Speech Synthesis System for Dutch, *IPO Annual Progress Report*, 19: pp. 32 - 35.

FROM PROMINENT SYLLABLES TO A SKELETON OF MEANING: A MODEL OF PROSODICALLY GUIDED SPEECH RECOGNITION

ROBERT BANNERT

Department of Linguistics and Phonetics, Lund University, Lund, Sweden

ABSTRACT

A model of speech recognition is sketched where the guiding role of prosody, especially the pathbreaking function of the accent syllables, is duly stressed. The relationships between the accent syllables and the root syllables of words provide the listener with a skeleton of meaning which will be completed and, if necessary, restored in further stages of the recognition processes. In a hierarchical organization of linguistic structures and processing levels, information flows between the acoustic-phonetic and the semantic level in a purposeful and optimal way interacting with phonological, morphological, syntactic, and pragmatic information.

INTRODUCTION

For quite a long time, speech perception and speech recognition has challenged the mind and skill of students of various fields of research like psychology, linguistics, phonetics, and engineering. In spite of all the enormous progress that can be witnessed, we have to state today that the problems of recognizing fluent speech, spoken by different speakers, are far from being solved as far as the fundamental principles characteristically employed by human listeners are concerned. The explanation for this state of the art has to be sought above all in our insufficient knowledge of the processes leading from the acoustic signal to the understanding of meaning conveyed by the speech signal.

In recent years, the significance of prosody in speech recognition has been recognized to an increasing degree [e.g. 1, 2]. The present paper is intended to contribute to a better understanding of the processes involved in speech recognition. Experimental data point to the important role, in relation to their acoustic and semantic features, that syllables made prominent by word accent play in the processing of the speech signal by the listener.

Every linguistic unit, like syllable, stress group, phrase, sentence, and text, has a specific structure, the knowledge of which is of central significance for speech recognition. The competence of the speaker/listener also contains, among other things, the knowledge of the phonotactic structure of syllables and words, their morphological structure (root, affixes), and their prosodic structure, e.g. the number of reductions and assimilations. The prosodic features are very often strongly interrelated with other phonological and morphological features, for instance phonotactic, morpho-phonological, and syntactic ones.

Models of speech perception have to cope with the fact that the speech signal is not always distinct and complete. Instead, most often the acoustic signal arriving at the listener's ear contains distortions of different kinds. These deviations appear as the consequences of at least three dimensions of indistinctness, namely of speech tempo (slow - fast), of articulation (distinct - lax), and of the linguistic distance between a norm or standard and the actual form (small - large) which contains regional, social, and individual features and foreign accent as well. Therefore it has to be assumed that the result of the acoustic-phonetic analysis not always amounts to a complete and unambiguous phonological form which will lead directly to the lexical element which, eventually, will be identified correctly. On the contrary, the phonological representation as the result of the working of the bottom-up processes has to be thought of as incomplete and deviant compared to the meaning intended by the speaker.

EXPERIMENTAL DATA

In a series of experiments, samples of Swedish, spoken with a strong foreign accent and deviating with respect to the distribution of word accent, were corrected temporally and tonally and thereafter presented to native Swedish listeners under various hard listening conditions. Deviant speech aggravates speech perception because the acoustic information contained in the speech signal and constituting the initial information to the processes of speech recognition may differ markedly from the normal and expected standard of pronunciation. Some interesting results emerged from manipulating certain features of the speech signal in a controlled manner by means of LPC-speech synthesis of high quality and then studying listeners' reactions to the manipulations. A detailed description of the method used and the results are given in [3, 4].

Evaluating listeners' responses to utterances manipulated in this manner, one observation is prevalent: It always seems to be the accent pattern that is picked up by the listeners. Accent pattern means the linear succession of accented and unaccented syllables in an utterance. The same accent pattern is to be found in the listeners' response, even if the accent pattern is incorrect in the stimulus, although the response differs from the intended utterance with respect to its semantic, syntactic, morphological, and phonological structure. If by way of speech synthesis the

wrongly positioned word accent is moved to the correct syllable,
listeners' responses will change in the same way. A typical
example of changes of accent pattern is illustrated in the follow-
ing (the position of word accent = primary stress is indicated
by the symbol '): The Swedish phrase "i 'samhä:llet" (in society)
showed the wrong prosodic structure "i sam'he:let" in the devi-
ant foreign accent rendering. This stimulus was heard as "i sin
'he:lhet" (on the whole) or "utan 'te:ve" (without TV) by
listeners who obviously focussed on the accent pattern or the
distribution of the word accent. However, when the tonal move-
ment, representing the most essential feature of word accent, is
moved from the wrong syllable "-'he:-" to the correct syllable
"'sam-", the pattern of listeners' responses is changed in accor-
dance with the correct accent pattern. Listeners now heard "i
'sandträ:det" (in the sand tree), "i 'samlingen" (in the collection)
or "i "handlingen" (in the action), all of them showing the ident-
ical distribution of word accent on the second syllable of the
stimulus. The number of syllables in the listeners' responses was
always identical. At the same time, it has to be noted, of course,
that the accent syllables of the listeners' responses share a cer-
tain amount of spectral features with the accent syllable of the
stimulus.

OUTLINE OF THE MODEL

Assuming certain general principles in some existing models of word and speech recognition (e.g. [5]), the most relevant features of a prosodically guided model of speech recognition will be outlined. A more detailed description is to be found in [4]. The model is summarized graphically in Fig. 1.

The hypothesized phonological structure resulting from the acoustic-phonetic analysis of the speech signal and the restructuring working of phonetic, phonological, and morphological knowledge is not totally specified. The possible phonological structure that acts as the search unit for lexical items is assumed to be an accent group, i.e. a chunk of a linguistic structure containing as its kernel the accent syllable surrounded by other unstressed syllables. No word boundaries are marked nor needed.

All information of linguistic and pragmatic kind may be used by the various stages and processes of speech recognition at all times and wherever necessary and useful. A close and optimal acting together of bottom-up and top-down information even at low levels where the first linguistic interpretation of auditive-acoustic information occurs and non-linguistic short cuts by-passing all the hierarchically structured acoustic and linguistic levels, are assumed for speech recognition.

The acoustic analysis of the speech signal is performed in two different channels, i.e. the prosodic and the spectral one (cf. [1]). Quite often the auditive-acoustic analysis cannot always result in a complete phonetic basic structure due to acoustic distortions from outside and assimilations and reductions in the signal itself.

The auditive-acoustic analysis is followed by the phonetic analysis which combines and integrates the auditive-acoustic parameters into chunks of approximately the size of a syllable and which labels it phonetically. The phonetic labelling, most often, cannot be performed in a refined way (cf. [5]). The phonetic interpretation provides the basis for the acoustic-phonetic basic information about the chunk of the speech signal to be processed.

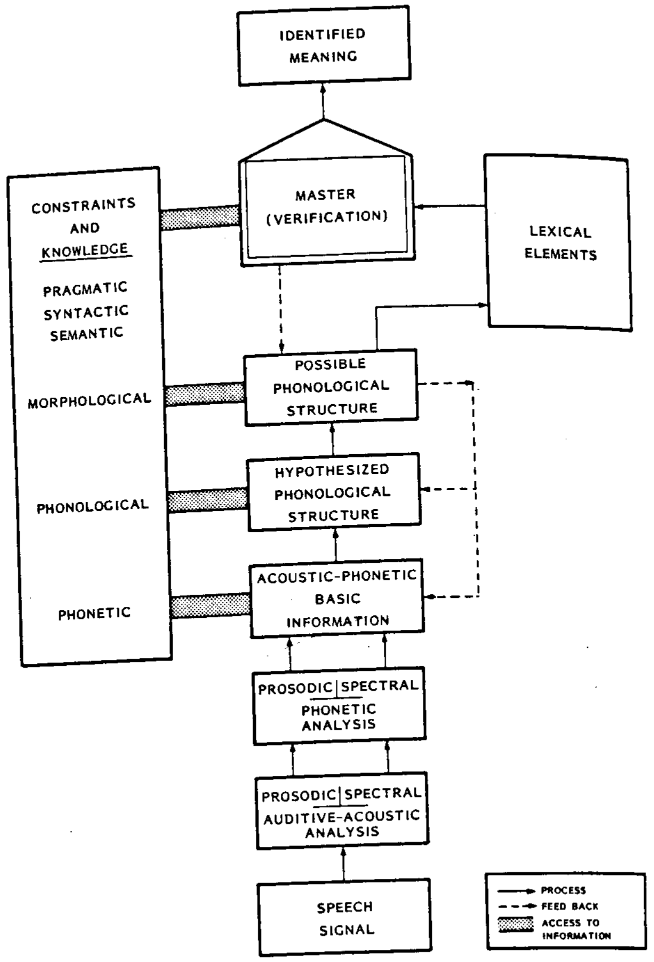


Fig. 1 A model of prosodically guided speech recognition

The acoustic-phonetic basic information is structured according to prosodic and spectral features. The prosodic features provide the position of the accented syllable or syllables in the chunk or chunks; the spectral features contain information about the spectral gestures of the segments. Taken together they provide information about the number of syllables in the chunks. There is, however, a clear difference between the two dimensions: while the accented syllable always appears correct in the basic structure, the spectral component often remains classified only in a gross manner.

This fact has certain consequences for the emergence of the hypothesized phonological basic structure on the following level: The spectral elements in the acoustic-phonetic basic information are subordinated to the prosodic structure of the accent groups. This subordination is brought about by the top-down constraints and the general knowledge of the listener which operate in generating the hypothesized phonological structure.

The hypothesized phonological structure is not generated only once and for ever but, instead, can be altered in a short period of time as a consequence of not only new acoustic-phonetic information but also of new top-down information which is flowing forth and thus becomes available all the time.

The semantic elements of the lexicon are arranged in a multidimensional fashion according to various phonological features and structural characteristics. These possible phonological structures provided by the analysis of the speech signal and the working of linguistic constraints, it must be assumed, normally do not look like orthographic words with clearly defined boundaries, which correspond exactly to a stored counterpart. They are not searched for like a numbered book in a bookshelf and found immediately by its distinctive digit. Approaching the lexical elements would rather amount to a search consisting of a large array of activities utilizing different features simultaneously. The possible phonological structure which emerged from the fragments of the acoustic-phonetic basic information contains the accented syllable as its most important search criterion. Therefore it can be assumed that the search starts out for phonological representations of lexical elements showing the identical accent pattern and some of the spectral features of the accentuated syllable. Of course, all the information concerning the surrounding syllables is used as a supporting criterion as well.

In general, it has to be assumed that speech recognition is characterized by an interplay of activities where all information available is processed simultaneously and optimally. This kind of search assumes explicitly that the boundaries in the possible phonological structure need not be defined exactly and in advance. The first aim of the search for lexical elements seems to be to find the syllables with the most distinct marking which, in turn, are identical with the basic meaning of the root or stem of a word, i.e. to find the skeleton or the corner stones of meaning.

As is generally known, languages use different principles for accent distribution in their information structure. In accent languages like, for instance Swedish, English, and German, word accent , in principle, exactly functions for signalling the word stem as the kernel of the meaning of a word. This is true both of morphologically simple and complex words. But also in languages with different principles for accent distribution, like for instance Finnish and Czech with initial accent or Polish with accent on the penultimate, the accentuated syllable represents a prominent feature of the phonological structure of lexical elements and thus a clear and distinct signal for starting the search and for the successful finding of lexical elements.

The information which is still needed at this point in order to be able to reconstruct completely the utterance containing several words will be processed and gained in the next step where verification is carried out by a component called the Master. Here, accessing the remaining information in the possible phonological structure and the top-down component, at this point especially syntax, pragmatics, and semantics, the missing parts of the phonological-syntactic structure are hypothesized and built into the total structure corresponding to (parts of) the utterance. After this verification, the process of speech recognition, hopefully, will end up with the identified meaning.

A verification component, the Master, has access to the linguistic constraints and the knowledge which, in turn, have access to the lower levels. For the Master there is also a feed-back channel to the possible phonological structure which, again in turn, feeds back to the lower levels. Thus it becomes quite clear that the top-down information is available to different and rather low levels of processing in speech recognition. It becomes also clear that, due to this fact, the speech signal need not be clear and distinct at every point in time. Of course, the more distinct the signal is, the easier and faster the lexical search can be because almost no support by the top-down component and no feeding-back is needed in this case. If the verification of some chosen lexical elements by the Master as to their linguistic and pragmatic correctness and of their semantic credability comes out negative, the feed-back channel to the possible phonological structure, the hypothesized phonological structure and, if necessary, to the acoustic-phonetic basic information will be activated. Then a change of the phonological structure already arrived at will be enforced by starting the searching process anew which, finally, will arrive at an acceptable result after having passed through a number of stages a second and maybe a third time.

In this interactive process of speech recognition, it is obvious that prosody, especially word accent, plays a direct and guiding part. Searching for lexical elements stored in the long-term memory takes place not by using words with clearly defined boundaries but rather by using prosodic features where word accent and phrase accent or focus distinctly point to the most important semantic elements of an utterance. The syllables which are prominent due to word accent represent reliable islands in the stream of sounds and there they function as the anchor or fixation points of speech recognition. Therefore it is easily understood that word boundaries are not a significant support or even a precondition for speech recognition. Phrase boundaries, however, play an important part in dividing the speech chain into appropriate processing units. It is interesting to notice in this respect that phrase boundaries are clearly marked, often by several prosodic means. In contrast, word boundaries, are not marked in any special way. Even where morphological word structure is concerned, unstressed syllables, especially at the end of a word, as markers of concord, normally contain linguistic information which can easily be derived. Therefore it is not astonishing to learn that speech recognition systems cannot find words in the signal of continuous speech if the words, even in longer texts, are not pronounced in a staccato way, i.e. surrounded by pauses. In the speech signal there are no word boundaries but acoustically more distinct and elaborated chunks of the size of a syllable, namely the prominent and accented syllables.

The model of speech perception outlined here differs from previous models in several respects. In contrast to the cohort theory, there is no activating of groups of possible word candidates all of them beginning with the same sound and the number of which will be gradually decreased as a consequence of acoustic information arriving later and of contextual constraints until, in the end, only one candidate will hold the floor. In my model, the spectral information of phonemes does not play a predominant part. Guided by the prosodic information pointing especially to the clearly marked accented syllable, one or more possible phonological structures not exactly defined by word boundaries, may start for the search of lexical elements. Very often they may even act as competitors (cf. [6]).

Rather as an amendment to the Phonetic Refinement Theory, in my model the strong part of prosody in finding the most significant and central elements of meaning is duly recognized. The process of speech recognition obeys the principle of clarity. The accent pattern, prominent in the signal and easily to be discovered and processed, forms a linguistic frame or skeleton which the spectral features are subordinated to and built into. Every part of the phonological structure which is missing or indistinct, if possible, will be restored or corrected later in the interactive processes.

Another virtue of this model lies in the fact that it is applicable to the whole range of different conditions of the speech signal in verbal communication and the bottom-up component of speech perception. The top-down component is always at work. It is obvious that a distinct and good speech signal makes speech recognition easier, faster, and accurate. If the speech signal is deviant with respect to a given norm or distorted by external sources, a larger period of time will be needed in order to identify a meaning because a larger burden is put onto all kinds of memory, information paths, and feed-back channels. An increased activation of search processes and memories explains the fatigue experienced by listeners who are exposed to speech in noisy environments or to strong foreign accent for longer stretches of time.

In conclusion, then, this model also covers speech recognition under different conditions: the optimal speech signal, spoken distinctly and free from external acoustic distortions; the speaker and listener using approximately the same standard of pronunciation; the indistinct pronunciation due to lax or fast articulation; the acoustically distorted signal; the perception of the hard of hearing and the deaf; the perception under inattentiveness and non-listening of the intended listener; the geographical, dialectal, social, and individual varieties of a language; the foreign accent.

REFERENCES

[1] Svensson S.-G. 1974. Prosody and Grammar in Speech Perception. University of Stockholm. Monograph from the Institute of Linguistics. No. 2

[2] Lea W.A., Medress M.F., and Skinner T.E. 1975. A Prosodically Guided Speech Understanding Strategy. IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. ASSP-23, 30-38

[3] Bannert R. 1984. Prosody and Intelligibility of Swedish Spoken with a Foreign Accent. In: Nordic Prosody III, C.-C. Elert, I. Johansson, and E. Strangert (eds). Acta Universitatis Umensis, Umeå Studies in the Humanities 59, 7-18

[4] Bannert R. 1986. From prominent syllables to a skeleton of meaning: A model of prosodically guided speech recognition. Lund University, Department of Linguistics and Phonetics, Working Papers 29, 1-30

[5] Pisoni D.B. 1984. Acoustic-Phonetic Representations in Word Recognition. Indiana University. Research on Speech Perception. Progress Report No. 10, 129-152

[6] Bannert R. 1980. Phonological Strategies in the Second Language Learning of Swedish Prosody. PHONOLOGICA 1980, 29-33. Innsbruck

THE ANALYSIS OF SPEECH PERCEPTION MECHANISMS ON THE MODELS OF AUDITORY SYSTEM

V.P. BONDARENKO, V.R. MOOR, A.N. CHABANETS

Research Institute of Automatics and Electromechanics
Tomsk, USSR, 634004

ABSTRACT

This article concerns the model investigations in auditory system. The model is synthesized on the basis of a little number of the raw data, with a limited system complexity, and an element reliability. Hence, the model structure concerns as a hierarchical system with a high potential reliability.

INTRODUCTION

The analysis of speech signal is associated with model investigations in auditory. The auditory system is divided into several schemical levels: mechanical conversion level of sound signal, sensory level and neuron processing level.

The auditory system complexity depends on the signal processing level. If the spinal ganglion cosists of about 30000 neurons, then the brain consists of about I0000000 neurons. The auditory system is a complex hierarchical system. It intends to the prediction systems /I/. There are two ways of the complex system simulation: analytical and synthetical.

An analytical method is based on the determinate of majority common real system parameters and their linkages. The most models of auditory system is based on the analytical approach /2,3,4,5/ The main disadvantage of this approach lies in practical impossibility to take into account the whole information about structure and behaviour of auditory system. Hence, the models obtained are private and explain only partial effects of auditory model operation. The analysis of the complex system must not be done by independent simulation of parts. Hence, the private models can not explain the basic behavioral principles of auditory system.

The synthetical approach is more preferable. It is realised by means of optimum models synthesis, which then approximate to the real system due to nearing objective functors and linkages between model -, and system parameters. The main difficulty of using of the synthetical approach lies in the term "optimum" for the auditory model system. Due to the analysis of the common processing principles of the information in auditory system /6/, we extracted as optimum criterion - minimum losed information of the input signal by its model processing. Besides, an auditory model system should hold a high potential reliability.

PROBLEM STATEMENT

The complexity of the model structure is limited upper

$$S=\sum_{i=1}^{n} N_i \leq S_0 < \infty$$

where N_i - information capacity of the i-th element;

n - generalized number elements.

The elements have threshold of sensitivity

$$\gamma^2 pt = C \leq C_0 > 0$$

where t - reaction time; γ - error of conversion; p - input signal power.

The model elements are not reliable and are not substitute after refusal. Probability of no-failure element operation is limited upper $p \leq p_0 < 1$.

The loss information of the input signal is calculated by functor

$$\delta = \int_W [f(w) - \varphi(w)]^2 \, dw$$

where $w = \Omega \times D$ - signal region; Ω - frequency range; D - dynamical range.

$$\varphi(w) = L^*[a(f)]$$

where L - signal definition operator a(f) in model; L^* - return operator.

The task is the information probability search of the model and the structural element linkage definition. The need of loss minimization defines a decision - making task.

Frequency- and dynamic ranges are divided into subranges, with their own data model element. The common decision task is in /6,7,8/, due to this decision task was synthesized the auditory model system. This model relates to the symmetrical system class m-order /I/. Hence, the system is operationable, till its every part (involving m-elements) is working.

The probability of no-failure model operation is:

$$R=\sum \prod_{i=1}^{n} p_i^{\omega_i} q_i^{1-\omega_i}$$

$$\sum_{i=1}^{n} \omega_i > m$$

where $w_i=0$ - for failured element; $w_i=I$ - for operatable element; p_i - probability of no-failure operation of i-th element; $q_i = 1 - p_i$.

In the present model the common element complexity is equal, hence: $p_i=p_j=p$

so

$$R=\sum_{i=m}^{n} C_n^i p^i q^{n-i}$$

The value estimations R are /I/, by $h \to \infty$, and $\frac{m}{n} < P_0$

$$R \begin{cases} > 1 - \exp(-kn) \\ < 1 - \exp[-kn + O(\ln n)] \end{cases}$$

where $k = \frac{m}{n} \ln \frac{m}{pn} + (1 - \frac{m}{n}) \ln \frac{1-m/n}{1-p}$

Thus, by $\frac{m}{n} < P_0$, the system reliability approximates to 1.

Modes Description

The synthetized model involves: filter system, threshold elements, spaceadding linkages, and time adding filters.

The filter system has the transfer function

$$y(j,\omega,x) = \exp[-1{,}44Q^2(ye^x-1)^2 - j5Qe^x y] \quad (I)$$

where x - space coordinate (filter number); Q - gain-bandwidth product of the filter system; $y=w/w_{01}$; w_{01} - resonance filter frequency with number x=0.

The threshold elements are equal - allocated along axis x in several rows. A threshold of element in i-th row is

$$a_i = a_1 \cdot \beta^{i-1}$$

where $B > 1$.

The element linkages are defined by functors:

$$q(x_1,t) = \int_{-\infty}^{t}\int_{0}^{x_0} h_1(t-\tau,x_1) h_2(x_1-x) L(\tau_i,x)\, dx \cdot d\tau \quad (2)$$

$$\gamma(x_1,t) = \int_{-\infty}^{t}\int_{0}^{x_0} h_3(t-\tau,x_1) h_4(x_1-x) M(\tau_i,x)\, dx \cdot d\tau \quad (3)$$

where h_1, h_3 - weight functions of the time summation; h_2, h_4 - weight functions of space summation; L,M - threshold element reactions; x_0 - upper filter number level.

The relations analyse (I7,(2),(3) has shown, that the model is not critical to value Q, since the filter time constants and weight function constants are matched. The model parameters estimation can be done by common psychoacoustical and neurodynamical data of auditory human system and by functional model analysis.

Parameter Model Improvement

There are many papers dealing with the subject parameter measurement and parameter estimate of the filter system of the auditory analyzer, but the results are contradictive /9,IO,II/.

The model described agrees well with relative levels of auditory system. It was found that the gain- band width product defines the curve of absolute sensetive level, a threshold curve type, amplitude-, and frequency modulation sensitivity, and etc. It follows, that the frequency group width (critical bandwidth) in auditory system agrees with the bandpass relative filter. Hence,

$$Q = f_{med} / f_{crit}$$

where f_{med} - medium group frequency; f_{crit} - critical frequency band for f_{med}.

Thus, gain- bandwidth product Q depends on filter resonance frequency w_{res}. Since w_{res} is a coordinate function x,

$$\omega_{res}(x) = \omega_{01} e^{-x}$$

we have, that Q is also coordinate function x.

It was found, that the gain- bandwidth means in the frequency range I-IO kc/s agrees with the artical results /IO,II/, and in range below 500 kc/s - with experimental data of Bekesy /9/.

INPUT SIGNAL CODING

Final signal description is defined with expressions (2) and (3). The functor g(x, t) defines the amplitude spectrum of the input signal, and γ (x,t) - differential from the phase spectrum.

In the bandwidth

$$\Delta\omega(x) = \frac{\omega_{res}(x)}{Q(x)}$$

g(x,t) and γ (x,t) can be represented as one count on the coordinate.

From condition (I) these counts must be taken in points of the largest value g(x, t). I.e., the restcounts it is necessary to suprress by means of the suppression function ξ (t, x)

$$\xi(t,x_1) = q(t,x) - P_{x1}(t,x) \quad ; \qquad (4)$$

$$P_{x1}(t,x) = \beta(x)\, q(x) \quad ;$$

$$\beta(x) = \begin{cases} <1 & x_1 - \Delta x \leq x \leq x_1 + \Delta x \\ >1 & x_1 - \Delta x > x > x_1 + \Delta x \end{cases} \quad ;$$

where P_x (t,x) - breakpoint in x, becoming by the excitation g (t,x) in the point x ; g(x) - model reaction coused by the sine signal.

Such a function is realized on the basis of the known lateral inhibition. A making-decision procedure on the basis (4) is: if $\xi\ (t,x) > 0$, so a signal in point x_1 exists, if $\xi\ (t,x) < 0$, so - isn't.

Values $g(x\ ,t)$ and $\gamma\ (x\ ,t)$ - are the result of the first level of the coding of the final description. It's adaptive, since $g(x\ ,t)$ and $\gamma\ (x\ ,t)$ are defined by input signal structure.

CONCLUSIONS

The model described and a number of signal processing mechanisms are very common with the data mentioned in neurodynamics and psychoacoustics dealing with human sound signal sensibility. At the same time, the model is optimal as to minimum criterion of the loss information with potential reliability, near to I.

The theoretical and the experimental mode investigations provided us to study the perception mechanisms particularly; simutaneous - and sequential mechanisms of disable, to-tone suppression, the vowel attribute determination and etc.

It was found, that the sounds formants are markedly changed during the base tone period, it allows one to obtain the information about the speech signal thin structure.

REFERENCES

I. Fleishman B.S. Theory elements of potential efficiency in complex systems. - M.: Soviet.Radio, I97I. - 223 p.

2. Labutin V.K., Moltchanov A.P. Auditory models. - M.: Energija,I973.- II0 p.

3. Flanagan J.L. Analysis, synthesis and speech perception. - M.: Svjar, I968.- 489 p.

4. Pozin N.V. Neuron structure simulation. - M.: Nauka, I970. - I56 p

5. Temnov V.L. A model of the results description of psychoacoustical experiments with stationary signals. - In the book: Speech signal analysis by human. - L.: Nauka, I97I, p. I8-25.

6. Bondarenko V.P., Razin V.M. On the modeling of human system perception. - Izvestija TPI, I976, v. 266, p. 38-40.

7. Bondarenko V.P., Razin V.M. Modeling of peripherial system perception. - Izvestija TPI, I976, v. 266, p. 4I-44.

8. Bondarenko V.P., Razin V.M. Peripherial structure optimization of perception system. - Izvestija TPI, I976, v. 266, p. 45-48.

9. Bekesy J. Experiments in hearing. - N.Y., I960.

I0. Johnstone B.M., Royle A.J.F. Basilar membrane examined with mossbauer technique. - Science, I968.

II. Boer E., Jongkees N.R. Computer simulation of cochlear filtering. Proceedings of the z-th International Congress on Acoustics, vol. 3, Budapest, I97I, p. 456-460.

Anurzej Plucinski

THE NORMALIZATION OF THE SPEECH SIGNAL SPECTRUM ENVELOPE

ANDRZEJ PLUCIŃSKI

Institute of Linguistics
Adam Mickiewicz University
61-874 Poznań, Poland

ABSTRACT

This paper proposes normalization of the speech signal envelope by means of multiplicative centralization. The method proposed is based on the assumption that speech signal analysers of the human nervous system identify such instantenous speech signal spectra which can be superimposed by means of multiplicative transformations. The method doesn't involve any initial classifications (e.g. into male/female voices, vocalic/consonantal sounds, etc.). An alternative representation of the speech signal spectrum by means of coefficients of an extension of the speech signals spectrum envelope into a power series is suggested. Such a representation give us a possibility to get rid off stationary contributions.

INTRODUCTION

An analysis of the influence of various distortions on speech intelligibility may help us to discover some mechanismus of sound perception. Simple and frequently met distortions are introduced by e.g. our electroacoustic equipment. We have no difficulty in finding out that the level of recordings being reproduced has almost no influence on the reception of the content being transmitted. Also dislocations of the spectrum in the frequency range, due to the change of the speed reproduction can reach considerable values with no effect on intelligibility. As seen from a formal point of view, these distortions consist in a multiplication of amplitudes or frequencies of the spectral components by certain constants. Apart from the above mensioned distortions we can also find linear distortions which consist in the attenuation of various spectral components, especially the extreme ones. On the basis of such observations we can infer that the received accoustic signal undergoes certain normalization in the process of perception. The normalization allows us to compensate for the interpersonal differences and for the influence of the conditions of the acoustic wave propagation.

1. A MATHEMATICAL MODEL

The distortions of the signal caused by both the change of amplification level and a non-uniform transmission of spectral components (tone quality) pertain to amplitude and can be described by means of a function dependent on frequency. Amplitudes of signal spectral components will be multiplied by values of that function. A constant component of the function will be responsible for the general amplification level. Distortions in time can be shown as multiplication of the frequency scale by the constant. Let $a(f,t)$ be a dependence which shows an envelope of the spectrum amplitude of speech signal. In accordance with the above remarks we can say that

$$a(f,t) = b(f)\varphi(f\vartheta,t), \qquad (1)$$

where b is a dependence which describes the signal transmitling characteristic, φ is an envelope of the primary signal spectrum, t denotes time, f-frequency and ϑ is a constant responsible for the displacement of the signal spectrum in the frequency range. The concept of primary signal will be clarified in the subsequent part of the paper.

As a result of such a formulation of the mathematical model we will be claiming that the distortions in question consist in multiplicative transformations of the signal in the amplitude and frequency range. We will show now that, by using multiplicative centralization, Fourier transform of the signal can be reduced to a certain standard form, free from the influence of these distortions. The centralization consists in dividing the amplitude and multiplying the frequency of spectral components by an appropriate weighted arithmetic mean.

2. NORMALIZATION BY MEANS OF MULTIPLICATIVE CENTRALIZATION

Distortions in amplitude will be eliminated according to formula (1) by dividing the envelope of the Fourier transform of the signal by the arithmetic mean

$$a^{o}(f,t) = a(f,t)/\mu(f), \qquad (1)$$

where $\mu(f)$ designates the weighted arithmetic mean.

Distortions in frequency will be eliminated if we multiply the arithmetic mean of the result of the previous operation calculated in the frequency dimension. Thus we have

$$a^{N}(f,t)=a^{o}(f\mu(t),t).$$

To justify operation (1), let us calculate, without going into details, the time mean of the envelope of the speech signal spectrum:

$$\mu(f)=(\int_{t_d}^{t_g} w(t)dt)^{-1}\int_{t_d}^{t_g} w(t)b(f)\varphi(fv,t)dt.$$

Let us substitute W for $\int_{t_d}^{t_g} w(t)dt.$

Using the properties of the integral, we can write that

$$\mu(f)=W^{-1}b(f)\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt = C_{\varphi}(f)b(f).$$

Thus, we can notice that as a result of operation (1) the multipier $b(f)$ is removed. Let us now calculate the same mean for the centralized process:

$$\mu'(f)=W^{-1}\int_{t_d}^{t_g} w(t)a(f,t)/\mu(f)dt =$$

$$= W^{-1}\int_{t_d}^{t_g} w(t)b(f)\varphi(fv,t)/$$

$$(W^{-1}b(f)\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt)dt =$$

$$= (\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt)^{-1}\int_{t_d}^{t_g} w(t)\varphi(fv,t)dt=1.$$

Hence, the next multiplicative centralizations will have no influence on the results. It follows from the above that the primary process φ is one which is invariable in relation to the multiplicative centralization of its amplitude, and that

$$C_{\varphi}(f) = 1$$

Let us calculate now the mean normalizing the position of the spectrum in the frequency dimension. Frequency is an independent variable; the only information on what frequency range the instantenous spectrum comprises is given to us by amplitudes of the components. For that reason it was decided to use them as a weight for each point in the frequency dimesnion. Such a selection of the weight function makes it possible to average only the process normalized in the amplitude dimension, because linear distortions will have a significant influence on the normalization in the frequency dimension. Thus we calculate

$$\mu(t)=(\int_{f_d}^{f_g}\varphi(f,t)df)^{-1}\int_{f_d}^{f_g}\varphi(f,t)fdf. \qquad (2)$$

If we further substitute f for s/v, we obtain

$$\mu(t)=(\int_{s_d}^{s_g}\varphi(s/v,t)/v ds)^{-1}\int_{s_d}^{s_g}\varphi(s/v,t)s/v^2 ds.$$

Using W for $\int_{s_d}^{s_g}\varphi(s/v,t)ds$, we write then that

$$\mu(t)= \frac{1}{v} W^{-1}\int_{s_d}^{s_g}\varphi(s/v,t)sds.$$

This result can be briefly written in the form

$$\mu(t) = \frac{1}{v} C_{\varphi_t}$$

If we then calculate the weighted mean (2) for the process already centralized in the frequency and amplitude domains, we will find that it will be equal to 1. It follows from it that

$$C_{\varphi_t} = 1 .$$

Thus, we can say that primary envelope φ of the speech signal spectrum is one which does not change under the influence of multiplicative centralization in the amplitude and frequency ranges.

3. SOME DETAILS OF THE NORMALIZATION PROCEDURES

I want to show now how to compute the means $\mu(f)$ and $\mu(t)$. It turns out that the computations will not be complicated when we use the extension to the power series of the spectrum envelope

$$a(f,t) = \sum_{i=0}^{I}\sum_{j=0}^{J}\alpha_{ij}f^{i}t^{j}.$$

3.1. Normalization in the amplitude range

Without affecting severely the previous considerations we can assume a week dependence of the function b on time. In consequence, a running mean $\mu(f,t_o)$ will be computed. t_o denotes the current moment.

The averaging of a signal by means of the running mean is equivalent to its filtering by means of a low-pass filter (Steiglitz 1977). In order to define the required averaging time, it suffices to determine the parameters of an equivalent filter. The parameters of an equivalent filter depend solely on the shape of the weighting curve w(t), i.e. on the so called time-window. Choosing the time-window of $\cos^2$ type we obtain the amplitude characteristic of the equivalent filter showed in fig. 1 (Pluciński 1986).

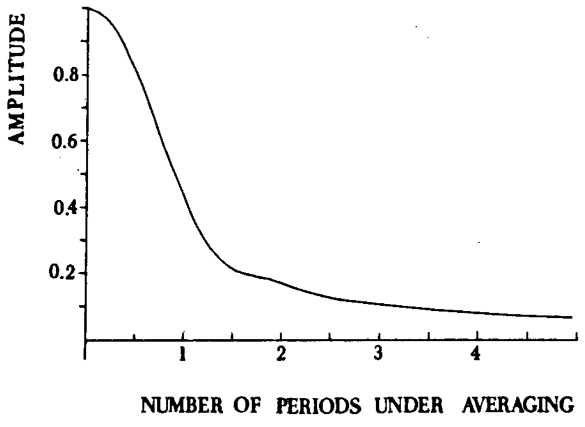


Fig. 1.

In effect the normalization, frequencies lying in the pass band of the equivalent filter will be compensated. Since the articulation time of individual speech sounds at normal speech tempo rarely goes over 100 ms, we can assume that changes in the spectrum envelope slower then 10Hz should be eliminated. Thus, as can be seen from fig. 1, the averaging time should not be longer than 200ms. The running mean for the time-window of $\cos^2$ type, i.e. if

$$w(t) = \cos^2(\frac{t-t_o}{t_o-t_d}\ \frac{\pi}{2}),$$

can be calculated according to the formula

$$\mu(f,t_o=0)= \sum_{j=0}^{J}((\frac{1}{j+1}(\frac{-1}{\pi})^{j+1}(j!\sin(\frac{\pi}{2}j) +$$

$$+\sum_{k=0}^{j}k!\binom{j}{k}(-\pi)^{j-k}\sin(\frac{\pi}{2}k)))t_d^{j}\sum_{i=0}^{I}\alpha_{ij}f^{i})$$

(c.v. Ryżyk 1964: 132, formula 2.513.4).

3.2. Normalization in the frequency range

The weighted mean over the frequency can be calculated according to the formula

$$\mu(t_j)=(\sum_{i=0}^{I}\alpha_{ij}(f_g^{i+1}-f_d^{i+1}))^{-1}.$$

$$\cdot\sum_{i=0}^{I}\frac{\alpha_{ij}}{i+2}(f_g^{i+2} - f_d^{i+2}),$$

where f_g and f_d denote borders of integration in the frequency range. This range should cover the acoustic band.

4. A PARAMETRIC REPRESENTATION OF THE SPECTRUM

In the computational technique applied here a development of the envelope of the instantaneous spectrum into a power series is used. The coefficients of this development can be used for a parametrical representation of changes of the signal spectrum in time. This has the following adwantages:
1) it allows for an uniform representation of both vowels and consonants by means of a trajectory in the space of those coefficients,
2) all information on the spectrum envelope, i.e. on the position of both maxima and minima, their amplitudes and widths is contained in this representation.

One can also expect that coefficients α_{0j} and α_{1j} of this development will not have any significant influence on automatic speech recognition. Using such a representation, we can easily remove from the spectrum a time independent component represented by coefficients α_{i0}. In the representation proposed herein we simply reject these coefficients which results in elimination of stationary noises.

We can find the parameters of the running polynomial approximation calculated on the basis of n samples taken at equal (time) intervals by the analysis of characteristics of an equivalent digital filter (Pluciński (1986)). In fig. 2 there are shown amplitude characteristics of filters equivalent to running approximation by means of the third-degree polynomial ($J=3$) on the basis of seven samples. There are shown amplitude characteristics for three types of time-windows, namely for:
1) rectangular time-window, i.e. $w(t_k)=1$,
2) $\cos^2$ time-window, i.e.

$$w(t_k) = \cos^2((k-n)\pi/(2(n-1)),$$

3) Gauss time-window, i.e.

$$w(t_k) = \exp(((k-n)^2/(n-1)^2)\ln p),$$

where $k=1,\dots,n$, $n=7$, p stands for the cut-off level of the Gauss curve, $p=0.01$.

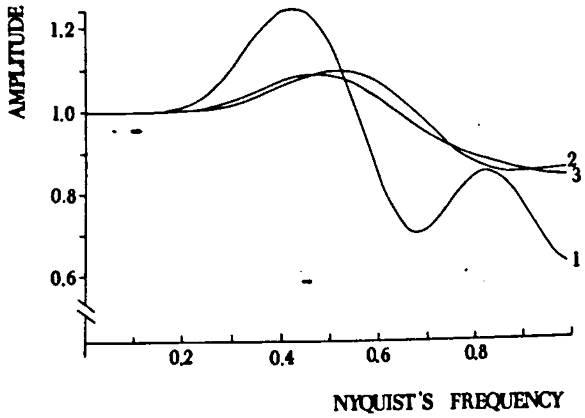


Fig. 2.

DISCUSSION

Hitherto known procedures for normalization of spectrum parameters concern formant frequencies. In order to motivate their proposals some authors refer to the anathomy of the organs of speech (Wakita (1977)), some authors to the properties of hearing (Syrdal (1986)) while others inform us only about the efficiency of a procedure of some kind (Lobanov (1971), Miller et al. (1980)). While forming normalizing rules, we aim at giving such rules which can help us identify some of numerical sets or sequences. Therefore, when analysing the rule that has been proposed by Lobanov, i.e. $F_i^N = (F_i - \bar{F}_i)/\sigma_i$, where $\bar{F}_i$ stands for a mean of i-th formant frequencies over all vowels and σ_i stands for standard deviation, we find out that this rule identify all numerical sets with elements y=ax+b, where a and b are any arbitrary constants. It may be proved by substituting F_i by aF_i+b. Analogously we can prove that Miller's formula - $F_{ij}^N =$ $= F_{ij} \prod_{i=1}^{3} F_{ij}^{-1/3}$, where i stands for the number of the vowel and j for the number of an observation - identify numerical sets with elements y=ax. The same concerns formulas proposed by (among others) Syrdal, Wakita and van Dijk. Like in the case of Miller's rule the procedure proposed in this paper identifies numerical sequences with elements y=ax. It is also a nonuniform procedure.

REFERENCES

Dijk, J.S. van. 1984. Conservation of vowel contrast in various speech conditions. Proceedings from the Institute of Phonetic Sciences of the University of Amsterdam, 8, 19-31.

Lobanov, B.M. 1971. Classification of russians vowels spoken by different speakers. J. Acoust. Soc. Am., 49, 606-608.

Miller, J.D., Engebretson, A.M., Vemula, N.R. 1980. Vowel normalization: Differences between vowels spoken by children, women and men. J. Acoust. Soc. Am., Suppl. 1, 68, S33.

Pluciński, A. 1985. The normalization of the speech signal spectrum; paper submitted to Studia Phonetica Posnaniensia 1, 57-68.

Pluciński, A. 1986. The parameters of running approximation and averaging; paper submitted to Studia Phonetica Posnaniensia, 2.

Ryżyk, I., Gradstein, I. 1964. Tablice całek, sum, szeregów i iloczynów. PWN. Warszawa.

Steiglitz, K. 1977. Wstęp do systemów dyskretnych. WNT. Warszawa.

Syrdal, A.K., Gopal, H.S. 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. J. Acoust. Soc. Am., 79, 1086-1100.

Wakita, H. 1977. Normalization of vowels by vocal tract length and its application to vowel identification. IEEE Trans. Acoust. Speech Signal Process, ASSP-25, 183-192.

PHONOLOGICAL RULE IMPLEMENTATION IN SPEECH RECOGNITION

CHARLES HOEQUIST, JR.

University of Cambridge
Linguistics Department
Cambridge CB3 9DA
United Kingdom

ABSTRACT

This paper compares the implementation of different types of phonological rules in a system providing limited dialect normalization. Dialect normalization will be sketched briefly, as a means of simplifying the speaker-normalization task.

Two phonological-rule implementations are compared: a representative of parsing by finite-state deterministic automata similar to those in Koskeniemmi [1] and of context-free phrase-structure rules like those proposed by Church [2].

INTRODUCTION

Automatic speech recognition (ASR) devices for continuous speech are forced to take account of rule-governed variation in the acoustic signal in a way that isolated-word recognizers are not. In the latter, recognition can be treated as a problem for sophisticated pattern matching. Continuous-speech recognizers, on the other hand, must attempt to cope with the inevitable acoustic variation resulting from, among other things, language-specific regularities governing the realization of segments or syllables in specific environments, that is, phonological rules.

The idea that rule-governed variation in the speech signal can best be handled by some analogue to a linguist's phonological rules is not in itself new. Several recognition devices which grew out of the ARPA project in 1971-1976 used rules to expand base dictionaries into dictionaries containing (it was hoped) all possible phonetic realizations of the dictionary's words [3]. With a realistically large rule set and large vocabulary, this sort of expansion is likely to become impractical. The implementations discussed here run in the other direction, that is, rules are applied to a labeled, segmented input string to produce candidates for matching to a fixed set of lexical entries. Our interest here is, however, not the direction the rules run, but the constraints on the power of the rule formalism and on constraining their application.

LINGUISTICS AND SPEECH RECOGNITION

Linguists' phonological descriptions of the last quarter-century have been overwhelmingly cast in the form of a single set of context-sensitive transformational rules, that is, rules which are capable of rewriting the phrase markers making up a string, in this case phonetic segments. The direction of operation should be irrelevant, so lexical items can be transformed into surface strings and vice versa.

Such a phonological grammar component has at least two major problems, concerning dialect and the formal power of the rules. It is impractical to try to design a single pandialectal set of rules for a recognition system to map inputs onto a single lexicon. Though it is imaginable that one set of rules could be written that would correctly map an input string onto the intended string of lexemes, it would in the process generate a considerable number of false mappings due to the application of rules which, by virtue of belonging to a different dialect, played no part in the production of the input. Training to an individual speaker might get around this problem (assuming for the moment that the mapping from one dialect's phonological system to another's is isomorphic), but at the cost of sacrificing generality that could be captured by the proper rules. This can be handled by the use of dialect normalization, in which an initial training phase establishes not only idiosyncratic but dialectal characteristics of the speaker, and uses these to determine what subset of rules is the most suitable. This strategy is used in the ASR section of the Cambridge Alvey project.

But even if the rules in a phonological component are marked for which dialect they apply to, the context-sensitive transformational rule formalism is itself problematic. This is because it is in some cases impossible to uniquely reconstruct the pre-transformation string, because of the ability of transformational rules to modify phrase markers. In practice, this can lead to a given phonetic surface string being traceable to several possible phonological strings, analogous to syntactic ambiguity (something similar can happen with context-free rules as well, when several structures are assigned to a given segment string; but in practice the problem is much less severe).

As in much recent work in syntax, one solution is to restrict the power of rules. As examples of context-sensitive transformations versus formally more constrained rules, two phonological parsers will be discussed here, one using exclusively context-free rules, the other allowing some context specification. The respective strengths and weaknesses of the two systems will serve to illustrate some of the requirements a phonological grammar needs to fulfill.

PHONOLOGICAL PARSERS

Context-free rules in a phrase-structure phonology

Using only context-free rules in the phonological component of a recognizer might seem hopeless at first; phonological processes are heavily context-dependent. As will be seen, the 'context-free' system discussed here does not ignore context, but encodes it in such a way that a context-sensitive formalism is claimed to be unnecessary.

The most developed example of a context-free phonological parser is to be found in Church [2]. There, a segment lattice (with no word- or syllable-boundary specifications) serves as the input to a chart parser operating with a set of context-free phrase structure rules. The parser outputs a hierarchical structure with a string of segments at its lowest level, dominated by nodes identifying syllable components, syllable boundaries, and stress-foot boundaries. This serves in its turn as input to a syllable-based lexicon.

Though the parser would be interesting to linguists simply on the grounds of the use of nonlinear phonological structure, the hierarchy of units is equally important for its restriction of the parser's power. The encoding of phonological processes, many of them contextually dependent, into context-free rules is done as a byproduct of the hierarchical structuring implicit in the chart parser. The contextual dependencies which obtain among successive segments are encoded into the higher-level units (syllables and feet), thus enabling the parser to handle many context-conditioned processes without requiring the generative capacity of a context-sensitive system. For example, the rule in English which aspirates voiceless stops in syllable-initial position would be expressed as a phrase structure rule such as

$$\text{onset} \rightarrow p^h \mid t^h \mid k^h$$

expanding some nonterminal symbol for syllable onset into aspirated voiceless stops. Thus, when parsing an input string, such segments would be labeled as syllable onsets, which in turn would allow the parser to hypothesize syllable boundaries for dictionary lookup.

This is fine, as far as it goes. However, the system (at least as Church presents it) suffers from two types of problem, one specific and system-oriented, one more general and based on the philosophy of context-free parsing.

The first difficulty can be characterized as a perfect-input requirement. This refers to the system's requirement of perfect, fine-grained phonetic labeling, and the consequences both of mislabeling, or even of labeling in too broad a fashion. Let us look at the initial-aspiration rule again. If the recognizer's front end ever fails to recognize a syllable-initial voiceless stop as aspirated, the parse fails. It is unreasonably optimistic to expect flawless performance from any front end. This difficulty weighs heavily, because a proper parse in this system depends on the accurate labeling of such phonetic detail. The problem can be evaded by making the rules more general, but this has the accompanying cost of dramatically increasing the number of valid parses formed from any given string.

Even if the initial segmenting and labeling process were to work as well as is necessary, the second problem, linked to context-free parsing itself, would remain. A parser of the sort described above decides whether a given string is allowed (that is, can be assigned a structure), given a particular rule set. Because this rule-set is composed of context-free phrase-structure rules, no rewriting of segment labels is allowed. Since segment labels cannot be altered, the phonological parser's job here is really to judge whether its input is phonotactically possible. Its alterations of input therefore only consist of the insertion of nonterminal symbols, in this case syllable and foot elements. This fails to address the existence of phonological processes involving segment insertion (e.g. epenthetic [t]) and deletion (e.g. disappearance of schwa vowels). Such rule-governed alterations in the segments are also troublesome for their ability to yield surface realizations which violate the phonotactics of English. To avoid rejection of such strings, a recognition system would have to broaden its allowable input, either by listing numerous alternative realizations of a word in the lexicon, yielding a system like that produced by the lexicon expansion rules mentioned above, or by increasing the number of rules to allow several optional rewritings of the proper non- terminal symbols. Either alternative increases the number of parses of a string and so reduces the advantage of context-free over context-sensitive rules.

Context-sensitive rules

The implementation used here as an example of context-sensitive parsing is the 'two-level' parser developed first by Koskenniemi [1]. Though his intention was to use it for morphological decomposition of an input string, it has qualities which suit it for phonological rule implementation as well.

The core of two-level parsing is its encoding of rules into nondeterministic, finite-state automata, which simulate context-sensitive rules. Since context-sensitive rules are formally more powerful than finite-state machines, it would be possible to write rules to generate strings which could not be correctly analyzed by any finite-state device. However, phonological processes do not seem to ever produce such outputs. These automata can be envisioned as moving simultaneously along two 'tapes' (hence the name 'two-level'), one representing the input and one representing a graph through the lexicon, which has a tree structure such that a node contains information about whether a lexical entry can end at that point, and if continuation is possible, what lexical characters can follow.

The rules check whether the current input character can be matched to the current lexical character. If all the rules allow the current pairing, the next character is taken from the input and checked for any allowable matches in the lexicon, and the process is repeated. This continues until the end of the input is reached (a successful parse) or no pairing of lexical and input characters is possible at some point. In contrast to Church's parser, the rules are context-sensitive, and the parse is left-to-right.

This shares with the context-free system the advantage of not needing word boundaries to be specified in advance, and so avoiding the problems caused by the reliance of a whole-word matcher on reliable word boundaries. It differs, however, in the set of strings it will pass and in the types of rules that it can express. The first difference is due to the structure of the two-level parser, and is not connected to context-sensitivity in rules. Church's parser first clusters the input segments into higher-level units, which are then used in lexical search. In the two-level system, the search is an integral part of parsing the input string. Since the rules are comparing input and lexical strings, a parse may fail simply because one of the input words is not present in the lexicon (compare this with the lexicon-independent output of the Church parser). The advantage of this is that many unusable analyses (consisting of phonotactically legitimate nonwords) are filtered out early. The second difference is that which we have been emphasizing, that rules are allowed which rewrite segment labels, thus allowing for the restoration of deletions, the undoing of neutralizations, etc.

CONCLUSION

It is the case that many phonological processes often expressed in a context-sensitive formalism can be expressed in a more restricted system. However, the inability of context-free phonological rules to alter the segments which constitute their input puts them at a severe disadvantage when faced with the output of those processes in speech production which alter segment identity (e.g. neutralization) or create surface violations of phonotactic rules through insertion or deletion of segments. Though it is possible to create a recognizer without context-sensitive rules between the input and the lexicon (recognizers can function without rules at all), the variation handled by the rules must be taken care of somewhere else. In this regard, it is significant that the context-free implementation discussed above includes a 'canonicalizer' level (Church, pp. 44-45), which removes phonetic detail and and tries to recover altered or deleted segments prior to lexical access. This latter amounts to putting in an extra set of rules to perform operations involving rewriting and inserting of phrase markers, i.e. context-sensitive transformational rules.

Nonetheless, a context-sensitive formalism cannot be said to be optimal. The very power which enables it to undo the effects of processes which defeat context-free rules also overgenerates mappings between the input and the lexicon. The obvious next step is to find ways to prune these mappings. Two methods are already in use. One is that of automatically checking the lexicon to see whether it contains the segment sequences which the rules produce, and ceasing to follow any hypothesized output which contains nonwords. The other is the reliance on an early decision as to the speaker's dialect to determine which subset of the existing rules will be applied to an input string, rather than simply trying to handle all dialects with one large rule set. Another possibility would be to apply some rules only when the system found evidence of fast-speech phenomena. Further reduction in the number of hypotheses could be achieved by the use of syntactic knowledge, to forbid sequences of lexical items that cannot be syntactically parsed.

REFERENCES

[1] Koskenniemi, Kimmo. "Two-Level Morphology", PhD dissertation, University of Helsinki, 1983. Printed in "Texas Linguistic Forum", vol. 22, June 1983, 1-164.

[2] Church, Kenneth W. "Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints", MIT dissertation, distributed by IULC, June 1983.

[3] Shoup, June E. Phonological Aspects of Speech Recognition. in: W. Lea (ed.), "Trends in Speech Recognition", New Jersey: Prentice-Hall, Inc., 1980, 125-138.

ACKNOWLEDGMENTS

This paper is based on work carried out as part of the Linguistics Department's component (SERC grant GR/D/42405) of Alvey research project MM1/069 on Automatic Speech Recognition, which involves Cambridge University, the MRC Applied Psychology Unit, and Standard Telecommunication Laboratories Ltd.

ADAPTATION TO REGIONAL ACCENTS IN AUTOMATIC SPEECH RECOGNITION

WILLIAM BARRY

University of Cambridge
Linguistics Department
Cambridge CB3 9DA
United Kingdom

ABSTRACT

A method of speaker-adaptive speech recognition is presented in which systemic differences are exploited to identify the speaker's gross regional accent: A small number of "calibration" sentences are spoken by the prospective user. Intra-sentence comparisons are made of selected vowels differing between dialects in their systemic value, and the speaker is scored on strength of adherence to one of four gross regional accents. The regional accent decision and the numerical data derived from the analysis of the calibration sentences are used to modify values in the vowel reference tables.

REGIONAL ACCENT DIFFERENCES

Speaker-independent automatic speech recognition requires a solution to the problem of regional accent differences. The accent has first to be identified, and then the reference values used in the recognition process have to be adapted towards the particular accent.

Differences between accents exist at various levels of description. Firstly, there may be differences in the phoneme inventory. For example, many speakers of Northern British English do not distinguish the vowels in "look" and "luck", or "put" and "putt"; many Scottish speakers have the same quality vowel in "good" and "food".Secondly, even in those parts of the vowel system that have equivalent phonemic oppositions, the lexical distribution of phonemes may differ. This may be due to different historical development in a large number of words such as /æ/ in "path", "grass", etc. in American and Northern British English while Southern British English has /a:/. Alternatively, there may be isolated incidences, such as "tomato", which has /ei/ in American and /a:/ in British English. Thirdly, regional accents differ in the phonetic quality of functionally equivalent phonemes. For example, although Southern and Northern British can both be said to have a distinctive contrast between the vowels in "cat" and "cart", that distinction is not carried to the same extent by the same phonetic properties. The qualitative difference between /æ/ and /a:/ in some areas of Northern England is very small, the distinction relying almost totally on the length difference; in Southern British the qualitative difference is very noticeable.

SYNTAGMATIC COMPARISON

These differences can be exploited for recognition purposes by comparing the acoustic characteristics of selected vowels within a known text. Two known words may contain different quality vowels in one dialect and the same quality vowel in another. Whether the reason is a difference in inventory, lexical distribution, or just a difference in the phonetic relationship of functionally equivalent phonemes, analysis will provide evidence for or against a particular regional accent. This principle of text-internal or 'syntagmatic' comparison has an obvious advantage over comparison with any external template values. The relational values are obtained from the individual's own realisational framework, avoiding the problem of having to normalise for non-dialectal inter-speaker differences.

DELIMITATION OF REGIONAL ACCENT

Although regional accent variation is strictly speaking non-discrete, both in geographical terms moving from one area to another, and in sociological terms within a given area, some people are categorisable according to their geographical background. Four gross accent areas were selected for differentiation: Southern Standard British (SSB), Northern British (NB), Scottish (Scot), and General American (USA). The differences within these regions may well be regarded by some (particularly those who live in them) as being at least as great as the differences between them. They do, however, constitute regional accents which are readily recognised in everyday speech communication by linguistically naive persons, and must therefore be considered to have some identity. General American, in particular, is not a natural regional accent associated with one geographical area. It is a standardised accent, roughly equivalent in the United States to SSB in Britain.

PHONOLOGICAL DIFFERENCES

These gross accent areas offer phonological and phonetic differences in their vowel systems which can be exploited for identification purposes. On the basis of evidence collated in Wells [3], the phonemic differences, tabulated in Table 1, are theoretically sufficient to distinguish between them. Words (in capitals) are used to represent the phonemic oppositions because differences in the lexical distribution of phonemes and variation in phonetic quality make the choice of one symbol rather than another confusing. The word labels used are those employed by Wells [3, p.127ff.].

Table 1. Primary vowel comparisons for dialect separation.
+ : phonemic opposition
– : no opposition exists

	SSB	NB	Scot	USA
TRAP-BATH	+	–	–	–
FOOT-STRUT	+	–	+	+
FOOT-GOOSE	+	+	–	+

These three oppositions differentiate the British accents more clearly than the American accent. Three further vowel comparisons provide additional dialectal definition for American English. They also provide additional characterisation of Scottish. These may be considered "secondary" comparisons (Table 2).

Table 2. Secondary vowel comparisons for dialect separation

	SSB	NB	Scot	USA
LOT-CLOTH	–	–	–	(+)
LOT-THOUGHT	+	+	–	(+)
LOT-PALM	(+)	+	+	–

The bracketed indication of an opposition for USA in the LOT-CLOTH and LOT-THOUGHT oppositions are a necessary acknowledgement of differences within North America. Although a distinctive contrast is claimed for both of them in General American, there are many speakers who make no contrast. The bracketed opposition for SSB LOT-PALM is an indication that the vowel quality distinction is unreliable; the opposition relies more strongly on the length difference of the two vowels.

Another type of difference provides useful additional sub-grouping, namely the incidence of the long monophthongs /a:, ɔ:/. In so-called 'rhotic' dialects, that is in our Scottish and USA speakers, they do not occur in words spelled with an <r> following the vowel. In addition, of course, these dialects do not have /ɜ:/, which occurs in words such as "bird", "hurt" or "heard", nor the centering diphthongs and triphthongs (as in "here, hair, tour, hire, hour", etc.).

CALIBRATION SENTENCES

The accent classifier operates on sentences containing the word classes given in Tables 1 and 2. Practical usefulness to a speaker-independent recognition system requires that the sentences satisfy two conflicting criteria: they have to be as short as possible yet provide all the vowel comparisons, in stressed position, necessary for differentiating the target accents, if possible more than once for greater reliability. Ideally, to provide a representative picture of a speaker's vowel space, they should also contain at least one token of the vowels not required for the comparisons.

The following four sentences satisfy all these requirements:

1. After tea father fed the cat.
2. Father hid that awful cart at the top of the park.
3. Father cooked two of the puddings in butter.
4. Father bought a lot of cloth.

In sentence 1 we have a difference in distribution. Although both SSB and USA have an /æ/-/a:/ distinction, /æ/ occurs in many words in USA which have /a:/ in SSB. Thus, when comparing "after", "father", and "cat", an American speaker will have a the same vowel quality in "after" and "cat" and a different quality in "father"; the SSB speaker will have the same quality in "after" and "father" and a different quality in "cat".

In sentence 2 the difference between rhotic /ar/ in "cart" and "park" and the non-rhotic /a:/ in "father" will signal a Scottish and an American accent. SSB and NB have a non-rhotic /a:/ in all three words. In addition, less difference between "awful", and "top" than between "father" and "top" would be evidence for a Scottish or a Northern British speaker.

Sentence 3 provides an example of complete neutralisation. Northern British, in contrast to SSB has no distinction between the vowels in "cook" or "pudding" and "butter". A Scottish speaker, on the other hand, will have the same quality vowel in "cook", "pudding" and "two", a strongly fronted, close, rounded vowel.

In sentence 4, minimal differences between the vowels in "bought", "lot", and "cloth" would signal Scottish; similarity between "bought" and "cloth", with both words differing from "lot" would indicate USA; the same quality in "lot" and "cloth", and a large difference between

each of these words and "bought" would be evidence for SSB.

In addition, the sentences also contain stressed words with the vowels /i:/, /ɪ/, and /e/, completing the inventory of stressed pure vowels, except for /ɜ:/. This is important if the accent identifier is to be used for anything more than pure diagnosis.

As comparison conditions are fulfilled, points are allocated for or against particular dialect categories. Positive and negative scoring aids differentiation. In some cases, fulfillment of a condition is evidence for one regional accent but strong evidence against another. For example, in sentence 1, a large difference in quality between "after" and "father" coupled with similarity between "after" and "cat" is strong evidence for an American accent and against Southern British. In other cases, an accent category is indifferent to nonfulfillment, and no negative points are allocated. For example, in sentence 3, Northern British will score positively, and Southern British negatively if "cook", "pudding", and "butter" have similar vowel qualities, but the USA score will be unaffected, due to a tendency for many American speakers to centralise both /ʊ/ and /ʌ/. Classification of a given speaker is based on the maximum accumulated score gained by any regional accent.

An obvious weakness in the practical application of the accent identifier is its present use of rigid and relatively gross criteria. Speakers with strongly modified accents can still be detected as non-standard by the human listener by means of other, perhaps finer regional features, which the accent identifier ignores. For example there are consonant and prosodic features which differ widely from one accent to another which have not yet been incorporated. At the moment, the only step towards differentiation of the degree of adherence to a regional accent is obtained from the continuous record of mark allocation, which can track the features which may deviate from the overall regional accent decision.

ANALYSIS PROCEDURE

Analysis is carried out in two steps. The first is a dynamic programming procedure to locate the vowels in the input sentences to be analysed. The second step is the comparison procedure itself.

The dynamic alignment uses a symmetric DP matching algorithm [2] operating after endpoint location on a combined measure of average amplitude per 20ms frame (normalised to compensate for differences in recording level) and zero-crossing count. After alignment, the analysis frames of the input sentence correspond to the frames in the reference sentence containing predefined comparison points; the comparison points are located manually with a speech-signal editor approximately one third through the selected vowels and the values stored.

The comparison procedure is an LPC-based, three-formant Euclidean distance calculated on an auditory (equivalent rectangular bandwidth) scale. The use of auditory scaling has the advantage of reducing the effect of F3 variation while giving very low F3 in rhotic vowels sufficient weight to influence the difference value. The formants are obtained by second derivative peak-picking, and cleaned by applying combinatorial constraints derived from phonetic theory. The constraints can be made extremely powerful by the fact that the vowels are known. It is also possible to inhibit individual vowel comparisons if a plausible formant structure is not found, thus avoiding totally spurious accent judgements.

In general, the formant analysis has proved very reliable, only falling down when the endpoint location of the input sentence, prior to the dynamic alignment procedure, fails due to extraneous noise. The use of a combined zero-crossing + amplitude measure for endpoint location provides considerable resistance to non-periodic disturbance.

ADAPTATION TO ACCENT

Accent identification itself is only the first step towards better recognition of non-standard speech. Adaptation is the necessary second stage. Part of this is possible on the basis of independent regional speech data, part depends on data on individual speakers gained from the calibration sentences during the identification process.

Independent vowel formant data for the regional accents have been collected from /hVd/ syllables. These provide regional group average values against which individual regional speakers' vowels can be matched. However, direct formant-to-formant matching assumes that, apart from vocal tract length differences, speakers differ only in regional accent. However, there can be other differences in long-term articulatory patterning [1] within accent groups resulting in differences in the exploitation of F1 and F2 space. In impressionistic terms this can be seen as the tendency of some speakers to speak without much jaw movement, or without much forward-and-back tongue movement. Adaptation to these differences is also possible on the basis of formant data gathered during the accent identification process.

Firstly, a group average F1/F2 "centroid" value is calculated from the average vowel values, each vowel in the regional system being related to the centroid by an F1 and an F2 factor. In addition, the maximum and minimum F1 and F2 values give the group F1 and F2 "dispersion" values. Individual "centroid" and "dispersion" values are calculated from the calibration-sentence data. Adapted vowel target values are calculated by applying the regional group vowel factors to the individual centroid values, using the F1 and F2 dispersion factors (= individual dispersion / group dispersion) to stretch or squeeze the vowel space in the F1 or F2 dimension.

SUMMARY AND DISCUSSION

The accent classifier is conceived as a first stage of a complex front-end component in a speaker-independent speech recogniser. The correct classification of a speaker's accent is essential information which will be passed up the system, enabling, for example, the subsequent front-end sub-components to adapt to the speaker. It may also be needed to trigger a particular subsection of phonological rules, and to direct accent-dependent lexical access. However, more than just the accent decision can be exploited in the speaker adaptation process, which can be envisaged basically as a 'mapping' of the acoustic space in which the particular speaker produces his vowels. Analysis data from the calibration sentences provides an economical basis for this mapping procedure.

Problems not addressed by the approach described here are, male/female speaker normalisation, and modification for degree of regional adherence. Progress in the latter depends to a large extent on long-term data obtained from the accent classifier revealing which oppositions most frequently differentiate the speakers. As data accumulates, statistical evaluation will determine the relative frequency of occurrence of particular regional features. The hierarchy thus obtained can be used to specify degrees of regional accent and associate them with particular vowel features.

REFERENCES

[1] Nolan, F. J., 1983 The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

[2] Sakoe, H. and Chiba, S., 1973 Comparative study of DP pattern matching techniques. Speech Research Group, Acoust. Soc. Japan Report S73-22.

[3] Wells, J.C., 1982 Accents of English. Cambridge: Cambridge University Press.

ACKNOWLEDGEMENTS

This paper is based on work carried out as part of Alvey research project MMI/069 on Automatic Speech Recognition, funded via SERC grant GR/D/42405 under the collaboration of STL, Cambridge University, and the MRC Applied Psychology Unit. I am grateful to colleagues on the project and in the Linguistic Department for discussion and helpful criticism.

USE OF THE ERB SCALE IN PERIPHERAL AUDITORY PROCESSING FOR VOWEL IDENTIFICATION

D. H. DETERDING*

Department of Linguistics
Sidgwick Avenue
Cambridge CB3 9DA
England

ABSTRACT

Some previous systems for using knowledge of peripheral auditory processing in speech recognition have used the Bark scale. Here, the use of the ERB scale is compared with the Bark scale.

Vowel spectra are transformed in the manner suggested by Bladon and Lindblom. The resulting vowel representations using the two different scales are then compared for a whole-spectrum approach to speaker-independent vowel recognition.

The success rate for correct identification is quite high with either scale; but it is unlikely that the remaining errors could be overcome using this kind of whole-spectrum approach.

INTRODUCTION

In recent years, many researchers have investigated the use of models of the peripheral auditory system as the first stage in automatic speech recognition systems. It is argued that, if the speech can be transformed in a manner similar to the processing of the ear, the task of recognition will be made easier.

If such a transformation is to be used, it is important that it be as accurate as possible. In their suggested auditory transform, Bladon and Lindblom [1] use a Bark scale. Moore and Glasberg [2] suggest that their ERB scale (standing for Equivalent Rectangular Bandwidth) is more accurate. In this paper, a comparison is made of the effectiveness of using these two scales in producing auditorily-transformed spectra for speaker-independent vowel recognition.

BARK SCALE vs ERB SCALE

Plots of the two scales against a log Hertz scale are shown in Figures 1 and 2.

The principal differences between the two scales are: the width of the critical band estimated by Moore and Glasberg is smaller, so there are more ERBs below 5000

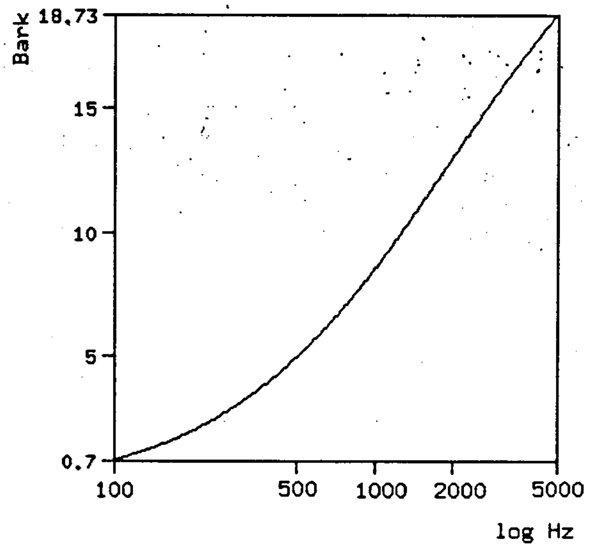


Figure 1. Plot of Bark scale against log Hz scale.

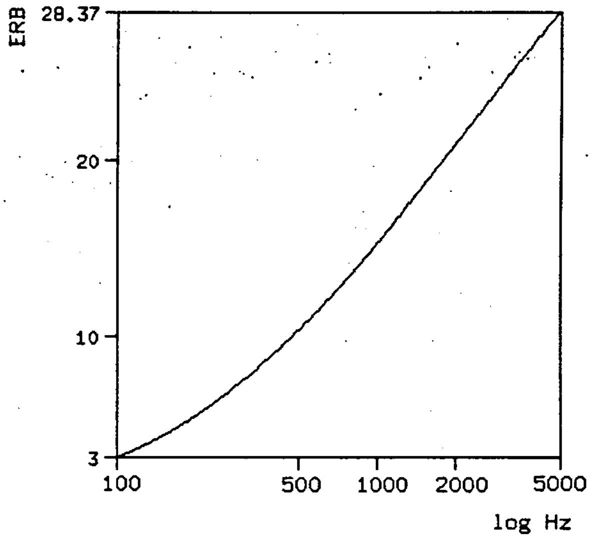


Figure 2. Plot of ERB scale against log Hz scale.

Hz than there are Bark; and the ERB scale deviates less from a logarithmic scale below 50C Hz.

One consequence of these differences is that, when vowel spectra are transformed to simulate aspects of peripheral auditory processing, the lower harmonics tend to be resolved on an ERB scale, whereas they are smoothed out on the Bark scale.

AUDITORY TRANSFORMS

In the experiment reported here, frames of 25.6 msec of speech were extracted from vowels uttered by a number of speakers. FFTs of these frames of speech then underwent transformations derived from models of the peripheral auditory system, and the final representations were used for attempts at automatic vowel recognition.

For the Bark scale representations, the various stages of the Bladon and Lindblom transform were performed according to the formulae in [1].

To derive comparable representations for the vowels on the ERB scale, the formula for calculating excitation patterns from Moore and Glasberg [2] was used in place of the convolution of the masking filter in the Bladon and Lindblom model; but Moore and Glasberg provide no formulae for db-to-phons or phons-to-sones conversions, so these were taken directly from the Bladon and Lindblom model.

Examples of the various stages of the two transforms on the FFT spectrum of a frame of speech are shown in Figures 3 and 4.

It can be seen that the final ERB scale representation is less smooth than the final Bark scale representation. This is a consequence of the narrower masking filter suggested by Moore and Glasberg. It is possible that any harmonic ripple that has not been smoothed out could interfere with vowel identification; so a wider masking filter was also tried with the ERB scale. However, the success rate for vowel recognition using this wider filter was worse, so the results presented here for the ERB scale are for the narrower filter.

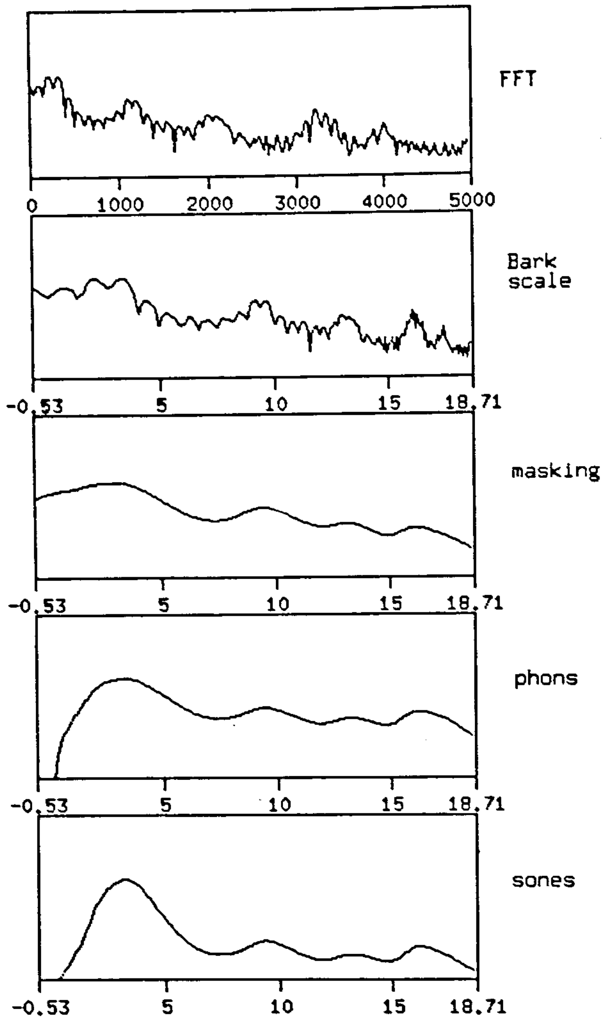


Figure 3. The effect of the various stages of the Bark transform for one token of "who'd".

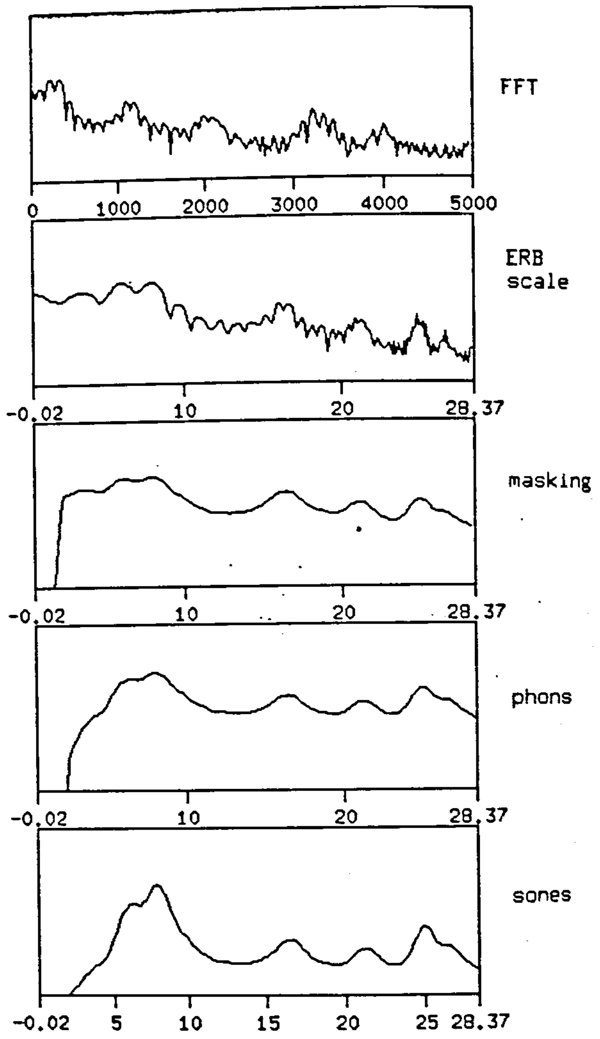


Figure 4. The effect of the various stages of the ERB transform for one token of "who'd".

NORMALIZING AUDITORY REPRESENTATIONS

Blomberg et al [3] find that, for vowel identification, the various stages in their auditory transform are actually destructive except for the last (DOMIN) stage; but they investigate recognition for each speaker independently, without attempting any kind of cross-speaker normalization. It is possible that an auditory representation only becomes important when speaker-independent recognition is attempted.

In the experiment reported here, identification of the vowels of each of thirteen speakers was based on templates derived from the vowels of the other speakers, so some kind of normalization was needed.

If speaker normalization can be achieved by a simple shift along an auditory scale to account for different vocal tract lengths [4], the shift required for adapting to one speaker from a set of templates should be appropriate for all the vowels of that speaker. Derivation of an appropriate shift can therefore be done on the basis of a single calibration vowel: the shift that allows the two representations of the calibration vowel to become most similar can be used for normalizing all the other vowels. This is comparable to the normalization scheme proposed by Nearey [5], though it uses an auditory scale instead of the logarithmic scale that he suggests.

Various vowels were tried as the calibration vowel for normalization, and the vowel from "hard" was found to provide the highest success rate. For the results presented, the calibration vowel was always "hard".

VOWEL RECOGNITION EXPERIMENT

Eight male and five female speakers, all using a Standard Southern British accent, each produced the words "heed", "hid", "head", "had", "hard", "hud", "hod", "hoard", "hood", "who'd", and "heard" in isolation. The frame of speech for use in the recognition was extracted from about one third of the way along each vowel. The location of this frame was determined manually, by examining the speech with a speech editor.

For identification of the vowels of each speaker, templates were derived by averaging the vowel representations of all the other speakers. For each vowel, identification was done by finding the template with a representation (after displacement by the normalizing shift) most similar to that of the vowel. The similarity of two vowel representations was determined by the Euclidean space between them.

RESULTS

The percentage of correct vowel identifications under various conditions is shown in Table 1. It is hard to draw clear conclusions about the superiority of either auditory scale from these results.

The success rate for vowel recognition after each of the various stages of the transforms is shown in Table 2. These figures suggest that each of the stages improves the recognition success rate, with the possible exception of the last stage. These findings differ from those of Blomberg et al [3].

The results in Table 1 show that the recognition performance for the female

	BARK	ERB
Normalized		
Male Only	89	92
Female Only	76	78
All	86	86
Un-normalized		
Male Only	90	94
Female Only	74	78
All	84	83

Table 1. Percentage of correct identifications under various conditions: in the "normalized" conditions, a normalizing shift was derived as described; in the "un-normalized" condition, no normalizing shift was used; in the "male" condition, the vowels of the male speakers were recognized using templates derived from the vowels of only the other male speakers; similarly for the "female" condition; in the "all" condition, the vowels of each of the speakers were used for identification of all the other speakers.

	BARK	ERB
FFT	64	64
auditory scale	74	73
masking	81	86
phons	83	87
sones	86	86

Table 2. Percentage of correct vowel identifications using the outputs of each of the stages of the transforms.

vowels was considerably worse than for the male vowels. Examination of the pattern of misidentifications showed that on both scales many of the vowels of one female speaker had been incorrectly identified. The possibility that the normalizing shift for this speaker was not optimal was then investigated.

All possible normalizing shifts, from minus 40 to plus 40 points, were tried. (One point represents 1/256 of the total spectrum, ie 0.075 Bark or 0.11 ERBs.) No shift allowed more than six (out of eleven) correct identifications on the Bark scale or seven on the ERB scale.

Even if, for this speaker, the templates were derived from only the other female speakers, the success rate was not perfect: no normalizing shift allowed more than eight correct identifications on either scale.

It seems that no simple normalizing shift will allow all the vowels of this speaker to be identified correctly.

It might be argued that the perception of some vowel distinctions lies mostly in the duration of the vowel, so, for example, for many speakers of Standard Southern British one cannot expect /ɑ:/ and /ɒ/ to be differentiated on the basis of a single extracted frame. But, with the best shift for this speaker using the female only templates, the remaining errors on both scales included:

/ae/	identified as	/3:/
/u:/		/I/

These errors could not be resolved by considering the duration of the vowel.

DISCUSSION

Many of the vowel representations looked like that in Figure 5, with much less distinct peaks than those of Figures 3 and 4.

Given the amorphous shape of the vowel in Figure 5, the high success rate of the recognition was surprising. If a single normalizing shift is used with a whole spectral matching, it is doubtful if the success rate could be improved much beyond its present level.

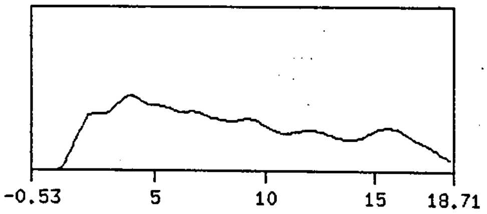


Figure 5. A Bark scale vowel representation of one token of "had".

Psychophonetic experiments [6] indicate that whole-spectrum-based vowel recognition is not likely to succeed because it retains spectral information that is relevant to the speaker's voice quality but not to the phonetic identity of the vowel. Spectral tilt, formant bandwidth, and even substantial changes in relative formant amplitude have little effect on phonetic vowel identity, but they have drastic effects on whole-spectrum matching scores. In obtaining better phoneme recognition scores than achieved here, Suomi [7] attempts to factor out the effects of spectral tilt from his whole-spectrum representations.

It is clear that some attempt must be made to find important features, principally the location of the formant peaks, and to use these for vowel recognition.

ACKNOWLEDGEMENTS

This paper is based on work carried out as part of the Linguistics Department's component (SERC grant GR/D/42405) of Alvey research project MM1/069 on Automatic Speech Recognition, which involves Cambridge University, the MRC Applied Psychology Unit, and Standard Telecommunication Laboratories Ltd.

REFERENCES

[1] R.A.W. Bladon and B. Lindblom, "Modelling the judgment of vowel quality differences", JASA 69 (5) pp. 1414-1422, 1981

[2] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", JASA 74 (3) pp. 750-753, 1983

[3] M. Blomberg, R. Carlson, K. Elenius, and B. Granström, "Auditory models as front ends in speech-recognition systems", in "Invariance and Variability in Speech Processes", (ed. J.S.Perkell and D.H.Klatt) Lawrence Erlbaum Assoc. pp. 108-114, 1986

[4] R.A.W. Bladon, C.J. Henton and J.B. Pickering "Outline of an auditory theory of speaker normalization", Proc. of 10th Int. Conf. on Phon. Sciences, Utrecht, pp. 313-317, 1983

[5] T.M. Nearey, "Phonetic Feature Systems for Vowels", Indiana University Linguistics Club, 1978

[6] R. Carlson and B. Granström, "Model predictions of vowel dissimilarity", STL-QPSR 3-4/1979 pp. 84-104, 1979

[7] K. Suomi, "Whole spectrum vowel normalization", Speech Communication 3, pp. 199-209, 1984

РАСПОЗНАВАНИЕ РЕЧИ ПРОИЗВОЛЬНОГО ДИКТОРА ПО КЛАСТЕРНЫМ ЭТАЛОНАМ

ВЛАДИМИР МАЗУР

Кафедра биофизики и математических методов в биологии
Львовский университет, Украина, СССР, 290005

РЕФЕРАТ

В работе предложен способ распознавания речи произвольного диктора по кластерным эталонам, являющих собой однй реализацию каждого слова словаря в произнесении диктора-центра кластера. Описан процесс создания кластеров, исследованы его характеристики.при изменении условий экспериментов. Определены оптимальные параметры для создания кластеров. Предложена классификация дикторов по их пригодности для работы с неадаптивной СРР. Получены результаты распознавания речи произвольных пользователей по кластерным эталонам.

ВВЕДЕНИЕ

Классификация различных подходов к построению неадаптивных систем распознавания речи приведена в работе /1/. Предлагаемый нами алгоритм распознавания речи произвольного пользователя является разновидностью подхода, использующего статистическое обучение, с введением более быстрого, экономичного и эффективного способа дикторской адаптации и представления эталонов. Принятый подход позволяет исключить необходимость большого набора эталонов при обучении за счет применения кластерных эталонов, являющих собой одну реализацию каждого слова словаря в произнесении диктора - центра кластера /2/. Задача создания кластеров, требующая статистический материал и основанное на нем обучение, решается на этапе исследования дикторских голосов. Созданные кластеры постоянны и не зависят от используемого словаря. Изменение словаря влечет за собой только запись эталонов для дикторов-центров кластеров. Голос произвольного диктора, желающего работать с системой, предварительно классифицируется по "парольной" фразе и система "настраивается" на эталоны наиболее близкого по речевым параметрам кластера, по которым происходит распознавание, либо система выдает отказ, что означает, что данный диктор может работать только с адаптивной СРР.

АЛГОРИТМ КЛАСТЕРИЗАЦИИ И РАСПОЗНАВАНИЯ

Для создания кластеров дикторских голосов был записан банк образцов речи различных дикторов. В экспериментах приняло участие 50 дикторов, из них 30 мужчин и 20 женщин. Каждый из дикторов произнес по 10 слов (цифры от 0 до 9), признако-временное описание которых было записано в банк образцов речи. Почти все дикторы, принявшие участие в эксперименте, впервые работали с речевым вводом, иначе говоря, были "несотрудничающими" дикторами. С целью устранения явно видимых дефектных эталонов, было предусмотрено 2 варианта коррекции. Первую коррекцию можно было осуществить во время создания банка эталонов путем повтора плохого эталона. Вторую - в режиме коррекции, записывая новый эталон вместо дефектного. Контроль за качеством записанных эталонов можно было осуществлять в процессе создания банка посредством анализа выводимых на дисплей параметров либо анализируя распечатки параметров созданного банка образцов речи.

Алгоритм кластеризации заключался в создании такого каждого последующего кластера, который был бы максимально отличным от всех уже имеющихся. В каждый из них включались все те дикторы, различие которых по измеряемым параметрам речи находилось в пределах ограниченной области, определяемой радиусом кластера R. Более подробно алгоритм описан в работах /3, 4/.

Выбор рабочего кластера для дикторов, принимавших участие в записи банка эталонов, осуществлялся по распечатке качественного состава кластеров. Если один и тот же диктор входит сразу в более чем два кластера, для него целесообразно выбрать тот кластер, где его порядковый номер после кластеризации (считая от центра кластера) наименьший. Если с системой хочет работать диктор, не принимавший участия в создании банка эталонов, для него нужно произвести экспресс-кластеризацию голоса и выбрать соответствующий его голосу кластер.

Распознавание осуществляется с использованием алгоритма динамического программирования, описанного в работе /5/ с применением метрики Чебышева.

ИССЛЕДОВАНИЕ И ОПТИМИЗАЦИЯ КЛАСТЕРОВ

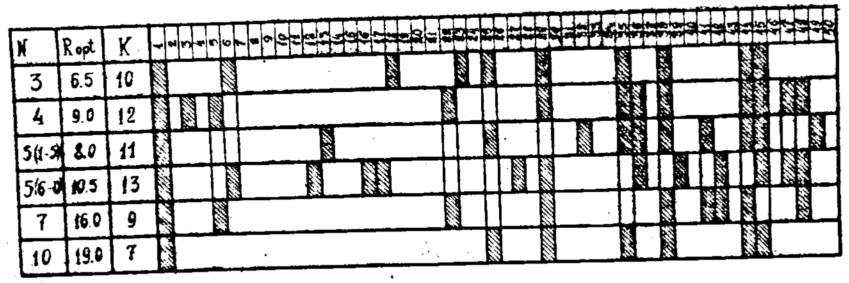
Для исследования качественного и количественного состава кластеров, их взаимосвязей и преобразований в зависимости от параметров алгоритма было проведено ряд экспериментов. Как уже отмечалось выше, голос произвольного диктора классифицируется по произнесенной им парольной фразе, являющейся как-бы его "визитной карточкой". Нас интересовало, как влияет длительность парольной фразы и ее акустико-фонетический состав на кластеризацию голоса, а в конечном счете на надежность распознавания речи по кластерным эталонам. Кроме этого хотелось определить оптимальную длительность парольной фразы. Для этого были использованы 6 различных по длительности и составу парольные фразы. Они были составлены из цифр от 0 до 9 и состояли от 3-х до 10-и слов. Состав отдельных парольных фраз следующий: парольная фраза из 3-х слов - 4,5,6; 4-х слов - 4,5,6,7; 5-и слов (1-й вариант) - 1,2,3,4,5; 5-и слов (2-й вариант) - 6,7,8,9,0; 7-и слов - 4,5,6,7,8,9,0; 10-и слов - 0...9. Средняя длительность различных парольных фраз находится в пределах от 1.5 до 4.1 сек, а ее распределение по отдельным парольным фразам видим на рис. 4.

Для каждой парольной фразы исследовался допустимый диапазон радиусов кластеров (R), изменяющийся, в зависимости от используемой парольной фразы, от 5.5 до 26 условных единиц. Диапазон изменения R исследовался начиная от появления первого кластера с количественным составом более двух дикторов (D > 2) и кончая величиной R максимально возможной при работе алгоритма.

Результаты кластеризации голоса 50-и дикторов при различных условиях экспериментов приведены в табл. 1. Для каждой парольной фразы, состоящей из N слов при оптимальном радиусе R было получено K кластеров дикторских голосов. Полученные кластеры обозначены штриховкой под номером, соответствующим диктору-центру кластера - Cp. По результатам, приведенным в таблице, можно выделить 3 группы дикторов - центров кластеров:

- "устойчивые" дикторы-центры кластеров (L = 5,6);
- "среднеустойчивые" дикторы - центры кластеров (L = 3,4);
- "неустойчивые" и "одиночные" дикторы-центры кластеров (L = 1,2),

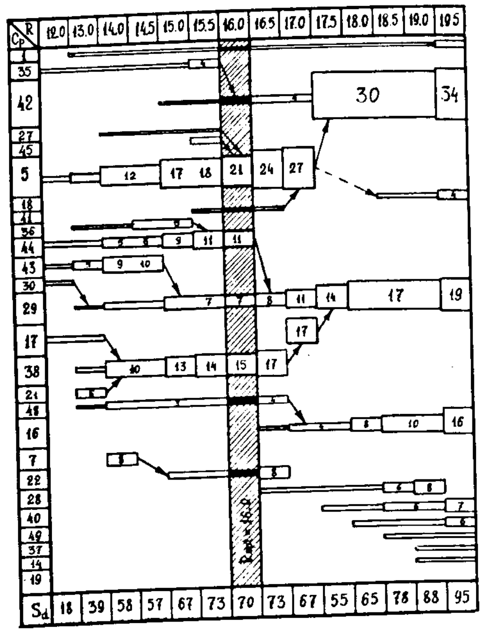
Табл. 1.



N	R_{opt}	K
3	6.5	10
4	9.0	12
5(1-5)	8.0	11
5(6-0)	10.5	13
7	16.0	9
10	19.0	7

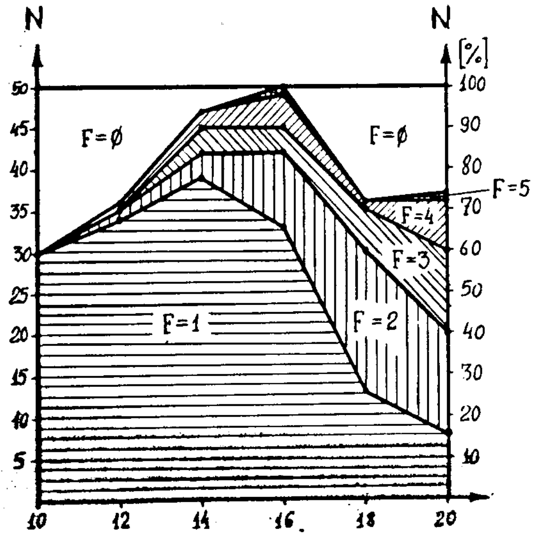
где L - частота появления диктора-центра в разных экспериментах. Дальнейший анализ показал, что приведенная классификация справедлива и для самих кластеров и полностью с ней коррелирует. Принятый подход классификации дикторов позволил получить 5 "устойчивых" кластеров с дикторами-центрами Cp = 1,29,38,44,45; 4 "среднеустойчивых" кластера с Cp = 25,35,36,48, а также ряд "неустойчивых" и "одиночных" кластеров. Независимо от использовавшейся парольной фразы, перечисленные кластеры в большинстве присутствовали после классификации исследовавшейся выборки дикторов. Кроме этого, с изменением R изменялся их количественный и качественный состав. На рис. 1 приведен граф динамики развития кластеров при изменении радиуса кластера для парольной фразы из 7 слов. На рисунке хорошо виден процесс образования, развития и распада отдельных кластеров и общая картина их преобразований. Кластеры обозначены прямоугольниками, толщина которых пропорциональна количеству дикторов в кластере, а само их количество приведено цифрой в прямоугольнике. Из рисунка видно, что при малых радиусах (R = 12.0 ... 15.0) происходит процесс формирования кластеров, при R = 15...16 наступает диапазон их оптимальности, а при R > 16.0 начинается их дробление и распад на ряд новых малочисленных кластеров /6/.

Рис. 1.



При выборе оптимальной величины радиуса нами учитывались как и общая картина развития кластеров, представленная на данном рисунке, так и отдельные параметры, такие как количество дикторов, входящих во все кластеры при данном R - Sd и частота попадания одних и тех же дикторов в различные кластеры - F. Величина Sd выбиралась таким образом, чтобы, по возможности, незначительно превышать число исследовавшихся дикторов (50). А F выбиралась из соображений минимального количества повторов дикторов в различных кластерах. Т.е. максимальным должно быть количество дикторов, для которых F = 1, и минимальным - для которых F = 0 и F > 2. На рис. 2 приведена диаграмма распределения количества дикторов N по частоте их попадания F в различные кластеры при изменении радиуса кластера R для парольной фразы из 7 слов. По перечисленным параметрам, приведенным на рис. 1 и рис. 2, был осуществлен выбор оптимального R, который для рассматриваемого случая равен 16 усл. единицам.

Рис. 2.



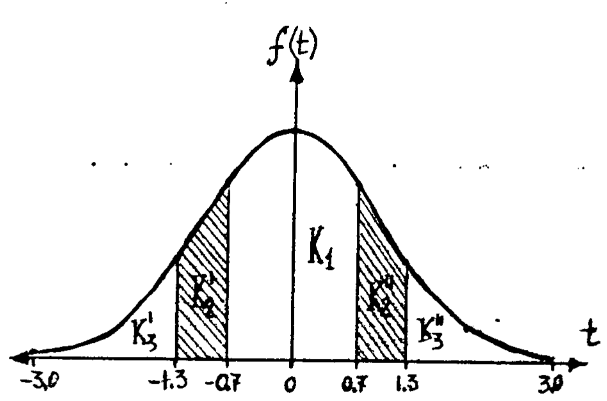
РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ ГОЛОСОВ

В результате проведенных исследований по кластеризации голосов различных дикторов можно сделать вывод, что произвольных пользователей по их голосу можно разделить на 3 группы:

- "устойчивые" к кластеризации дикторы;
- "неустойчивые" к кластеризации дикторы;
- "некластеризуемые" дикторы.

На рис. 3 показано распределение плотности F(t) от их нормированного отклонения t от средних кластерных эталонов. В первую из перечисленных выше групп дикторов, изображенной на рисунке областью K1, входит 50% дикторов, во вторую (K2',K2") - 30% и в третью (K3',K3") - 20%. Дикторов, вошедших в первые две группы, можно объединить в группу условно кластеризуемых дикторов. А дикторы, вошедшие в группу "некластеризуемых", работать с неадаптивной СРР не смогут. Их речь можно распознавать только пользуясь адаптивными СРР. Интересно, что "некластеризуемыми" становятся как "плохо сотрудничащие", так и "хорошо сотрудничащие" с СРР дикторы, которые в одинаковой мере, но с разной полярностью удалены от средних кластерных эталонов (зоны K3' и K3").

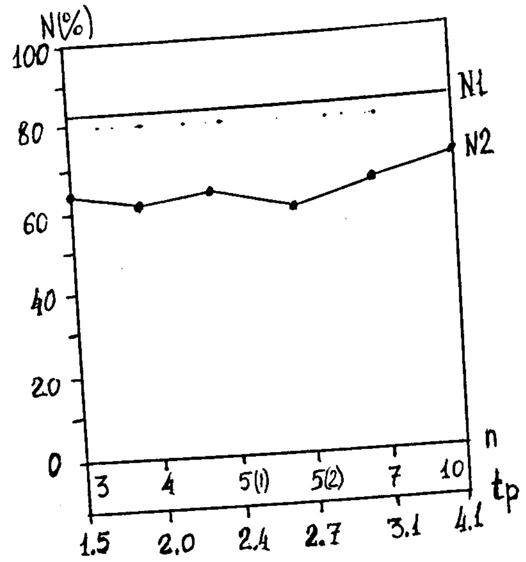
Рис. 3.



РЕЗУЛЬТАТЫ РАСПОЗНАВАНИЯ РЕЧИ ПРОИЗВОЛЬНЫХ ДИКТОРОВ

В экспериментах по распознаванию речи произвольных пользователей по кластерным эталонам приняло участие 50 дикторов, записывавших свои эталоны в банк эталонов речи, а также 10 новых дикторов. Каждый диктор произнес десять слов - цифры от 0 до 9. Усредненные результаты распознавания голоса всех дикторов по кластерным эталонам для каждой парольной фразы приведены на рис. 4. и обозначены кривой N2. Минимальная надежность распознавания получена при использовании парольной фразы из 4-х слов - 63.1%, максимальная - для парольной фразы из 10-и слов - 68.6%. Следует сделать оговорку, что все результаты получены для дикторов, не проходивших никакого предварительного обучения работы с СРР. По данным некоторых исследований при наличии предварительного обучения дикторов и их адаптации к работе с СРР, надежность распознавания их речи возрастает на 5...15% соответственно после 3...9-и часового обучения. Учитывая эти данные, можно надеяться, что средняя надежность распознавания речи "произвольного-обученного" пользователя нашей системой будет составлять около 85%. Это предположение подтверждается также тем, что для 2-х "сотрудничающих" дикторов, принимавших участие в эксперименте и образовавших "свой" кластер с Cp = 1, надежность распознавания во всех экспериментах была равной 100%.

Рис. 4.



n - количество слов в парольной фразе
tp - средняя длительность парольной фразы

Средняя надежность распознавания речи произвольных дикторов в адаптивной ССР, когда они имели "свои" эталоны, также приведена на этом рисунке и обозначена N1. Эта величина равна 82%.

ВЫВОДЫ

В результате проведенных исследований можно сказать, что к вопросу о распознавании речи произвольного диктора нужно подходить дифференцированно. Сначала необходимо определить возможность эффективной работы конкретного диктора с неадаптивной системой распознавания речи и только в случае позитивного результата этот диктор может приступать к работе с ССР. Определить "пригодность" произвольного диктора для работы с неадаптивной СРР можно по произнесенной им парольной фразе. Надежность распознавания речи произвольного пользователя сильно зависит от его подготовки к работе с СРР, иначе говоря его "сотрудничества" с системой. Проведенные исследования показывают возможные пути совершенствования неадаптивной СРР и несмотря на ряд трудностей при решении проблемы позволяют применять такие системы в ограниченных практических целях.

ЛИТЕРАТУРА

/1/ P. Fonsale "Connected-word recognition system using speaker-independent phonetic features", Proc. ICASSP-83, Boston, pp. 312 - 315.

/2/ Р.Я. Гумецкий, В.Н. Мазур "Диалоговая система обучения программированию и работе на ЭВМ с речевым запросом и ответом, ориентированная на массового пользователя", Труды сов.-франц. симпозиума "Акустический диалог человека с машиной", ИППИ АН СССР, Москва, 1984, стр. 51 - 54.

/3/ Р.Я. Гумецкий, В.Н. Мазур, В.А. Марченко "Система распознавания дискретной речи произвольного диктора". В кн.: "Автоматическое распознавание слуховых образов", Тезисы докл. и сообщ. АРСО-14, Каунас, 1986, ч. 1, стр. 94 - 95.

/4/ Г.Г. Гюльназарян, Ф.Е. Коркмазский, В.Н. Мазур "Использование систем автоматического речевого ввода", ВЦ АН СССР, Москва, 1986, стр. 19 - 25.

/5/ H. Sakoe, S. Chiba "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. on ASSP, v. 23, N1, 1978, pp. 43 - 49.

/6/ В.Н. Мазур "Исследование кластеризации дикторских голосов для распознавания речи произвольного диктора", Тезисы докл. и сообщ. АРСО-14, Каунас, 1986, ч. 1, стр. 72.

The Realization of Semantic Focus and Language Modeling

RODOLFO DELMONTE

Istituto di Linguistica - Università di Venezia

Ca' Garzoni-Moro - S.Marco, 3417 - VENEZIA(I)

ABSTRACT

Italian is a language in which discourse level informational strategies are easily detectable at sentence level. When arguments of a certain predicate do not constitute new information they are adjoined as clitic to the front of the verb; subject arguments constituting the theme of a discourse or text are left unexpressed. All relevant information on the contrary is highlighted by means of a variety of structural means: these are usually accompanied by phonetic signals mostly at the level of intonational contours. Semantic focus can be characterized by phonological structure, syntactic structure and pragmatic or full semantic representation. Only emphatic and contrastive focussing requires pragmatic or full semantic representation: this is not generated by available grammatical components of rule systems for speech synthesis, currently presented in the literature.The two remaining levels of representation, the phonological and the syntactic ones, enable a system of synthesis by rule to realize focus structure in most cases. Relevant semantic information is passed on to the syntactic component from the lexicon, which must be highly articulated. The remaining components activated in a system for synthesis by rule are the morphological and the phonological ones.

Phonetically speaking, the focussed constituent can be characterized by a peak with Low or High tone, aligned with word-stress, accompanied by a preceding H/L tone and sometimes followed by a L tone in coincidence with an Intonational Group boundary. Intonational Groups (IGs) constitute the higher phonological structure and are defined on a syntactic-semantic level, as the root sentence including the higher S node and its complements and modifiers. Moreover, we found out that to obtain a satisfactory definition of focus the highest-lowest peak in F_o value is not sufficient as an acoustic correlate. Focus is defined as a relation over two adjacent tonal assignments, in terms of the rate/s of change of the F_o curve.

INTRODUCTION

In a previous paper[1]we distinguished between Phonological Focus (FF) which gives rise to unmarked Focus Assignment Rules (FAR), and Logical Focus (LF) which gives rise to marked FAR. The former constitutes a case of default sentence level rule which associates a certain basic pitch contour with each Intonational Group(IG). Basic intonational contours of a certain language are usually defined generalizing over a set of illocutionary types (or tunes as defined in[2]) which are language-specific. In Italian there are at least the following: declaratives, questions, exclamatives and parentheticals. IGs constitute the higher phonological structure and are defined on a syntactic-semantic level, as the root sentence including the higher S node and its complements and modifiers.

Logical Focus (LF) is conceived as the pitch induced by syntactically governed discontinuities of constituents which can and usually are - affected by discourse level rules, as to their interpretation. These structures are however detectable at sentence level and give rise to a syntactic representation in which grammatical functions are assigned to constituents which do not occupy their canonical position in superficial or constituent structure. FF and LF generate focus structures which define the boundary of a sense unit at a discourse grammmar level: with FF focus structure includes the arguments of the predicate as they are normally associated by lexical frames, where syntactic or functional subcategorization, selectional restrictions and other feature information is listed for each lexical entry. In the case of LF this is also taken into account, plus the marked structures of Italian in terms of syntactic discontinuities. No pragmatic or extragrammatical knowledge is required, however, since no emphatic or contrastive structures are generated by the rules.

We take for granted that the system will generate an adequate structural description of marked structures(but see[3]). In order to investigate its relations with an acoustic-phonetic model of focus structure wee built a test set made up of sentences inlcuding the following structural types:

1. Neutral declarative followed by a subordinate hypothetical clause;
2. Topicalized version of 1.

3. Clitic left dislocation version of 1.
4. Clitic right dislocation version of 1.
5. Sentence with an Extraposed Subject NP;
6. Sentence 1 with Postposed Subject;
7. Sentence 1 with Inverted Subject;
8. Cleft construction;
9. Wh- question;
10. Yes-no question.

Sentences have been read aloud by an expert phonetician who repeated them until he judged to have performed the best rendering: F_0 and short-term power (both on a log scale) were computed each 10 ms at the CSC of the University of Padua.

Sentences are listed below with underneath their phonological marking:

```
1. Gli industriali devono pagare i decimali
            HL*                H*      H*LL%
 se vogliono che le trattative continuino.
                                L%
2. I decimali gli industriali devono pagare
     H*L                                L%
  se vogliono che le trattative continuino.
                                L%
3. I decimali gli industriali devono pagarli
   H L*             H L*            H*L L%
  se vogliono continuare le trattative.
                           L%
4.Gli industriali devono pagarli i decimali
         H L*           H* L              L%
  se vogliono che le trattative continuino
                             L%
5. Questo accordo non possono accettarlo
         H L*    H*        H            H*LL%
   i sindacati
       L%
6. Devono pagare i decimali gli industriali
   H        H   H      H*L                L%
 se vogliono continuare le trattative.
                            L%
7. Devono pagare gli industriali i decimali
   H        H                H   H*L      L%
   se   vogliono   la   continuazione   delle
            trattative.
                L%
8. Sono i decimali che gli industriali non
              H*L L%
    vogliono pagare.
              L%


9. Chi hanno detto che hanno intenzione di
    H    H        H*L L%          aiutare gli
industriali?
                L%
10. Hanno detto che avrebbero aiutato i
     H     H*L          H*   H     H*
terremotati gli industriali?
  H   H*L            L H%
```

Sentences 1. and its variants can be translated roughly as follows: "The industrialists must pay the decimals if they want the negotiations to continue"; sentence 6 as follows: "The unions cannot accept this agreement"; sentence 9 as follows: "Who did the industrialists say they intended to help?" and finally 10 as follows: "Have the industrialists said they intended to help the earthquake victims?"

As to the underlying phonological model, the reader is referred to [1].

In Pierrehumbert system [2], only two tones in combination make up the intonational contour specification: T* where T=H,L, the star indicates alignment with the prominent syllable. As a first approximation we adopt P's binary notation, plus the tone associated with IG's boundaries: T%=H,L where H% is usually associated with yes/no questions and L% marks the end of non-interrogative IG's. As in her system, when focus is associated to a prominent syllable there is a couple of tones which appear, as follows: [+Focus]---> H*l; HL*. It must be noted that the other two allowable sequences (L*H, LH*) are less frequent in Italian, or belong to emphatic and contrastive utterances which we do not take into account here.Also, we did not see the need for introducing a phrase accent, which should accompany the final nuclear pitch accent as happens in English.

ANALYZING THE DATA

From a linguistic point of view we can divide sentences into two parts: the one following and the other preceding the focussed constituent.First of all we look at the sentence section following the focussed constituent, which on a first approximation we take it to coincide with the rightmost T*T/TT* tonal marking. The portion to be considered varies remarkably from one sentence to another: it is constituted by a subordinate clause in sentences 1. and 3.; the subordinate clause plus what remains of the major clause, once the topicalized constituent has been fronted, in sentence 2. the subordinate clause plus the constituent which has been extraposed, either the subject or the object NP of the main clause, in sentences 4. 6. 7.; the presupposed relative clause attached to the clefted constituent in sentence 8.; the extraposed NP subject in sentence 5.; and the right dislocated NP object in sentence 4.

All this sentence material can be treated homogeneously from an intonational point of view even though it contains syntactic and semantic elements differing quite markedly from one another. These components of the intonational structure can be opposed to the material which precedes the focussed constituent/s which we discuss below. The phonetic characterization of post-focus linguistic elements can defined as follows: there is a downstep pattern in the F_0 contour which reflects a somewhat global line starting from an upper limit and reaching a baseline value about 5 half-tones below it (hereafter h.t.)[1].The declination line associated with each such portion of F_0 patterns does not lend itself easily to defining a constant decaying rate. In fact, lowering seems to apply randomly to prominent/non-prominent syllables looking at sentence stretches of a certain syllable length. Local variations may take Phonological Words[2] as their domain, with the only restriction that local F_0 jumps cannot override F_0 jumps of the previous PW. The first sentence, declarative, is made up of two IG's, the first of which ends with the main sentence and the second with the subordinate clause. Only the main clause contains focussed material, i.e. the assertion of the underlying semantic proposition; the subordinate expresses an hypothesis based on given information.In sentences moving the focussed constituent to the front, after FAR has applied, the declination line is set at approximately .5 h.t. above the final value(L%). Also these sentences (see 2 and 3) are made up of two IG's, the first of which ends with the main sentence and the second starts with the subordinate clause. The only noticeable difference from the simple declarative consists in the decrease in the final lowering at the end of the first IG: the degree of final lowering is higher in the simple declarative than in its marked variants and this is due to differences in semantic representation. In the former case, the main clause contains an assertion and the whole proposition constituted by the main predicate and the subject of predication are elements of focus structure: the pitch range correlated with the main sentence is higher than the one correlated with the subordinate clause. Marked variations of this utterance concentrates the predication onto a single constituent which marks focus structure: in sentence 2 it is the object NP, as in sentence 8; it is the VP in sentence 6 and the subject NP in sentence 7, and so on.As in [5] focus is the representation of the variable x such that P(x), where P(x) is a predication in x corresponding to the dominant or main Verb.What is needed then is a condexing rule to associate the predication with the entity in focus, as in [6]: Coindex NP and X where X = an AP, PP, NP, VP or S. Coindexing tells us which thing x is being predicated about. In case of sentence 2, a topicalization, what we have is:

2i. [[$_{NP}$i decimali]$_i$ [$_{VP}$devono pagare gli industriali]$_i$]

X P(X)

FOCUS=X=[$_{NP}$i decimali]

As Berwick rightly remarks: "there is certainly not much in the way of constraint in this proposal. What is missing is the machinery telling us which NP's and X's are to be coindexed"(ibid.,53). This would require discourse structuring rules, obviously; but at sentence level a lot can be done in Italian on the basis of syntactic structuring, as discussed above.

We are left with the portion of the intonational contour which precedes the focussed constituent. From a phonetic point of view, to achieve a satisfactory definition of focus it is not sufficient to look at peaks in the pitch contour. Variability in the topline -- or the maximum value for the F_0 contour in a phonological phrase - which can be constituted either by a peak, H, a maximum, or a fall to a very low pitch, L, a minimum in the pitch range of a given intonational contour in absolute terms, do not constitute the correlate of the focussed constituent. Other factors not relatable to focus can contribute to the creation of peaks, such as the length of the utterance or the beginning of a new discourse. We found and verified in synthesis experiments, that the steepness of the dipping following/preceding the focussed segment (usually a syllable), i.e. the rate/s of change or number of h.t. for the segment/s constituting the sequence relevant to the definition of focus structure, is the viable discriminating correlate of focus. In this way focus is defined as a relation over two adjacent tonal assignments, in terms of the steepness of the dipping of the F_0 contour. If we look at our examples, we find easily that in the first portion of the sentence there are two or even three combinations of T*T/TT* - and indeed, potentially there could be an infinite number. Only if we adopt our criterion we can account for sentences in which two or more constituents seem to be structurally marked and semantically relevant in the overall informational structure. This is the case of sentences 3,4,5 in which a constituent is moved to TOP position or is left/right dislocated and is bound to a resumptive clitic within the sentence, as shown in:

iii. [$_{S"}$[$_{TOP}$[$_{NP}$Questo accordo] [$_S$non [$_{NP}$e] [$_{VP}$possono [$_{VP}$[$_V$accettar] [$_{cl}$lo]]]] [$_{NP}$i sindacati]]]

The constituent in TOP does not count as new information as is the case with topicalized sentence 2. Rather, it qualifies as secondary focus even though it has been fronted: primary focus is associated with the VP and is marked as H*LL% at the offset of the IG.

The grammatical representation is thus confirmed by the data we collected in that focus is characterized by three features: a L/H peak/fall, aligned with word-stress, accompanied by a preceding or trailing H/L tone followed by a L tone in coincidence with an IG boundary(not necessarily), the steepness must be the highest in the sentence. If we look at the steepness we have the following data: in sentence 3. HL*=6 h;t. whereas H*LL%=8 h.t.; in sentence 4. HL*=3 h.t. but H*L=7 h.t.; in 5. HL*=4 h.t. but H*LL%=8 h.t.; in 2. the steepness associated with H*L=9 h.t.; in 8. H*LL%=8 h.t.; in 6. H*L=9 h.t. and finally in 7. H*L=9 h t.

We shall concentrate now on the two interrogatives: the wh- question in 9 and the yes/no question in 10. As to 9 we note that the wh- word constitutes the questioned object and the NP subject "gli industriali" is extraposed beyond three sentence boundaries, as shown below:

ii. chi [shanno detto] che [shanno intenzione] di [saiutare] ...

The intonational contour of the wh- question is clearly identifiable in that it doesn't possess a final peak nor a single peak at the onset: in wh- questions all the fronted constituent is in focus and is raised to a H plateau. What follows is a very steep F° drop: 10 h.t. in our examples. This pattern sharply separates the remaining sentence portion, which is uttered on a low declination line. It must be remarked that wh- words do not possess word stress unless they are contrastively emphasized - so they build a PW with the following head in this case the word "detto".

On a semantic level, wh- questions are partial questions and the H portion of the sentence is solely constituted by the questioned material, the remaining part of the question no longer constituting part of the question because presupposed or already known. On the contrary, yes-no questions are total questions and the whole sentence is uttered homogeneously on a H level.

[1] This characterization of F° variations in terms of half tones has been suggested to me by G.A.Mian and G.Tisato; each half-tone corresponds to $\Delta F/F=6\%$.

[2] We define a Phonological Word as a structural component of IGs made up of one stressed lexical element, the head of the PW, preceded by as many unstressed lexical elements as there are within a Phonological Phrase. Phonological Phrases in turn correspond to major syntactic constituents(see Selkirk, 1984).

BIBLIOGRAFIA

[1] Delmonte R.(1983), A Phonological Processor for Italian, in Proc. 1st Meeting European Chapter of the ACL, Pisa, 26-34.

[2] Hirschberg J., J.Pierrehumbert(1986), The Intonational Structuring of Discourse, Proc.24th Annual Meeting of ACL, Columbia Univ.New York, 136-144.

[3] Delmonte R., G.A.Mian, G.Tisato(1986), A Grammatical Component for a Text-to-Speech System, Proc.ICASSP 86, Tokyo, 45.2.1,2407-2410.

[4] Selkirk E.(1984), Phonology and Syntax, The MIT Press, Cambridge Mass.

[5] Berwick R.(1983), Introduction: Computational Aspects of Discourse, in M.Brady & R.Berwick(eds), Computational Models of Discourse, The MIT Press, Cambridge Mass., 27-106.

[6] Williams E.(1980), Predication, Linguistic Inquiry, 11(1), 203-238.

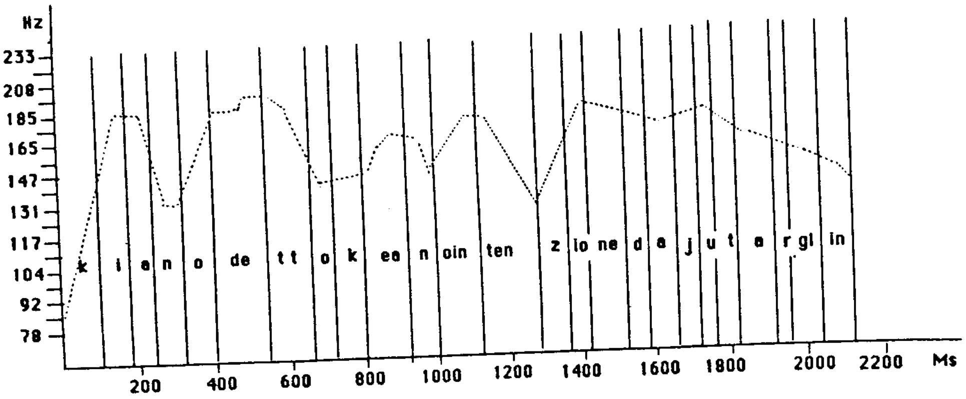


Fig.1 F° plot of wh- question

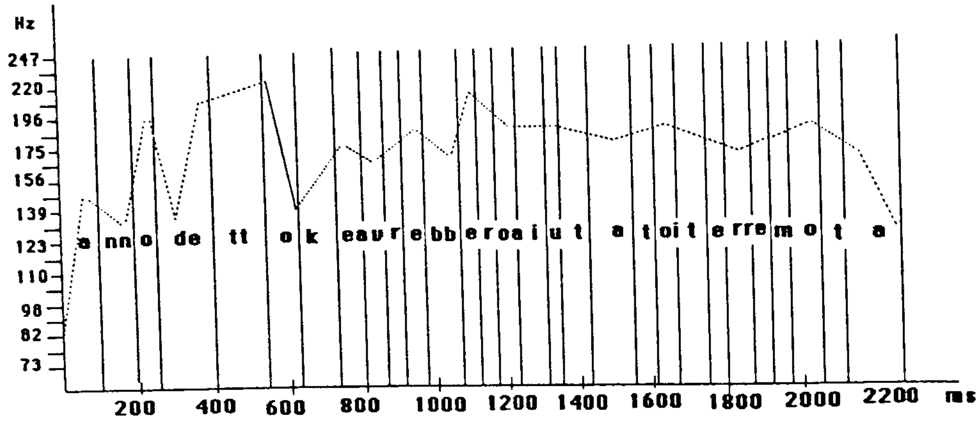


Fig.2 F° plot of yes/no question

TERNARITY IN PIRAHA PHONOLOGY

DANIEL L. EVERETT

Dept. of Linguistics
State University of Campinas
And
The Summer Institute of Linguistics
Caixa Postal 129
Porto Velho, RO, BRAZIL 78900

ABSTRACT

Rules for Pirahã primary stress, stress shifts in morphemic combinations, and extrametricality crucially refer to ternary feet, requiring us to admit ternarity as a primitive of metrical theory.

INTRODUCTION

A central concern of linguistic theory is to be able to describe the range of relevant phenomena within parameters sufficiently restricted so as to provide a meaningful explanation of the data. Consequently, we must resist the temptation to introduce novel theoretical devices unless absolutely required by the facts. In metrical phonology, for example (cf. [7]; [4]; and others), most researchers would agree with Hammond's [5] (p.193) assertion that "...bounded feet ... are maximally binary." This means that we would need only binary and unbounded feet in our theoretical tool box. It is tempting to speculate that if this is true it is the result of a deeper principle, viz., that heads must be adjacent to their domains. This would then elevate the notion of adjacency to the position of a cross-modular organizing principle, since, for example, the importance of adjacency in the syntax has been noted by various researchers (e.g. [1]; [11]). However, in this paper, a preliminary report on research in progress ([3]; cf. also [2] and [4]), I argue that primary and secondary stress placement in simple and morphophonologically derived words in Pirahã, an Amazonian language, crucially depends on the postulation of ternary feet. Corroborating evidence for this analysis is adduced from extrametricality. This analysis is important for phonological theory in that it provides the clearest evidence to date that bounded feet are not maximally binary and that ternarity must be admitted as an underivable theoretical primitive (cf. [6] for a suggestion that ternarity can be derived, based on the erroneous conclusion that all ternary trees are amphibrachs).

PRIMARY STRESS

The first evidence for ternarity is found in Piraha's rule of primary stress placement:

(1) Primary Stress Placement: Stress the rightmost token of the heaviest syllable type encountered in the rightmost three syllables of the word (—— = 'primary stress'; ═══ = 'secondary stress'; . = 'syllable boundary'. See [4] on the determination of syllable weight in Piraha):

(2)	.?a.ba.gi.	'toucan'
(3)	.?a.ba.pa.	'Amapa' (city name)
(4)	.bii.sai.	'red'
(5)	.ho.aa.gai.	'species of flower'
(6)	.ka.pii.ga.ii.to.ii.	'pencil'
(7)	.pia.hao.gi.so.ai.pi.	'cooking banana'
(8)	.kao.ai.bo.gi.	'evil spirit'

In examples like (7) and (8), where a heavier syllable (CVV) occurs to the left of the antepenult, rule (1) will overlook it, seeing only the final three syllables.

Stress is realized phonetically by some, but not all, speakers as intensity. Its phonological relevance is strongly supported by two optional, low-level rules:

(9) [+vd] ⟶ ([-vd])/ following stress
(10) syl ⟶ (Ø)/ following stress

We can derive the restriction of (1) to the final three syllables via the algorithm in (11):

(11) Tree Construction: Build a right-dominant, ternary, Obligatory-Branching (OB) foot (See [5] for a discussion of OB feet) at the right margin of the word.

Conditions: (a) The rightmost syllable of the tree must dominate a segment leftadjacent to]. (b)This algorithm applies from right to left.

MORPHEMIC COMBINATIONS

Not only will (11) correctly account for primary stress facts, it also derives the facts of secondary stress and stress shift in morphemic combinations ('[...]' = morpheme; '(...)' = phonological foot'; a. = base form; b. = derived form):

(12)a. [.?a.pi.pai.] [.ho.ao.ba.]
'watch' 'give'
b. ([.?a.pi.pa.]) ([.ho.ao.ba.])
(13)a. [.?a.pi.bai.] [.tio.hio.?io.]
'proper name' 'next'

b. ([.?a.pi.ba.]) ([.tio.hio.?io.])
(14)a. [.ka.hai.] [.?o.ga.ba.gai.]
'arrow' 'want'
b. ([.ka.hi][o.)(ga.ba.gai.])
(15)a. [.bao.sai.] [.bii.sai.]
'cloth' 'red'
b. ([.bao.sa.][.bii.sai.])
(16)a. [.?a.pa.pai.] [.?ii.ta.ha.]
'head' 'hurts'
b. ([.?a.pa.pa.]) ([.ii.ta.ha.])

As seen in these examples, resyllabification occurs in noun + adjective and noun + verb sequences, following deletion of the final vowel in the noun and the initial /?/ in the verb. Secondary stresses are produced by constructing a rightdominant phrase tree over the resultant sequence. These processes, in conjunction with postlexical, ternary foot construction produce the stress changes between the a. and b. examples above. Example (16) shows that stress shift cannot be explained via 'stress clash avoidance' (cf. [10]). Examples (14) and (15) show that the algorithm in (11) does not stop at [. They further illustrate the necessity of condition (a) in (11), since material from the leftmost morpheme has been incorporated into a foot with material from the rightmost morpheme. That is, an independent foot could not be formed at] because after the rightmost tree is constructed, there is no segment left in the noun which is adjacent to] (segments cannot simultaneously belong to separate trees since this would result in "crossing association lines" - out in just about anybody's theory).

EXTRAMETRICALITY

Extrametricality facts offer independent evidence for (11) (Note that the following data also appear to support the proposals in [9], wherein it is claimed that extrametrical syllables may be overlooked by certain rules yet still be relevant to other metrical processes or representations). In Pirahã, the nominalizer /-sai/ may not be stressed when phrase final, although it is always relevant for determining the ternary domain of (1) ('{...}' = extrametrical):

(17) a. [.?oi.boi.bii.{sai}.] 'sp. of fish'
cf. b.*[.?oi.boi.bii.{sai}.]
(18) [.?i.bi.{sai}.] 'hammer'
(19) a. [.?ii.to.pi.{sai}.] 'remover'

cf. b.*[.?ii.to.pi.{sai}.]

In (17), since { -sai} is extrametrical we would normally expect it to be irrelevant to the rule (1) above, falsely predicting that /.?oi./ will receive stress. Condition (a) of (11) correctly stresses /.bii./ To account for this, we can assume a filter along the lines of:

(2Ø) * ...{σ}]Ø

It should also be observed that examples like (18) eliminate an alternative hypothesis, namely, that only the final /i/ of -sai is extrametrical /-sa{i}/, since this would incorrectly stress this word as an oxytone rather than a proparoxytone.

CONCLUSION

In this paper, an analysis of stress placement in Piraha has been presented which demonstrates the necessity of enriching metrical theory to include ternary constituents. This means that either the notion of adjacency is not as important to linguistic theory as previously thought or that we must weaken our conception of it to include systems like Pirahã. Unfortunately, the data presently available on the world's prosodic systems is too scarce in my opinion to favor either possibility.

REFERENCES

[1] Emonds, Joseph, The Unitary Nature of Syntactic Categories, Foris, 1986.
[2] Everett, Daniel, "Pirahã", IN: Desmond Derbyshire and Geoffrey Pullum, (eds.), Handbook of Amazonian Languages, Mouton de Gruyter, pp. 2ØØ-326, 1986.
[3] Everett, Daniel, "Ternary Constituents and Multiple Obligatory Branching Feet in Pirahã", in preparation.
[4] Everett, Daniel and Keren Everett "On the Relevance of Syllable Onsets to Stress Placement", Linguistic Inquiry 15 7Ø5-711, 1984.
[4] Hammond, Michael, Constraining Metrical Theory: A Modular Theory of Rhythm and Destressing, Indiana University Linguistics Club, 1984.
[5] Hammond, Michael, "The Obligatory Branching Parameter in Metrical Theory", Natural Language and Linguistic Theory 4, 185-228, 1986.
[6] Levin, Juliette, "Generating Ternary Feet", mss., University of Texas, Austin, 1986.
[7] Liberman, Mark and Alan Prince, "On Stress and Linguistic Rhythm", Linguistic Inquiry 11, 511-562, 1977.
[8] McCarthy, John and Alan Prince, Prosodic Morphology, in preparation.
[9] Poser, William, "Invisibility", GLOW Newsletter, February, 1986.
[1Ø] Prince, Alan, "Relating to the Grid", Linguistic Inquiry 14, 19-1ØØ, 1983.
[11] Stowell, Timothy, On the Origins of Phrase Structure, unpublished Ph.D. dissertation, Massachusetts Institute of Technology.

Three Classes of "+" Boundaries

Kenneth W. Church
AT&T Bell Laboratories
Murray Hill, NJ, USA

It is well-known that English morphology has two classes of affixes: "+" morphemes such as *in+*, *ad+*, *ab+*, *+al*, *+ity* and "#" morphemes such as *un#*, *#ness*, *#ly*. The two classes differ in a number of respects, including: (1) Etymology: "+" morphemes are (often) historically correlated with Latin; "#" with German and Greek, (2) Stress Assignment (e.g., *parênt+al* vs. *pârent#hood*), and (3) Word Formation: + morphemes can attach to bound morphemes (e.g., *crimin-* as in *criminal*); # cannot (**criminhood*). This paper will extend this reasoning in dividing the first class into three parts, Ia, Ib and Ic (see table).

Class Ib contains what we generally think of as "typical" + boundary forms (e.g., *parênt+al*, *divin+ity*), both with respect to stress assignment and word formation. It will be argued here that Class Ia obeys a different set of word formation rules and that Class Ic obeys a different set of stress assignment rules.

The notion of compositionality provides a unifying theme across classes. Just as it is often observed that "#" forms have compositional semantics and stress assignment (e.g., *divine#ness* means "the state of" composed with "divine"; the stress of the whole is the concatenation of the stress of the parts) unlike "+" forms (e.g., *divin+ity* has religious implications that cannot be attributed to its parts; the stress of the whole is not the concatenation of the parts because of stress retraction), we would want to say that Class Ia is less compositional than Ib which is less than Ic which is less than II.

1. Word Formation Rules (WFR)

Aronoff proposed two distinct types of word formation rules in his thesis [Aronoff]: stem based wfr and word based wfr.

- Stem Based WFR: *subsume/subsumption, consume/consumption, resume/resumption, expense/expensive, conduce/conductive*
- Word Based WFR: *nominate/nominee, nominate/nominal, feminine/feminism.*

Stem based wfr rules relate pairs of words sharing one of a short (100-1000) list of latinate stems, e.g., *fer, mit, sume, duce, scribe*, whereas word based wfr apply to a large (possibly open) class of forms, often ending with *-ate* or some other archaic affixes such as: *-ine, -uli, -us, -um* that may be stripped off or "truncated" as part of the word formation process. Aronoff distinguished the two types of word formation rules in order to account for the fact that some generalizations, especially productivity and allomorphy, are clearly associated with stems, whereas other generalizations are associated with words.

This paper will use Aronoff's distinction in order to separate Class Ia from other "+" boundary forms. First, though, it may be worthwhile to review Aronoff's reasons for hypothesizing two types of word formation rules.

1.1 Productivity

The contrast in productivity between stem based and word based wfr is very striking. Note that there are very few gaps in stem paradigms:

	0	0 (pp)	-ion	-ive
duce	adduce	adduct	adduction	
	deduce	deduct	deduction	deductive
	conduce	conduct	conduction	conductive
	educe	educt	eduction	eductive
	induce	induct	induction	inductive
	introduce		introduction	
	produce	product	production	productive
	reduce	reduct	reduction	
	seduce		seduction	seductive
	transduce		transduction	
scribe		conscript	conscription	
	describe	nondescript	description	descriptive
	prescribe	prescript	prescription	prescriptive
	subscribe	subscript	subscription	subscriptive
ceive	conceive	concept	conception	conceptive
	deceive		deception	deceptive
	perceive	percept	perception	perceptive
	receive	recept	reception	receptive
here	adhere		adhesion	adhesive
	cohere		cohesion	cohesive
	inhere		inhesion	inhesive

	+ Boundary			# Boundary
	Class Ia	Class Ib	Class Ic	Class II
Examples	ion, ive, ent, or, ory	ity, ic al, ian	ize, ee, itis,ism, ist istic, ment, mental	ness, wise hood, ship
Etymology	Productive in Latin	Norman French	Scientific Literature and Enlightenment	Anglo-Saxon
Stress Retraction	+	+	−	−
Attaches to	stems	bound/free	bound/free	free

In contrast, word based alternations are full of gaps. For example, the word based *-ate/-ee* alternation (e.g., *nominate/nominee, designate/designee*) is limited to just a few cases; the vast majority of words ending with *-ate* do not have variants ending with *-ee*.

1.2 Allomorphy

Stem based word formation rules attempt to capture both productivity and allomorphy generalizations. In stem based forms, allomorphy (e.g., *scribe* vs. *script*) is purely a function of the stem and the suffix, and is independent of derivational history (cyclicity), prefix, part of speech, semantics, phonology, etymology, dialectical variation, etc. In contrast, allomorphy may have more complicated sources in word based forms. Consider, for example, the word *education* which does not follow the stem based pattern found in *adduction, deduction, conduction, eduction, induction, introduction, production, reduction, seduction* and *transduction*, because *education* is derived from the word *educate*, not from the stem *duce*. This example illustrates that derivational history can play an important role in explaining allomorphy, but only in word based derivations, and not in stem based derivations.

Mark Aronoff noticed that stem based allomorphy depended only on the stem and the suffix and attributed this fact to (the mythical) Ben Moshe.

> "The form of the suffix is never determined by a specific word. It is never the case that one verb in a given root will allow one variant, and other verb in the same root a different variant. The form of the suffix is root governed, that is, morphologically governed. There are no exceptions to this. It is the first law of the root originally discovered by the great Semitic grammarian ben-Moshe (ms) [sic] and called Ben-Moshe's First Law.
>
> We will illustrate ben-Moshe's first law in (28) with the root *sume*. The variant of *ion* which appears after *sume* is +*tion*:" [Aronoff, p. 102]

(28)	subsume	subsumption	*subsumation
	consume	consumption	*consumation
	resume	resumption	*resumation
	presume	presumption	*presumation
	consume	consumption	*consumation [sic]
	assume	assumption	*assumation

Aronoff uses Ben Moshe's Law to cover both cases like *sume/sumption* above where the allomorphy alternation is extremely clear as well as cases like *vert/version* and *sert/sertion* where the allomorphy is somewhat more subtle. Note that the orthographic "t" in *invertion* is realized as /zh/ whereas the corresponding "s" in *insertion* is realized as /sh/. Aronoff attributes this distinction to the allophorphy of the stems *-vert* and *-sert*, and then observed that Ben Moshe's Law correctly predicts that this voicing contrast is maintained in related forms such as *diversion, conversion, perversion* which contain /zh/ as in *inversion*, and *desertion, exsertion, assertion* which contain /sh/ as in *insertion*.

Ben Moshe's Law can also be used to cover quantity changing allomorphy as in *confide/confidence*. The "Confidence Puzzle" is intriguing because *-fide* is heavy in *confide* (as evidenced by the long vowel) but light in *confidence* (as evidenced by the stress retraction before the weak retractor suffix *-ence*). Other stems also use allomorphy in order to change quantity (see table). Consider *-side* and *-pel*. Both change their underlying quantity before the suffix *-ent*. *-side* is underlyingly heavy, but acts light in *resident*, whereas *-pel* is underlyingly light, but acts heavy in *repellent*. Note that Ben Moshe's Law correctly predicts that the choice of allomorphy is independent of prefix. The same light *-side* found in *resident* also appears in *president* and *dissident*; the same heavy *-pel* found in *repellent* also appears in *expellent* and *propellent*.

	Acts Light	Acts Heavy
Tense	-fide, -side, -spire, -tain, -stain, -cide, -pare	-hale, -grade, -plain, -flame, -vade, -praise, -rade, -suade, -place, -claim, -rive, -vive, -dign, -mise, -scribe, -quire, -vise, -prise, -fice, -pugn, -clude, -prove, -sume, -lude, -trude, -fuse, -plode, -close, -mote, -pose, -void, -join, -plore
Lax	-fer, -cel	-pel, -mit, -gress, -press, -cess, -cuss

1.3 (Almost) No Exceptions to Ben Moshe's Law

Ben Moshe's Law, according to Aronoff, is exceptionless. After some computer assisted investigation, it appears that the rule is, in fact, nearly exceptionless, if not completely so.[1] Many apparent counter-examples can be dispensed with by attributing the counter-examples to word based wfr, as opposed to stem based wfr, as we did in order to account for *education* which is problematic since most combinations of *duct* and *-ion* yield *duction*, not *ducation*. Aronoff himself uses the word based escape hatch in order to dispense with *consummatation*, which would ordinarily be a problem for Ben Moshe's Law, since *sume* plus *-ion* normally produces *sumption*, not *summation*.

> "Note that the form *consummation*, as in Shakespeare, is not an exception. Rather it is derived from the base *consummate*, by truncation." [Aronoff, p. 102]

Compensative is very much like *consummation*; *compensative* is formed from *compensate* via truncation, as opposed to *expensive* which is stem based and obeys Ben-Moshe's Law. *Friction/frication* also demonstrates the contrast between stem based and word based wfr. *Preventive/preventative* and *interpretive/interpretative* illustrate another class of (apparent) counter-examples to the law. Again, these apparent counter-examples can be accounted for by showing that one of the forms has an alternative source. In this case, *preventative* is from the latin frequentative; the frequenative *-ative* should not be confused with *-ive*.

[1] A dictionary search for orthographic sequences taking both *-ation* and *-ion* produced: *legation (legion), domination (dominion), oration (orion), duration, conversation, cessation, dilatation, natation, labefactation, retractation, affectation, dictation, volitation, indentation, notation,* and *potation*. Of these, *legation, domination, oration, duration, cessation, potation, natation* and *notation* are spurious. *Conversation* is from *converse*, not *convert*. *Indention* is an archaic form of *indentation*. *Dictation* is truncated from *dictate*. *Labefactation, retractation* and *volitation* are extremely rare forms, whose status is dubious. This leaves only *dilatation* and *affectation* as possible problems for Ben Moshe.[2]

In general, forms obeying Ben-Moshe's Law show up with a large number of latinate prefixes, as opposed to form like *compensative, expectation, education* and *preventative*, which violate the Law. Thus, for example, *conducive*, another exception to Ben-Moshe's Law (cf., *conductive, deductive, inductive, productive*), is not found with very many other prefixes (e.g., **educive*, **deducive*, **producive*). Exceptions are unlikely to show up with very many prefixes because prefixes are only productive on stems and these exceptions are word based.

1.4 Class Ia and Stem Based WFR

This paper provides additional evidence in favor of Aronoff's two types of word formation rules by proposing that some affixes (namely, Class Ia affixes) are (generally) associated with stem base wfr and that other affixes (namely, Class Ib and Ic) are associated with word based wfr. Note that Class Ia affixes (e.g., *-ion, -ive, -ent, -or*) are often found after latinate stems (e.g., *permission, permissive, confident, conductor*) but not generally after truncated morphemes (e.g., **nominion*, **nominive*, **nominent*, **nominor*). Similarly, Class Ib and Ic affixes (e.g., *-al, -ee*) are often found after truncated morphemes (e.g., *nominal, nominee*), but not generally after latinate stems **subsumal*, **subsumptal*, **subsumee*, **subsumptee*.

- *The Distributional Claim*: Class Ia affixes (e.g., *-ion, -ive, -ent, -or*) attach to latinate stems (e.g., *fer, mit, sume, duce, scribe*) whereas Class Ib and Ic affixes (e.g., *-al, -ity, -ic, -ee, -ism, -ist*) attach to words (possibly via truncation).

One of the consequences of this claim is that *feral, feric, ferity, ferrous* and *ducal* cannot be related to the latinate stems *fer* and *duce* because Class Ib affixes such as *-al*, *-ic, -ity* and *-ous* do not attach to latinate stems. This observation may be important for practical computer applications of morphological analysis to unknown words, especially for speech synthesis.

In addition, this distributional claim forces a form of level ordering [Kiparsky], [Mohanan]. Note that Class Ia affixes affixes can be found inside Class Ib affixes (e.g., *festivity, conventional*) but not the other way around (e.g., **fest+ity+ive*, **convent+al+ion*), because Class Ia affixes (e.g., *-ive, -ion*) must be attached to latinate stems and therefore, they cannot follow Class Ib affixes.

1.5 Multiple Class Membership

The distributional claim is somewhat weakened, unfortunately, by the fact that some affixes such as *-able* share membership in more than more class. Just as others (e.g., [Aronoff, section 6.2] have assumed that *-able* belongs to both "+" and "#", it will be assumed here that *-able* belongs to all three classes: Ia, Ib and Ic. The difficulty is that *-able* may or may not feed allomorphy, truncation and stress retraction:

- Allomorphy: (with) *circumscriptible, extensible, defensible, perceptible, divisible, derisible* (without) *circumscribable, extendable, defendable, perceivable, dividable, deridable*
- Truncation: (with) *educable, irrigable, navigable, regulable, demonstrable, operable, separable* (without) *educatable, irrigatable, navigatable, regulatable, demonstratable, operatable, separatable*
- Stress Retraction: (with) *cómparable, réparable, préferable*[3]

  (without) *compárable, repárable, preférable*

Aronoff assumed that forms which feed allomorphy, stress retraction and/or truncation contain a "+" boundary and that forms which block these processes contain a "#" boundary. The present proposal would assign *divisible* to Class Ia in order to account for the observed allomorphy, *demónstrable* and *coómparable* to Class Ib in order to account for the observed stress retraction, and *compárable* to class Ib in order to account for the observed lack of stress retraction.

2. Class Ic

The introduction suggested that Class Ib contains what we generally think of as "typical" + boundary forms (e.g., *parént+al*, *divin+ity*), both with respect to stress assignment and wfr. Section 1 argued that Class Ia obeys a different set of stem based wfr. This section will argue that Class Ic obeys a different set of stress assignment rules.

Within words, one expects to find stress clashes resolved by a rule which forces stressed syllables to alternate. Thus, for example, *degráde* plus *-ation* yields *dègradátion* with alternating stressed syllables, not *degràdátion* with the two adjacent clashing stresses. This prohibition against stress clashes applies to most "+" boundary forms (Classes Ia and Ib), but not to Class Ic. Note, for example, that *depàrtméntal* and *emplòyée* do not become **dèpartméntal* and **èmployée*, as would be predicted if these stress clashes had to be resolved.

Class Ic forms are also exceptions to most so-called "+" boundary rules. Note, for instance, the contrast between *concain+ism* and *profan+ity*. Tri-syllabic laxing, a typical "+" boundary rule, forces the tense vowel in *profane* to become lax in the Class Ib *profanity*, but tri-sylabic laxing does not apply in Class Ic and therefore the tense vowel in *concain* does not become lax in the Class Ic form *concainism*.

It will be assumed here that Class Ic forms are stressed much like compounds. *Assignee*, for example, is formed by combining the two pieces *assign* and *ee* with a right dominant foot [W S] so that the main stress falls on *ee*. Other Class Ic forms such as *cocainism* are combined with a left dominant foot so that the main stress falls on *cocain*.[4] In both cases, the internal metrical structure of the left piece is kept intact. Note that the

[3] By reasoning employed above to account for the Confidence Puzzle, *cómparable, réparable, prerable* may be considered examples of allomorphy along side *divisible*.

[4] Just as with compounds, it is extremely difficult to decide when to use a left dominate foot and when to use a right dominant foot. We will not attempt to address this question here.

stress on *sign* in *assign* is preserved in *assignee* and the stress on *cain* in *cocain* is presevec in *cocainism*; *assignee* does not become **àssignée*,[5] *cocainism* does not become **cócainism*, *employee* does not become **èmployée*, and so on. Similarly, the internal structure of the left piece is kept intact in *generalize*, *mineralize* and *federalize*, which do not become **genéralize*, **minéralize* and **fedéralize*, respectively.

The following table is presented as further evidence for the claim that Class Ic boundaries do not destroy metrical structure. The table lists a number of words ending in *-ist*, *-ism* and *-ize*. Notice that the stress pattern of the left piece is fixed across all three forms; for example, *romantic* has 010 stress in *romanticist* (010-0), *romanticism* (010-20) and *romanticize* (010-2).

-ist	-ism	-ize	Stress
romanticist	romanticism	romanticize	010
exorcist	exorcism	exorcize	10
humanist	humanism	humanize	10
antagonist	antagonism	antagonize	010
unionist	unionism	unionize	10
communist	communism	communize	10
militarist	militarism	militarize	100
terrorist	terrorism	terrorize	10
systematist	systematism	systematize	100
stigmatist	stigmatism	stigmatize	10
dogmatist	dogmatism	dogmatize	10
hypnotist	hypnotism	hypnotize	10

In this respect, Class Ic affixes differ from most other "+" boundary affixes which induce stress retraction. Strong retractors (e.g., *-ate*, *-ation*) often mung metrical structure: *design* (01) / *designate* (102). Even weak retractors (e.g., *-ent*, *-ant*, *-ence*, *-able*, *ance*, *al*, *ous*, *ary*) can modify metrical structure: *confide* (01) / *confident* (100). Class Ic affixes are unusual, because they do not induce either mode of stress retraction.[6]

Many so-called cyclicity arguments can be used as further evidence that Class Ic boundaries do not destroy metrical structure. Consider *capitalistic* and *militaristic*, where it has been noted [Withgott] that the /t/ can flap in *capitalistic* but not in *militaristic*, presumably because *capitalistic* comes from *capital* where the /t/ flaps, whereas *militaristic* comes from *military* where the /t/ does not flap. These facts are completely consistent with the observation that *-istic* is a Class Ic affix and that Class Ic affixes do not destroy metrical structure. The same flapping facts hold across a wide number of Class Ic affixes; *capitalist*, *capitalism*, *capitalistic*, *capitalize*, *capitalization*, *capitalitis* and *capitalite* all flap, unlike *militarist*, *miliarism*, *militaristic*, *militarize*, *militarization*, *militaritis* and *militarite*.

In conclusion, this section has argued that Class Ic cannot be stressed the same way as other "+" boundary forms and therefore they should be assigned a separate class. The previous section argued that Class Ib requires its own word formation rules and therefore, it, too, should be assigned its own class.

[5] *Designee* might be considered a counter-example to the claim that Class Ic boundaries do not destroy metrical structure. The contrast between *dèsignée* and *assignée* is accounted for by noting that *designee* is truncated from *designate* (and keeps that structure), whereas *assignee* is formed from *assign* (and keeps that structure).

[6] Admittedly there are a few forms ending in *-ist*, *-ism* and *-ize*, where the affix does not appear to be stress neutral (e.g., *immunize*). These forms are extremely problematic for our proposal since they appear to display classic "+" boundary stress alternations (e.g., *immúne/immunize*).

References

Aronoff, M., *Word Formation in Generative Grammar*, MIT Press, Cambridge, MA., 1976.

Chomsky, N., and Halle, M., *The Sound Pattern of English*, Harper and Row, 1968.

Kiparsky, P., *Lexical Morphology and Phonology*, unpublished ms., MIT, 1982.

Liberman, M., and Prince, A., *On Stress and Linguistic Rhythm*, Linguistic Inquiry 8, pp. 249-336, 1977.

Mohanan, K., *Lexical Phonology*, MIT Doctoral Dissertation, avaliable from the Indiana University Linguistics Club, 1982.

Withgott, M., *Segmental Evidence for Phonological Constituents*, unpublished Dissertation, University of Texas, Austin, 1982.

Appendix: Lexicon of Stems and Affixes

- Archaic Affixes (Victims of Truncation): ate, us, um uli, ii, ae, ine, ar, ure
- Class Ia: ion, ation, ive, ative, ent, ence, ency, ant, ance, ancy, or, ory, atory, able, ible
- Class Ib: ity, al, ality, ation, ative, ator, atory, ic, ian, able, ous, osity
- Class Ic: ist, ism, istic, itis, oid, ine (scientific), ate (scientific), ite (scientific) ite (non scientific), ish, able, ability, ee, eer, ette, ify, ize, ization, ification, ment, mental, mentary, mentarian, mentation, er, ery, ectomy, ology, olysis, ometer, imeter, ographer, oscopy, esce, ique, ess
- "#" boundary: wise, less, ness, hood, ship, way, land, ful, most, ly, man, ward, ling, like, dom
- Latinate Stems: act, bate, carp, cast, cave, cede, ceed, ceive, cel, cent, cept, cern, cess, cess, cide, cinct, cise, cite, claim, clam, cline, clive, close, clude, cluse, coct, crease, create, crete, cult, cumb, cur, cure, curse, cuse, cuss, dic, dict, dite, duce, duct, dure, empt, ept, face, fact, fame, fect, fend, fense, fer, fess, fest, fice, fide, firm, fit, fix, flame, flate, flect, flex, flict, flu, flux, form, fort, found, fract, front, funct, fuse, fute, gest, grade, gress, hale, here, hes, hibit, hort, hume, ject, join, joint, junct, lapse, late, lease, lect, lege, licit, lide, lige, line, lise, loc, lude, lume, luse, mand, mend, mense, merge, merse, miss, mit, mote, mount, mune, mute, nate, note, nounce, opt, pact, pand, panse, pare, part, peal, pel, pend, pense, place, plain, plan, plant, plaud, plause, plead, plete, plex, plic, plode, plore, plose, ply, pone, port, pose, posit, pote, pound, press, prise, prize, prove, pugn, pulse, punct, punge, pute, quest, quire, quisit, quit, rase, rect, rode, rog, rose, rupt, scend, sciss, scribe, script, sect, sense, sent, sert, serve, sess, sever, side, sign, sist, sole, solve, sorb, sorpt, spect, spense, sper, spire, spond, sponse, stance, stant, strain, straint, strate, strict, stroy, struct, strue, suade, suase, sult, sume, sumpt, sure, surge, tact, tail, tain, tect, tempt, tent, tense, tent, test, text, tin, tinct, tire, tone, tort, tract, train, treat, trice, trite, trorse, troverse, trovert, trude, truse, turb, twine, vade, vail, vase, vene, venge, vent, verge, verse, vert, vest vice, vide, vince, vise, vive, voc, voke, volve, vulse

THE SERBO-CROATIAN PHONOLOGICAL SYSTEM AND PROBLEMS IN PRESENTING IT TO FOREIGN LEARNERS

ČASLAV STOJANOVIĆ

The Institute of Foreign Languages
11000 Belgrade, Yugoslavia

ABSTRACT

The paper deals with the Serbo-Croatian phonological system, covering all its prosodic features occurring in words.

The four melodic accents, the lack of reduction, as well as the sequence and distribution of vowels present a lot of challenges to foreign learners.

Six consonants have a vocalic nature in the phonetic sense.

A more precise description has been provided here as to the place or manner of articulation of several consonants.

Problems of presentation are coupled with the wrong rendering of the SC sounds on the part of foreign learners whose mother tongues belong to various language groups.

INTRODUCTION

The Serbo-Croatian phonological system⊕ covers 5 relatively pure vowels and 25 consonant-type sounds. A more detailed analysis, however, reveals a few extremely interesting points.

VOWELS

Phonetic Description

There are 5 vowels. According to the place and manner of articulation they are as follows:

1/ front, close, unrounded [i]
2/ front, half-close, unrounded [e]
3/ central to back, open, neutral [a]
4/ back, half-close, rounded [o]
5/ back, close, rounded [u]

There are no nasalised vowels.

Suprasegmental Features

In some words there may be one of the four melodic accents:

1/ long-rising ´ 2/ short-rising `

3/ long-falling ^ 4/ short-falling "

The accents may occur:

a) only the falling ones:
in one-syllable words
b) only the rising ones:
in the middle of a word
c) any one of them:
on the first syllable
d) none of them:
on the last syllable

Some words bear no accent, having unaccented long or short vowels only.

A theoretical problem could be raised at this point:
Could one - following an arithmetical transaction -come to a conclusion that there are virtually 25 vocalic phonemes in Serbo-Croatian?

Reduction

By definition, there is no vowel reduction in standard SC. In some subdialects, however, there is a lot of reduction, even elision, of some unaccented vowels. Cf. standard v. reduced:

S: Jesam li ti rekao šta ima da bude?

R: S'm ti rek'o bre štima da bidne?

S: Što ćeš raditi večeras?

R: Što'š radit' većeras?

In the teaching process, on the contrary, emphasis is laid on the length of every unaccented long vowel, particularly for grammatical reasons, e.g., in the plural genitive: žénā (of women). For practical purposes, a horizontal line is placed above the long vowel.

Sequence

By definition, there are no diphthongs in SC, but only vowel clusters, that is,

● All the 30 phonemes are always spelt in the same way as they occur in the written language. The Cyrillic Alphabet, as used in SC, is phonemic.

sequences of two vowels, each of which has retained its syllabic value, e.g.,

radio (ra-di-o) , video (vi-de-o)

There are some interesting cases. Some men-of-letters spell some words differently, e.g., some spell the last syllable of the word meaning "tomato" with an "ai" cluster = paradaiz, whereas some spell it with a vowel + consonant group "aj" = paradajz. Taking into account the exact pronunciation of that word, the "ai" cluster could be treated as a diphthong.

In this connexion, a theoretical question could be raised:
Are there diphthongs in SC after all, now that there are foreign words which have become integral part of the language? E.g., auto (au-to).

Distribution

Every one of the SC vowels can occur in all the positions within a word.

Problems of Presentation

Theoretically, the description of any vowels embraces the following features of articulation:
1/ the position of the tongue
2/ the shape of the lips
3/ the position of the soft palate

Practically, it is impossible for the learner to judge the position of the tongue or that of the soft palate. Only the shape of the lips is visible. It is advisable to mention that there are no nasalised vowels.
Although an empiric approach is of major importance in language teaching, it would prove useful if the teacher of SC knew the vocalic system of the native language of his student.
Foreign learners tend to mispronounce the SC sounds due to the prejudices of their mother tongues.

In Arabic there are only three vowels: /i/, /a/, /u/. Arabian students confuse between SC /o/ and /u/, and between /e/ and /i/,respectively. They mispronounce the words "bio" (was) and "beo" (white) in the same way: [biu].
The Spanish have no problems, as there is the same vocalic system in Spanish: "este, hijo, hasta, hombre, lunes".

In English there are 12 relatively pure vowels, 8 diphthongs and 5 triphthongs.

Instead of the SC /a/ the English often pronounce RP No.5 vowel: sam = [sɑ:m], RP No.4: Ana = [æ nə] , RP No.10 : čak = [tʃʌk] or RP No.12:sam= [sɜm] ,the last one being the problem of reduction. All SC vowels can be reduced that way.

In Russian there are a few varieties of all the vowels / i, e, a, o, u/ as well as two reduced vowels, [ə] and [ɩ] . The Russian vowels are never as long as the SC ones. There are also two degrees of reduction in unstressed syllables. That is why Russian learners of SC tend to make SC vowels shorter or reduced. They introduce their varieties of vowel sounds. E.g., selo = село́ [sjiˎɫo] ; bik = быk [bɨk] ; delo = де́ло [ˎdjeɫə] ; eho = э́хо [ˎɛxə]; nauka=нау́ка [naˎukə].

There is no melodic accent like in SC, yet there is a strong dynamic stress (ударе́ние), indicated by the mark ́ . It can occur in all the positions within a word. The SC accent and the Russian stress are related in a specific way in the words of the same meaning:
a)syllables bearing falling accents in SC correspond to R stressed syllables, e.g., mȇso = мя́со; ȕlica = у́лица;
b) syllables bearing rising accents in SC differ more from R stressed syllables, insomuch as in Russian the stress is shifted on to the following syllable e.g., rúka = рука́; sȅstra = сестра́.
There is also a minor stress in Russian marked ̀ : ра̀дио-переда́ча.
The Russians generally stress such similar SC words in their way.

Foreign learners find it very difficult to distinguish between the SC accents. While contrasting pairs of accents can be of some help, e.g. a long-rising accent against a short-rising one:

žénā (of women) v. žȅna (woman),

mere imitation of the accents occurring in words which cannot be contrasted does not necessarily prove efficient.
Foreign learners are usually aware of length distinction only.Because of the presence of both accented syllables and unaccented long vowels in a single word foreign learners often cannot determine which syllable bears the accent. Therefore, at each new attempt they may lay the accent on another syllable, e.g.: "vrabācā"(of sparrows) may become: vra-ba-ca, vra-ba-ca, or vra-ba-ca.

As for reduction, the English generally take care of the accented syllable only. Thus, they pronounce "A Happy New Year" " Srećna nova godina " as [ˈsretʃnə ˈnɔvə ˌgɔdnə].

There are diphthongs in French, Spanish and English. Speakers of these languages tend to make the SC vowel sequences into diphthongs, the English even into triphthongs occasionally: bio = [biou]. As for distribution of vowels,there are 5 English vowels which cannot occur finally: RP No.3/e/, RP No.4/æ /, RP No.6 /ɔ/,RP No.8/u/, RP No. 10/ʌ /. For that

reason,the English make similar SC vowels longer or into diphthongs:
ovde = [ˎɔvdei]; znate = [ˎznæ tæ i];
neko = [ˎnekɔ:] or [ˎnekou·];
robu = [ˎrɔbu:]; čeka = [ˎtʃekɑ:].

CONSONANTS

There are 25 consonants in SC.Fifteen of them are voiced, ten are voiceless.

Plosives

Bilabial:	[p] [b]	(spelt p,b)
Dental:	[t] [d]	(spelt t,d)
Velar:	[k] [g]	(spelt k,g)

The voiceless stops are not aspirated.

Affricates

Alveolar:	[ts]	(spelt c)
Palato-alveolar:	[tʃ] [dʒ]	(spelt č,dž)
Alveolo-palatal:	[tɕ] [dʑ]	(spelt ć,dj)

Nasals

Bilabial:	[m]	(spelt m)
Dental:	[n]	(spelt n)
Palatal:	[ɲ]	(spelt nj)

Phonetically, they are nasal vowels.

Apical sounds

Alveolar: rolled	[r]	(spelt r)
Alveolar: flapped	[r]	(spelt r)

Rolled [r] is formed by rapid intermittent taps of the tongue tip against the teeth ridge.It is coupled with the central neutral vowel [ə] .
The formation of the flapped alveolar [r] is similar, but involves one tap only.

Laterals

Dental:	[l]	(spelt l)
Palatal:	[ʎ]	(spelt lj)

Being continuant and non-fricative, the two laterals are vowel sounds.

Fricatives

Labio-dental:	[f] [v]	(spelt f,v)
Alveolar:	[s] [z]	(spelt s,z)
Palato-alveolar:	[ʃ] [ʒ]	(spelt š,ž)
Velar:	[x]	(spelt h)

Semi-vowel

Palatal:	[j]	(spelt j)

Being a variant of [i] , this sound is phonetically a vowel.

Theoretical Problems

Several consonants have not been described adequately in previous works, because they are evaluated on the phonetic and phonological levels simultaneously. Thus, in all SC sections on Phonetics, the nasals,laterals and semi-vowel are dealt with first as they function in language, their phonetic description being almost neglected. There is no mention of their vocalic nature.
From the point of view of phonetic description, the nasals, laterals, and semi-vowel are vowels. Linguistically, they occur marginally in the syllable, thus, they are included in the consonantal category on functional grounds. At times, however, owing to the phonetic context, they are accompanied by friction, thus becoming allophones of consonantal nature. /2/
In SC books the tongue is divided into the apex, front and back. According to the IPA books the apex is subdivided into the tip and blade.That has in turn brought about a more precise division of palatal sounds. (Cf. č,dž with ć, dj).

The two apical sounds stand in complementary distribution, thus being two allophones of one phoneme. Rolled [r] occurs between two consonants;thus, it is considered a vowel on the linguistic level.Flapped [r] is considered a vowel if it precedes a consonant initially. In other positions, it is considered a consonant. The central neutral vowel[ə] does not exist in SC, consequently SC speakers are not aware of its existence and role in releasing the apical sounds.

In SC books [s] and [z] are treated as dental fricatives. Actually, they are alveolar sounds because the air-stream escapes by means of a narrow groove in the centre of the tongue, causing friction between the tongue and the alveolar ridge. /2/

Errors Made by Foreign Learners

Some of the consonants present a lot of problems to foreign learners. They tend to stick to the phonemes existing in their own mother tongues.
In Russian there are even doubled consonants, which are rather long. Russian speakers tend to pronounce similar SC words in their way: masa = [ˎmassə].
In Russian and German,voiced consonants are devoiced finally. That is why R and G learners devoice SC final consonants:
nov = [nof] ; vod = [vot] .
The problem of palatalisation is worth considering. In Russian there are a lot of consonants which can have both hard (non-palatalised)and soft (palatalised) varieties. There are 18 palatals in R, and only 9 in SC. Only 3 palatals are the same in both languages:[ɲ],[ʎ],[j]. Three are similar: [tʃ] is softer in R, whereas [ʃ] and [ʒ] are harder. SC ć , dj and dž do not exist in Russian.
R learners often palatalise SC sounds in those positions which call for palatalisation in R: delo = [ˎdjeɫə] .

Plosives. Arabian speakers confuse [p] and [b]: Palestina = Balestina.

The Germans devoice stops in all positions: biti=piti; dobro=topro; gad=kat. The English often hear and reproduce SC /p,t,k/ as [b,d,g] due to the lack of aspiration in the SC voiceless stops:

Papić= Babić; klatno= gladno.

All Germanic languages have regressive assimilation, whereas Slavic languages have progressive assimilation. Compare the SC and English renderings of " pet banana" (five bananas):

SC rendering: ped banana
E rendering: pet panana

The Spanish confuse [b] with their bilabial fricative [β] : subota= [suβota].

Affricates. Initial [ts] is difficult for the English. They sound [s] instead:

cvet = svet

The French can never pronounce [ts], so they utter [s] instead: lonac = lonas.

All foreigners (and a lot of SC speakers too) find it almost impossible to distinguish between č and ć, and dž and dj, respectively. R Iearners neutralise č and ć into [tʃ]: kuče,kuće=[ˌkutʃe] , and dž and dj into [dj] :džak,djak=[djak].

It is advisable to point out to the double articulation of these palatals. The tip, blade, and rims of the tongue touch the upper alveolar ridge and side teeth. At the same time, the front of the tongue is raised towards the hard palate, less(for č,dž)or more(for ć,dj). That explanation should help anyone, particularly the Greeks, who pronounce [ts] instead of [tʃ] and [tɕ], and [dz] instead of [dʒ] and [dʑ]: čuti=[tsuti]; noć=[nots]; džep=[dzep]; Djura=[dzura].

Nasals. The French and Chinese do not sound final [n], but they nasalise the previous vowel: slon = [slõ].

Russian, French, Italian, Spanish and Portuguese speakers find it easy to pronounce [ɲ], yet it is very difficult for the Germans and English. They sound it as [n] or treat it as [n/j], e.g., konj= [kon]; njuška=[nu:ʃka,n/ju:ʃka].

Rolled [r]. This sound is easy for the Spanish and Scottish. The French and Germans sound uvular [r] instead. The English pronounce RPNo.11 vowel[ə:] coupled with a kind of [r]:krv=[kə:rv]. The Japanese confuse [r] and [l]:

gorak= golak; red = led

The English pronounce t+r and d+r like their affricates [tr] and [dr]:

tri = [tri:]; drug = [dru:g].

Laterals. As [l] involves a double articulation (a tap of the tongue tip against the upper teeth, coupled with a rise of the middle of the tongue toward the hard palate), the French shift the second articulation forward, thus forming their clear [l]: šal = [ʃal+].

While pronouncing final [l], the English sound their dark [ɫ], by shifting the second articulation backward: šal=[ʃɑ:ɫ].

The Russians always use their dark [ɫ]: lampa=[ˌɫampə];sila=[ˌsjiɫə];bal=[baɫ].

The Japanese confuse [l] and [r] :

Split = Sprit

Russian, Italian, Spanish and Portuguese speakers find it easy to pronounce [ʎ], yet it is difficult for French and German learners. Their rendering is always [j] : ljubav = [ˌjubav]. The English rendering is either [l] or [l/j]:

ljubav = [ˌlu:bav], [ˌl/ju:bav].

Fricatives. Initially, Spanish learners confuse [v] and the stop [b]:vino=bino. Arabic speakers do so medially as well:

navijam = nabijam

The Greeks and Spaniards have no [s] , [z], [ʃ], or [ʒ] sounds, but the sounds in between [s] and [ʃ] , and [z] and [ʒ], respectively. Thus, they pronounce:

uzela sam sok = [użela ṡam ṡok]

doživela sam šok=[dożivela ṡam ṡok].

The Spanish often assimilate [s] or [z] to the following nasal or lateral:

pismo = [pim̥mo]; razlog = [rallog]

The French have no [x], so they drop it:

hitno = itno

The English have the glottal fricative [h], so they hear the velar fricative [x] as the voiceless velar stop [k], and pronounce it thus: hleb=kleb; Čeh= Ček.

REFERENCES

/1/ S.N.Dmitrenko,Section on Phonology, " Grammar of Contemporary Russian", USSR Academy of Science,Moscow,1970

/2/ A.C.Gimson, "An Introduction to the Pron.of English,Arnold,London, 1965

/3/ J.Jovanović,Z.Vukadinović,"SC Elem. Course for Foreigners",FLI,Bgd,1982

/4/ K.L.Pike, "Phonetics",Un.Mich.,1943

/5/ K.L.Pike, "Phonemics",Un.Mich.,1947

/6/ "Principles of the IPA";IPA,UCL,1964

/7/ Č.Stojanović, "Pronunciation Errors Made by SC Learners",.FLI, Bgd,1967

/8/ Z.Vince, Section on Phonetics, from "Language Advisor", MH ,Zagreb,1971

/9/ R.Aleksić,M.Stanić, Section on Phonetics, "SC Grammar", Textbook Publishing Company of Serbia,Bgd.,1962

o
o o

A PHONOLOGIC-PHONETIC COMPONENT OF A DYNAMIC LINGUISTIC MODEL

Georgi Chikoidze

Linguistic Modelling Laboratory
Institute of Control Systems of the GSSR Academy of Sciences
Tbilisi, Georgia, USSR 380042

ABSTRACT

As a basic quantitative criterion for a phonologic description choice, an average phonologic code length of a text is suggested. Capabilities of such an approach are demonstrated on the example of a Georgian phonologic system.

A dynamic linguistic model is a system, which fulfils direct and inverse transformation of a "sense-speech (text)". A phonologic-phonetic component of the direct transformation (synthesis) produces to the given phonemes corresponding phonetic characteristics, i.e. descriptions, which must serve as an immediate basis for a choice of articulatory commands. A natural basis of a phonetic description is a set of articulatory features (in the case of synthesis), or their acoustic correlates (for analysis).

These processes are essential components of a complete linguistic model, being the basis of many important practical applications. According to this, conditions of simplicity and description economy acquire not only abstract-theoretical meaning but also practical value, stipulated by natural demands, made by technical realization of such systems. Technical realization of a phoneme in a linguistic model is its code, the choice of which, generally speaking, is arbitrary, not taking into account a trivial condition of noncoincidence of different phonemes codes. According to this, a problem of the code choice naturally arises, providing a chance to create simple and economical coding-decoding procedures, which direct and inverse phonologic-phonetic conversion.

Concretizing this code choice criteria, on the basis of quite general consideration we can suppose, that they may be reduced to the conditions of a code structure simplicity and to minimality of its some quantitative characteristics (of average length) and also to demands of simplicity of its correlation with phonetic characteristics. A binary code, for which a structure of coding-decoding procedure is represented by a dichotomic tree, obviously, possesses the simplest structure. Methods of agreement of such a code with frequency of coded symbols appearance, providing the code construction with minimum average length, are well-known. However, these methods are not intended for taking into consideration an additional condition, which is a requirement for simplicity and directness of relation of a phonological code to a phonetic description.

Evidently, this last demand is satisfied more completely by a code, a coding-decoding tree of which can simultaneously be regarded as the basis for a phonological system description in terms of phonetic features. By the tree of such a type, constructed on the basis of acoustic characteristics, a Russian phonological system is described in /1/. Terminal nodes of such a tree are phonemes; some phonologically meaningful binary features are connected with each nonterminal node, one of the branches, coming from it (in our case -left), is connected with the positive value of the feature, and the other - with the negative one.Now, by attaching a value "1" to each "positive" branch, and "0" to each negative one, for every phoneme can be produced a binary code, which is created in the process of passing the route from the root ("top") of the tree to the terminal node corresponding to this phoneme. A procedure of passing by such tree routes, leading from the root to terminal nodes, can be performed on the basis of the given code sequence, creating the corresponding phonemic features (a process of decoding), as well as on the basis of the given phonetic features succession, providing the construction of the corresponding code (a process of coding).

A set of phonetic features, associated with nodes of his tree, makes up some subset of a meaningful phonetic characteristics set, enough for distinguishing all the phonemes of the given phonologic system. Sets, possessing such a property, can be chosen by many different ways, and thus, some additional condition of minimality is imposed in order to determine some "marked" sets. In the majority of cases, and particularly, in /1/ and /2/ a requirement for minimum of number of different features, forming a set, serves as a restriction.

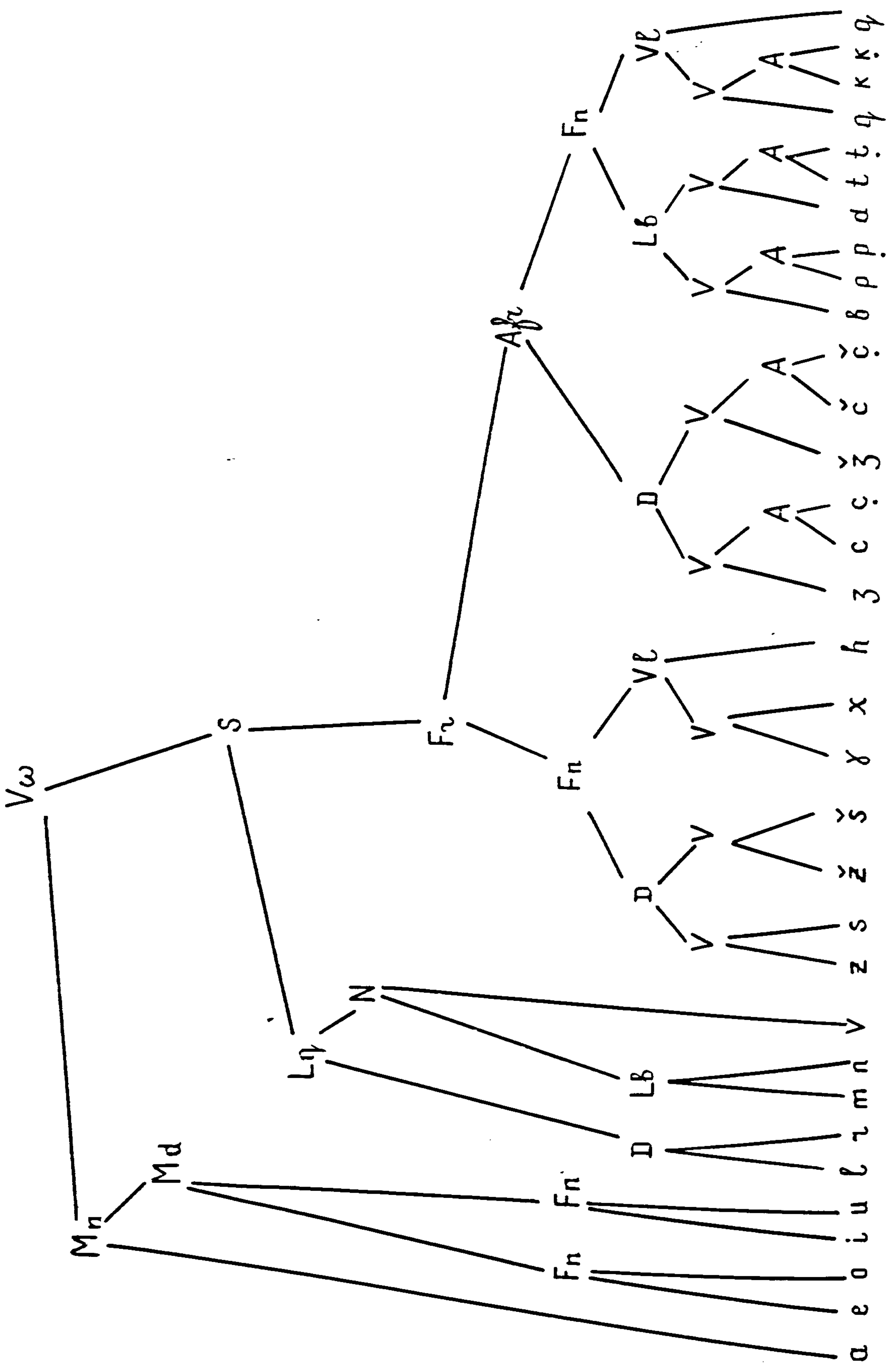

Vw
S
Mn
Md
Fn
Fn
Lq
N
Lb
D
Fr
Fn
D
D
Afr
Fn
Vl
Lb
Vl
V
A
a e o u i l r m n v z s ž š γ x h ʒ c c̣ ǯ č č̣ b p ṗ d t ṭ q k ḳ q̇

Fig. 1. The coding-decoding tree for Georgian phonemes.

However, taking a standpoint of the dynamic linguistic model, it should be recognized, that a condition of minimality of the phonemes codes average length, is more essential, because exactly this parameter determines the average number of steps of the coding-decoding procedure and, therefore, characterizes time expenses, connected with the phonologic-phonetic component of the model. A degree of this condition fulfilment for this set of features is convenient to determine by a value Q:

$$Q = \frac{L - H}{H} \; 100\%,$$

where L - is the average length of the phoneme code in the text, and H - is entropy of the phonemes distribution in the text, representing a theoretical lower limit of values L. Thus, Q is the redundant average length of the code in percentage of its theoretical minimum H, and hence, a system with less Q value must be prefered. As the orientator let us note, that for the Russian phonologic system, constructed in /2/ the value Q was 21%.

A tree of dichotomized articulatory features, describing a variant of the Georgian phonologic system, for which Q has the value of 10%, is attained on the basis of the given in /3/ distribution of Georgian phonemes frequencies in the text, to which the meaning of entropy H = 4,31 b.u. corresponds. The phonetic material used in the system construction almost entirely is taken from /4/, though in the choice of the definition for correlation of air passage common features, corresponding to the upper levels of the tree, some consideration were taken from /1/ and /5/.

The top feature of the tree on Fig. 1 is Vw vowelness, positive value of which corresponds to vowel phonemes a,e,o,u. Set of vowels, in their turn, is structurized by usual features of minimum (Mn) and mean (Md) raising of the tongue. Phonemes characterized by the negative value Vw, i.e. consonants, first of all, are divided into two subset by the sonority feature, (S). With common representative of sonors (l,r, m,n) a phoneme V is included in this class, that corresponds to the contemporary point of view on the Georgian phoneme, and also to separate remarks from /4/. First of all, from the class of sonors are distinguished liquids (Lq), and then nasals (N), and nonsonors are divided into fricatives (Fr) and nonfricatives, coinciding with the class of stops; the last in their turn - into affricatives (Afr) and nonaffricatives, representing by them a class of pure stops.

Configuration of the upper part of the tree and distribution of the corresponding nodes, almost entirelly reproduces gradations of the opening degree, defined in /5/: Mn corresponds to the sixth gradation, Md - to the fifth, and its negation ($\overline{Md}$) - to the fourth; then liquids (Lq) are characterized by the third degree of the opening, nasals (N) - by the second, fricatives (Fr) - by the first, and stops (Afr and $\overline{Afr}$) - by zero. Exception is the phoneme V, combining characteristics of a sonor and a fricative; on the scheme of Fig. 1 it occupies a sonor position near nasals, being a single representation of the additional to them class, that defines a degree of its opening, as an intermediate between the first and the second, but more close to the last of them. General structuralizing features in this zone are a top feature Vw, dividing a set of gradations of opening degrees into three upper and four lower gradation, and also S feature, dividing these last ones in two.

Below the considered zone there is a two-level zone of features of the formation place: the upper level is represented by a feature Fn with meanings "front" - "nonfront", and the next level - by a bilabiality Lb, a dentality D and a velarity Vl. Analogous to the scale of opening degrees regulates common features of air passage, features of the formation place correlates to the following consequent regions of the vocal tract: labial (Lb), dental ($\overline{Lb}$ = D), alveolar ($\overline{D}$), velar (Vl) and transvelar ($\overline{Vl}$), which can be realized as pharyngeal (the phoneme q) or laryngeal (h). Note, that uniting these two last' regions into one, we shall be able to say, that not only the positive but also the negative meaning of these features always points to one and the same region of the formation. The same is true for the feature Fn, the positive meaning of which always corresponds to the set of bilabial dental, alveolar and palatal regions, and the negative - set of velar and transvelar. Joining a palatal zone to the positive region of Fn also justifies this feature utilization for the opposition of front and back vowels, that conforms to the corresponding remark in /4/ about the equivalence of this opposition to the opposition: palatal - velar.

Two lower levels create features of voiceness (V) and aspiration (A).

Let us note some alternative possibilities of the tree structure choice, illustrating considerations, which have led us to the variant depicted in Fig. 1. So, for example, the phoneme V is related to fricatives on the concluding scheme of the Georgian phoneme classification in /4/. Equally with the already stated considerations, a choice of a position V was stipulated by the circumstance, that its inclusion in the class of fricatives rather deteriorates the value of Q. accepted by us, as a criterion, particularly, in this case Q = 11%. Further deterioration of the value Q is connected with the accepted in /6/ variant of the tree "top" construction according to the pattern /1/. In /6/ firstly vowels and liquids are opposed to other phonemes by the feature of vocality, and then liquids - to vowels like consonants to nonconsonants. In this case Q reaches 16%. On the contrary, concession to the traditional approach, apparently, is a refusal from the variant, opposing firstly fricatives and affricates to simple stops, and then fricatives - to affricates, since this variant provides lower value of the criterion: Q = 8%.

At the same time, economy of description basically defined by the value Q, must not conflict with its completeness, i.e. on the basis of the accepted scheme of the phonologic-phonetic transformation all the characteristics, necessary for the functioning of phonetic, phonologic and morphological rules, must be produced. So if the synthesis of the sound is provided by the modelling of the speech tract configuration in the process of the sound articulation, then it is necessary to enlarge, for example, the sonor characteristics by information about, common to them voiceness and their formation place, and also - to note such specific features as laterality of l and vibrantness of r; voiced consonants deafening in front of voiceless stops needs, differentiating of q by a voiced-unvoiced feature; finally, the formulation of the morphological rules of a stem truncation and contraction will be simplified by the feature, common to a and e, and so on. The most natural is including of all useful characteristics creation into the process of decoding. This can be expressed graphically by adding their symbols to the corresponding branches, for aims of minimizing a number of repetitions; such additions must be made on the maximally high level, for which they are relevant; so the additional feature of voiceness must be created, according to the above adduced examples, when passing the positive (left) branch, coming from S. Let us emphasize, that these additional symbols will be simply ignored when coding.

REFERENCES

1. M. Halle. The sound pattern of Russian. 's-Gravenhage, 1959, p. 19 - 46.
2. C. Cherry, M. Halle, R. Jakobson. Toward the logical description of languages in their phonemic aspect. "Language", vol. 29, N 1, 1953, p. 34 - 47.
3. D. Tzkitishvili. Statistics of binary letter combinations (digrams) in the Georgian language. Col. of Proceedings of the Institute of Control Systems of the GSSR Academy of Sciences "Machine Translation", XI-XII; 3, publ. "Metzniereba", Tbilisi, 1974, p. 141 - 146.
4. G. Ahvlediani. General Phonetics Foundation, publ. TSU, Tbilisi, 1949.
5. F. de Saussure. A course of general linguistics. "Satzekgiz", Moscow, 1931.
6. G. Chikoidze. A system of articulatory features and Georgian phonemes coding. Proceedings of the Institute of Control Systems of the GSSR. "Machine Translation", XII: 3, publ. "Metzniereba", Tbilisi, 1975, p. 5 - 37.

PROLEGOMENA TO DIACHRONIC PHONOLOGY

V.K. Zhuravlyev

Institute of Linguistics
Moscow

The problem of neutralization of phonological oppositions as a corner stone of phonology was for the first time put forward by count Trubetzkoy at the first Congress of phonetic sciences (1932).By that time Jakobson had published the first trial of historic phonology (1929) and clearly formulated its "Principles"(1931).

The XI Congress is proposed with the synthesis of these outstanding achievements of the 20th century linguistics which was expected for a long time. This synthesis makes it possible to construct the diachronic phonology paradigm which remained uncompleted up to now.

An unprejudiced analysis of the state of affairs in our science reveals striking contradictions between synchronic and historic phonology in general and the classical Prague concept in particular. The former one has been worldwide recognized and has become a kind of an epicentre of the 20th century linguistic thinking whereas the diachronic phonology has not won proper recognition even among specialists of the history of the language despite the fact that it is one of the first attempts of the special theory of structural transformations whose importance has been realized only nowadays.

It is realization of the central system-forming role of neutralization that made Trubetskoi revise all his earlier material in phonology, namely, all its notional apparatus (Viel, pp. 175-176; 183-188) and this made it possible to cauplete in 1933-1935, the construction of the paradigm of the general phonology as an integral science on phonological oppositions and conditions of removing these oppositions, i.e. neutralization. Jakobson's conception of historical phonology was established in 1927-1930. The fundamental concept of position was not properly ducidated in his work and,quite naturally, the concept of neutralization was not introduced at all. This concept as well as more recent "distinctive feature theory" (1952/56) is "a paradigmatic type phonology, and ignored the problems of syntagmatic relations" (Stankewicz, 1967, 394-5).

Many adherents of historical phonology, deliberately or instinctively and may be even for reasons of principle, focus their attention just on paradigmatics (Hoenigswald et al.). Martinet who paid great attention to neutralization could not, howerer, find a proper place for it in his diachronic phonology (1955), there is no place in it for the concept of position either. Nevertheless outstanding successes of the young-grammarians arc due to the fact that they tried to explain phonetic changes just by "phonetic environment" in positional conditionality, i.e. in syntagmatics. Therefore one more contradiction is revealed, namely, contradiction between historical phonetics and historical phonology.

Diachronic phonology could come into the world only having positively broken all ties connecting it with classical historical phonetics in whose depth it was generated. Jakobson gave a brilliant interpretation of the historical phonetics by Shakhmatov and contraposed "integral method" of historical phonology to "isolationism" of young-grammarians. From that time on "diachronic phonology is still in its infancy" (Wartburg, p.49). It can, at last, pass the adolescence age and assimilate the richest heritage of the previous generations. It is necessary to remove contradiction between historical phonetics and phonology, rehabilitate the postulate of immutability of phonetic laws and to bring all empirical material derived with their help into the most valuable capital of our science. It is necessary to remove the contradiction between the classical general and historical phonology by completing the construction of its paradigm.

From time to time attempts are made to give phonological interpretation of the phonetic law concept (A. van de Groot, J.Fourquet et al.), Sometimes it is stated that all phenomena covered by the Jakobson "mutation" formula can be due to an influence coming from the syntagmatic level "and any change in phonemic inventory" comes through a "rephonologization" of certain positional facts" (Ivić, p.52-53). Rather seldom the attention is paid to the role of neutralization in phonological changes: "from neutralization to neutralization the opposition disappeared completely" (Fourquet, p. 131). However attempts to remove cardinal contradiction of our science, to reconsider its entire notional apparatus and to perform the proper synthesis have never been made.

The proposed synthetic conception of phonology is based on neutralization as a nucleus of oppositions of the phonological system: there is an opposition (in position of differentiation) and there is no opposition (in position of neutralization). Neutralization connects paradigmatics and syntagmatics by means of positional (syntagmatic) removal of paradigmatic opposition. Being strictly synchronic neutralization is turned to diachrony, convergence or divergence, dephonologization of a neutralized opposition or phonologization of potential opposition in past or future. Thus it connects synchrony and diachrony removing, at last, the Saussure's antinomy.

Neutralization has actually turned out to be the most powerful system-forming factor. It integrates uniting phonemes and allophones, positions, oppositions and correlations, vocalism and consonantism as a single whole.

The centre control the periphery via the neutralization mechanism thies stimulating the corresponding phonetic laws as means of generation of allophonic variation aimed to create potential phonological oppositions which can increase the integrative force of central correlations (cf traditional concept of the "system pressure").

Chaos of accomodations becomes Cosmos of regular neutralizations determinated by the particular system as a tendency of growth of its integrative prorerties.Thus one more contradiction of our science is removed, namely, the young grammarians managed to find regularity in the past,in the history of the language and we cannot reveal phonetic laws in the present, in the observed synchronous state. Accomodations, assimilations, dissimilation and even neutralizations which were considered to be a destructive factor weakening the distinctive (differentiative) force now turn out to be the system integration factor. The loss in differentiation is compensated by the gain in internal integrity, coherence of the system.

Being very sensitive to integrative needs of the system neutralization, this "demiurge" of Trubetzkoy, gives rise not only to new allophones but also to the rules of their positional functioning at the given synchronous state of the language, i.e. creates the phonological essence of the socalled phonetic laws of the young-grammarians.

The neutralization mechanism employs quantitative and qualitative change in differentiation and neutralization positions thus performing convergence-divergence of phonemes and allophones as a main way of phonologization of potential opposition and dephonologization of obsolete ones. "Demon" of Polivanov "enables or disables" the phonetic laws by taking off the former allophones from the state of the complementary distribution and by removing their positional dependence (the law is disabled). It also selects those potential convergents and divergents from the allophonic variation which are able to take part in the following convergent-divergent process (the law is enabled).

Now it is not typology of correlations and not survey of the inventory of distinctive features but typology of neutralizations, mechanisms and rules of its performance that are put in the foreground of the diachronic phonology. Since it (typology of neutralizations) is constructed and tested using the material of different languages irrespective of their genetic affinity it can serve as a more reliable base of the diachronic reconstruction than isolated facts of similar changes in unrelated languages of the conventional typology.

To complete the paradigm of the diachronic phonology means to give the main role not to a phoneme or a distinctive feature, or an opposition, or a correlation, or a position and not event to a neutralization but, at last, to the phonological system as a whole, to its integrating properties and the system-forming factors and their dynamics. In this case an investigator will pay his attention not to the aspect of mutability (cf "phonetic changes" of traditional historical phonetics and phonology) but to stability, to dialectics of self-preservation and self-motion of the phonological systems. And only now it becomes possible to reveal the profundity of Trubetzkoy's idea that "the phonological evolution makes sense only if it is applied to the reasonable reconstruction of the system... Many phonetic changes are caused... by the necessity to form stability... to correspondence to structural laws of the phonetic system (1929, p. 65). Revision and intensification of "the integral method" of diachronic phonology make it possible to recover and enhance its explanatory function. The notional apparatus of the modern diachronic phonology allaws us to reconstruct continuous sequence of phenomena and processes of the phonological system history as a continuous chain of causal-resultative relations.

References

1. В.К.Журавлев.Диахроническая фонология. М., 1986.
2. V.K.Žuravlev. Zwei Paradoxa der modern Komparativistik. Indogerm. Forschungen, 91 Bd. 1986.
3. M.Viel. La notion de "marque" chez Troubetzkoy et Jakobson. P. 1984.
4. В.К.Журавлев. Роль нейтрализации в фонологических изменениях. - В кн.: Фонология. Фонетика. Интонология. - Материалы к IX Международному конгрессу фонетических наук. М. 1979.

ON METHODS OF RECONSTRUCTION

ALBERTAS STEPONAVICIUS

Department of English Philology
Vilnius University
Vilnius, Lithuania, USSR, 232734

ABSTRACT

The inventory of the methods of reconstruction must be extended so that it should comprise the methods of comparative, internal, graphic and external reconstruction. The scope of each method needs further specifications as well.

INTRODUCTION

The reconstruction of earlier states and processes, their absolute and relative chronology in the history of a language ranks among the central problems of diachronic linguisfics and diachronic phonology in particular. In a wide sense, reconstruction is synonymous with diachronic linguistics. In a narrower sense, reconstruction means techniques, or methods, of recovering earlier forms of a language. Traditionally, there are distinguished two methods of reconstructions: comparative (CR) and internal reconstruction (IR). However, the wide range of the technical means used in diachronic linguistics cannot be reduced only to these two methods. In addition to CR and IR there are theoretical as well as practical reasons for distinguishing the methods of graphic and external reconstructions (GR, ER). GR is in fact distinguished by Lehmann /1, pp. 63-81, 83/ when he discusses the use of written records as one of the methods of determining linguistic change. Milewski /2,pp.137-138/ distinguishes the "traditional philological method based on comparative analysis of old texts", whichkakin to GR. Birnbaum /3, p.97/ singles out ER, which as the first method of diachronic linguistics is differentiated between his three fundamental types of reconstruction as based on extraneous linguistic elements (borrowings, loan and foreign words, nonnative proper names, etc.).

GRAPHIC RECONSTRUCTION

In the case of alphabetic writing, GR is the most reliable method among all the possible methods of reconstruction. GR is especially effective when based on what is called phonemic alphabets (as opposed to morphophonemic alphabets), in such languages as Old Greek or Old English. The essence of GR is in establishing the graphs and the graphemes, the relationship between the graphs and the sounds, between the graphemes and the phonemes in the language of the texts under analysis (cf. the use of spelling evidence when establishing the sounds and phonemes of Old, Middle and Early Modern English, as in /4/). Thus alphabetic writing provides the most valuable evidence for the inventory and distribution of sounds and phonemes in a language at a certain stage of its development. Moreover, graphic evidence helps reconstruct sound changes and their chronological order, However much depends on what is actually reconstructed. Graphic evidence may be used in reconstructing paradigmatic and syntagmatic, segmental and prosodic, phonological and phonetic systems and changes, but such evidence is more scarce and less reliable for reconstructing prosody and phonetics. The loss of phonemes and oppositions is almost immediately reflected in spelling by indiscriminate use of formerly contrasting graphs, or by the use of one symbol instead of several initial ones, or by reverse spellings. The rise of phonemes and oppositions is usually reflected in spelling by the creation of new graphs, or by a contrastive use of two available graphs, though writing in this case is more conservative. Spelling usually reflects purely syntagmatic changes. Yet it is necessary to bear in mind that some phonological changes, both paradigmatic and syntagmatic, are not attested by spelling at all. First of all, this is true of many mutually related sound shifts which lead to replacement of oppositions and correlations (cf. the Great Vowel Shift, or the replacement of the consonantal correlation <u>voiced</u> vs. <u>voiceless</u> by the correlation <u>fortis</u> vs. <u>lenis</u> in Modern English, /4, §§ 194-196, 199/). Yet even in such cases occasional spellings may occur, indicating sound change of one type or another. Purely phonetic changes regularly are not reflected in writing, yet in special cases writing gives ample evidence of phonetic changes as well. Thus diphthongs as gliding phonemes are regularly spelt with digraphs. The choice of letters for the elements of gliding may indicate phonetic realizations of diphthongs, as well as changes in their phonetic realizations (cf. Old English diphthongs, /4, §§ 153-154/).

GR provides important evidence for other methods of reconstruction, so we may say that it precedes IR and CR; on the other hand, it may equally need a support by evidence provided by other methods of reconstruction.

INTERNAL RECONSTRUCTION

IR is based upon the comparison of genetically or structurally related elements from the same language and the same dialect. The method of IR is important in that taking no outside language into account it helps reconstruct earlier sound elements and patterns (quite recent and prehistoric ones as well) together with most important data concerning the distribution of the sound elements. IR helps establish, however, only relative, but not absolute, chronology. This method takes into account first of all morphophonemic alternations, such as Old English dæg - dagas(/æ / - /a/, /4, § 112/), fyllan - full (/ü/ - /u/, /4, § 142/), Modern English was - were (/z/ - /r/), frost - frozen, house - houses (/s/ - /z/), break - breach (/k/ - /č/), long - longer (/ŋ/ - /ŋg/), without regard to morphological classes. In this case the effectiveness of IR depends upon the paradigmatic similarity of alternating phonemes and the possibility to recover the conditioning phonological factors of alternation. The method may be complicated and restricted in its application (cf. such cases as bring - brought) and finally made altogether inapplicable by successive changes of sounds and morphemes, and, naturally, complete mergers of allomorphs. One has to admit that sometimes alternations exist as morphological interchanges from the very beginning without any sound change involved (cf. ablaut of the type sing - sang). IR may also be based upon the principle of pattern congruity. Such reconstructions considerably widen the scope and possibilities of IR and they may be no less cogent than those based on alternations (cf.

the different treatment of the "second fronting" in West Mercian and Kentish proceeding from the different patterns of the short vowels of the two dialects, /4, §§ 147-150/). We still remain within the limits of IR when we base our assumptions on relationships between subsystems of the sound structure, e.g., between prosody and segmentics, paradigmatics and syntagmatics, or on interlevel relationships between the sound structure and morphological, syntactical and even semantic patterns, as well as on typological maxims. Typological maxims impose two constraints on reconstruction: the sound changes must be typologically acceptable as processes and the proto-forms and the proto-language must be typologically acceptable in a static sense /5/. This broad treatment of IR is much in accord with Kuryłowicz's approach to it /6/.

Ideally speaking, IR should precede CR:in the first place pre-forms and pre-languages are established by means of IR and then the CR of proto-forms and proto-languages is carried out (cf. /7, p.156/).

COMPARATIVE RECONSTRUCTION

The traditional method of CR hardly needs any further elaboration (see, among other works, /8/). It is based upon the comparison of genetically related elements from cognate languages and dialects of the same language. Otherwise it may be said that CR deals with the facts of different dialects of the same language or different languages within the same language family. Moreover, it must be added that a contrastive treatment of evidence from earlier and later stages of the same language should be considered as belonging to the method of CR as well, for such evidence is drawn actually from different linguistic systems. The comparative method has proved to be of special importance in prehistoric reconstructions.

EXTERNAL RECONSTRUCTION

ER may be based on linguistic and non-linguistic data. Linguistic data may be provided by language contacts, in the form of borrowings, loanwords, foreign words and names. The interpretation of the descriptions by orthoepists in terms of modern linguistics may also be considered as a procedure of ER based on linguistic data. Non-linguistic data may be provided by archaeology, history, onomatopoeia (e.g., records of animal cries), etc.

CONCLUSION

From the methodological point of view, it is important and possible to distinguish and define more exactly four methods of reconstruction: GR, IR, CR and ER. Practically, however, it is possible to achieve reliable reconstructions only as a result of a combined use of several methods of reconstruction. In´a sense it is true that there is no "method of internal reconstruction as distinct from a method of comparative reconstruction" /9, p.116/, or from any other method of reconstruction.

REFERENCES

/1/ W.P.Lehmann, "Historical Linguistics: an Introduction", New York, 1962.

/2/ T.Milewski, "Językoznawstwo", Warszawa, 1969.

/3/ H.Birnbaum, "Problems of Typological and Genetic Linguistics Viewed in a Generative Framework" The Hague-Paris, 1970.

/4/ A.Steponavičius, "English Historical Phonology", Moscow, 1987.

/5/ H.M.Hoenigswald, "Notes on Reconstruction, Word Order and Stress"; Linguistic Reconstruction and Indo-European Syntax, Eds.P.Ramat et al. Amsterdam. 1980. pp.69-87.

/6/ J.Kuryłowicz, "Internal Reconstruction", Current Trends in Linguistics, Ed. Thomas A.Sebeok, vol.11, Diachronic, Areal and Typological Linguistics, The Hague-Paris, 1973, pp.63-92.

/7/ R.D.King, "Historical Linguistics and Generative Grammar", Englewood Cliffs, New Jersey, 1969.

/8/ H.M.Hoenigswald, "The Comparative Method", Current Trends in Linguistics, vol.11, pp.51-62.

/9/ P.Kiparsky, "On Comparative Linguistics", Current Trends in Linguistics, vol.11, pp.116-134.

ARTICULATORY PHONETICS AND RECONSTRUCTION VERIFICATION
(Indo-Iranian data)

D.I. EDELMAN

Institute of Linguistics, USSR Academy of Sciences
103009, 1/12 Semashko St., Moscow

Abstract

Articulatory phonetic data of living Indo-Iranian languages may be used to verify reconstructed subsystems of Indo-European, Proto-Aryan and Proto-Iranian phonological systems. Analysis of the articulatory aspect of historical changes helps to solve some disputable problems in the history of Indo-Iranian languages.

Using typological evidence to verify or correct data obtained through comparative-historical method allows to reveal systemic relationships between reconstructed units, including phonetic units, and to make the reconstructed system more probable /1/. Thus, articulatory phonetics and phonology of living Indo-Iranian languages discloses interesting typological parallels to various reconstructed subsystems of Indo-European, Aryan and Iranian protolanguages. This makes for a better understanding of the functioning of these subsystems in the synchronically viewed reconstructed protosystems and diachronic processes accompanying both the emergence of these systems and their subsequent changes.

One example of such parallels is a system of triads of consonants: "palatalized - simple - labialized" of the type C' - C - C°, which is observed in several living Indo-Iranian languages: in one of the Iranian languages - Yazghulami - in West Pamir (consonants ḱ - k - k°, ǵ - g - g°) /2/, in one of the Nuristani (Kafir) languages - Kati - in the province of Nuristan in Afghanistan (cerebral shibilants ṣ̌' - ṣ̌ - ṣ̌°, ẓ̌' - ẓ̌ - ẓ̌°) /3/ and in one of the Dardic languages - Kashmiri (consonants of almost all the zones and series, e.g. t' - t - t°, tʿ' - tʿ - tʿ°, d' - d - d°, ṭ' - ṭ - ṭ°... k' - k - k°... s' - s - s° ... h' - h - h°) /4/. These triads, particularly in Yazghulami and Kashmiri, are typological parallels to early Indo-European triads of gutturals, laryngeals and, possibly, sibilants /5/.

Attention to the articulation of the consonants forming these triads in the abovementioned languages helps to better understand the functioning and evolution of corresponding Indo-European triads.

"Simple" members of the triads of the "guttural" group (of the type k' - k - k°) appear to be unmarked members of the oppositions, and are represented by velar consonants characterized by wide variation and ability to merge with palatalized and labialized consonants in certain positions. With predominance of the palatal focus, the palatalized members of the triads easily shift to the front zone and become affricated (by the type k' → ḱ → č and so on). This accounts for the tendency towards turning into affricates and further changes of Indo-European palatal consonants: *ḱ > Aryan *č > *ç from which derive Nurist. c, Indo-Aryan and Iranian ś (its transition into s is a relatively later phenomenon which did not occur in all the Iranian languages). Depalatalization and a back-lingual shift k'>k is less frequent. The labialized members of the triads are represented by velar consonants with a second labial focus, not always synchronous with the main one: labial articulation may begin earlier than velar implosion and end later than velar explosion. Labiality may disappear completely or transfer to the neighbouring vowel. Its predominance and the change of a labialized consonant into labial is rare.

This evidence supports the possibility of the functioning, in early Indo-European, of triads of guttural consonants, and discloses the main principles of their change in the eastern area, i.e. in satəm languages: a) the shift forward of palatalized - palatal consonants (with affrication and possible assibilation), b) the loss by the labialized consonants of the labial articulation component leading to their merger with the "simple" consonants /6/. The reverse pattern in the frequency of processes, i.e. depalatalization of palatalized consonants and their merger with "simple" ones, and the predominance of the labial focus in the labialized consonants with their transition to the labial group - characterizes the changes in these triads in the western area, i.e. in centum languages.

Positional presentation of sibilants as shibilants in various languages (e.g. see /7/) points to the pattern of transition of Indo-European *s > š following *i, *u, *r, *k, *ḱ in satəm languages: appearance of a secondary focus (additional point of articulation) of *s - a palatal focus after *i, *ḱ and velar or postalveolar focus - after *u, *r, *k. The secondary focus brought about shibilants (of *š, *š type), the phonologization of which (and consequently, the phonologization of the opposition s ~ š /8/ occured much later, after the satəm group had diverged into a number of subgroups (partly even after the divergence of Aryan protolanguage).

The difference of articulation of *š in different regions of Proto-Iranian language brought about differences in sibilant subsystems of various Iranian languages. In the Western and North-Eastern subgroups, "soft" (palatalized) articulation of *š with a secondary palatal focus prevailed. The result was an appearance in these languages of a two-member opposition s - š (except Ossetic which lacks this opposition, and where the single phoneme /s/ has different dialectal realizations [s, ś, s', š]). In the South-Eastern subgroup, the influence of a substratum similar to the substratum for Indo-Aryan languages resulted in the predominance of "hard" articulation of *š, with a secondary velar focus, which, in this region, was associated with the cerebral phonological zone. As the opposition /č/ - /č̣/ developed in the same region, an "empty slot" for the "soft" /š/ appeared in the phonological system of these languages, which was later filled with positional variants of other consonants - reflexes of Proto-Iranian *ś, *č etc. The result was the establishment of a three-member opposition s - š - ṣ̌, to some extent similar to the Old Indian opposition.

One of genealogical features differentiating the East-Iranian language group from the West-Iranian - is a reflection of Proto-Iranian *b-, *d-, *g- as West-Iranian b-, d-, g- ~ East-Iranian v-, δ-, γ- in word-initial position. Individual exceptions in East-Iranian languages, such as the initial b-, d- instead of *v-, *δ-, can be explained by relatively late articulatory tendencies already within these languages themselves. Thus, Ossetic at some stage became to be characterized by the strengthening of articulation of word-initial voiced consonants. As a result, borrowings from Old Ossetic (Alan) into Hungarian display the complementary distribution, which existed at that period, of voiced stops and fricative consonants in word-initial and middle positions: b- : -v-; d- : -δ- (the latter being represented by Hungarian z). Significant in this respect, are cases of reflection of the Old Iranian resonant *u̯- in Ossetic b-, and almost complete absence in Modern Ossetic of original words with initial v-. This evidence shows that the articulatory tendency characterizing a one-word speech segment, i.e. articulatory "border mark", resulted in a "deviation" in historical development of consonantism, a "violation" of the phonetic law. A similar tendency is observed in Khotanese. In several East-Iranian languages (Yagnobi and Ishkashmi) "deviations" occur only in the reflection of *δ - and connected with the general instability of articulation of *ϑ, *δ in these languages area: *ϑ changes, relatively early, into t or s, and *δ- - usually into d- (probably, not without Tadjik influence). In the neighbouring area the unstable articulation of *δ brought about the transition *δ > l. This occured in Iranian languages - Pašhto and Munji, and also in one of the Nuristani languages - Prasun, - which shows that this phenomenon is regionally rather than genetically conditioned.

As to reflexes of East-Iranian *γ-, such deviations in the form of its reflection as *g- are non-existent. The reason of this is its articulatory characteristic: very early and virtually across the whole of Iranian linguistic area *γ shifted to the uvular (postvelar) zone, therefore its "return" to the velar stop *g- became impossible. Even the word-initial strengthening in Ossetic resulted only in its transition in one of the dialects, into the unvoiced uvular q- - the only stop in this phonemic group (the other dialect retains γ-).

The tendency toward a spirant character of word-initial voiced consonants in East-Iranian languages may be rooted in the ancient past It is known that Proto-Iranian *b, *d, *g are reflexes of the two Proto-Aryan consonant series merged: aspirated *bh, *dh, *gh and non-aspirated *b, *d, *g (corresponding to series I and II of the Indo-European model suggested by T.V. Gamkrelidze and V.V. Ivanov). During Proto-Iranian, as well as during Proto-Aryan and Indo-European periods, there was no fricative/stop opposition for voiced consonants. Even in late Indo-European "dialects", and later periods - up to individual Indo-Aryan languages - voiced aspirated consonants may be phonetically realized as voiced fricatives /9/ and/or as freely varying sounds of *bh/v type, etc. A similar articulation type may be assumed also for those early Aryan dialects from which Iranian languages later originated. This clarifies the further development of articulation of voiced consonants in early Iranian dialects:

two phonological series were merging differently in different dialectal groups. In Western dialects, owing to a strong occlusive component and weak aspirated one, the consonants of the aspirated series quickly lost their aspiration and merged with non-aspirated consonants, becoming similar to the latter (*b, *d, *g). Eastern dialects which had a longer contact with Indo-Aryan languages, retained for a longer time the aspirated component in the articulation of voiced consonants (and/or their spirant articulation). As a result, when the two series merged into one, the articulation characteristic to the other series prevailed, i.e. that of the series of aspirated (and/or spirant) consonants.

Thus, the phonetic - articulatory tendency toward a spirant realization of voiced aspirated consonants, which could have operated already in dialects of Late Indo-European, continued to operate also in dialectal zones of Proto-Aryan, remaining for a long time at the phonetic level. Word-initial voiced spirants became phonemic much later - after the two Proto-Aryan series of voiced consonants had merged into one series in Proto-Iranian and after the opposition "stop/fricative" in the subclass of voiced phonemes become - much later - phonologically relevant, in the period when Iranian languages were already divided into the main groups. Different factors - internal and external - contributed to the phonological independence of these oppositions, which progressed unevenly in different language subgroups and areas, and even - within one language system - in different articulatory zones.

Thus, attention to the articulatory phonetics of living Indo-Iranian languages provides a typological background for verification and correction of reconstructed phonetic and phonological systems and their change patterns, whereas attention to the articulation aspect of historical changes sheds light on the character of the main tendency in these changes, possible deviations from it, and relative chronology of a number of processes. This helps to narrow the gap between the comparative-historical postulates and the studies of living languages.

The above-described procedures may also help to solve some specific disputable problems in the history of Indo-Iranian languages. They help to reveal the lack of uniformity of Proto-Iranian in different regions (even with respect to different phonetic presentations of phonemes forming a single phonemic series), and possible areas of substratum and adstratum influences on Indo-Iranian languages. These procedures help to differentiate internal and external factors in the development of this language family, starting from an earliest period.

References

/1/ R.Jakobson, Typological Studies and Their Contribution to Historical Comparative Linguistics, - "Reports for the Eighth International Congress of Linguists (Oslo, 5-9 August 1957). Supplement", Oslo, 1957, p. 1-11.

/2/ В.С.Соколова, "Очерки по фонетике иранских языков", М.-Л., 1953, ч. II, с. I33-I35; Д.И.Эдельман, "Язгулямский язык", М., I966, с. I4-I8

/3/ А.Л.Грюнберг, "Языки Восточного Гиндукуша. Язык кати", М., I980, с. I70-I72.

/4/ Б.А.Захарьин, Д.И.Эдельман, "Язык кашмири", М., I97I, с..34-4I.

/5/ Т.В.Гамкрелидзе, Вяч.Вс.Иванов, "Индоевропейский язык и индоевропейцы. Реконструкция и историко-типологический анализ праязыка и протокультуры", Тбилиси, I984, т. I, с. 85-I5I, I70-I72, 2I4.

/6/ Д.И.Эдельман, К типологии индоевропейских гуттуральных, - "Известия АН СССР, Серия литературы и языка", т. 32, вып. 6, М., I973.

/7/ Г.А.Климов, Д.И.Эдельман, К формированию фонологического ряда шипящих, - "Phonologica 1976 (Akten der dritten internationalen Phonologie-Tagung, Wien, 1-4. Sept. 1976)", Innsbruck, 1977.

/8/ "Принципы описания языков мира", М., I976, с. 260-26I.

/9/ J.Knobloch, Concetto storico di protolingua e possibilita e limiti di applicazione ad esso dei principi strutturalistici, - "Le «Protolingue»: Atti del IV Convegno Internazionale di linguisti", Milano, 1965, p. 153; Т.Я.Елизаренкова, "Исследования по диахронической фонологии индоарийских языков", М., I974, с. I8I.

terms of description, the senior terms being semantico-phonological classes. Classes of lexical etymologies and grammatical classes comprise the temporal parameter - they are spread out ("smea - red") in time, the functional load of distinctors in the system being measured in their actual or potential transformations. Distinctors can combine into hypermerismata - or they can discompose into hypomerismata. This idea was expressed for the first time in the Tbilissi Phonological School. Synthesis (resp. discomposition) of features presupposes absorption (resp. emanation) of information and functional load /5/. The functional load freed from the segmental (phonemic) level, is used in phonotactics or prosodic structures. These transformations expose an important property of surface distinctors, which are revealed in neutralizations. They interchange, and their mutations are most clear when the distinctors transgress from one phonological level to another. We explain this phenomenon in the following way: surface distinctors are "incarnations" or "reincarnations" of deep-structure merismata, the latter are less numerous and quite invariable. Group phonemes, syllabophonemes and units of word prosody may be treated as transforms of phonemes, and vice versa.

The possibilities of the model were investigated by G.S. Klychkow and L. Hertzenberg /6,7/. The model comprises variables x, y ..., the slots C and V as parts of the syllable CV, stages segmental Cx or Vx, suprasegmental C^x or V^x , and connected /CV/ vs. disconnected /C-V/. If "x" is laringality, then $/CV/^x$ is a syllable with stød-like accent, /C‿V = x_1/ should be interpreted as H or schwa, C^xV - syllable with aspirated initial, CV^x - syllable with long vowel, CxV as HV, CVx as CVH, where H represents the segmental laryngeal. It is worth noting that one feature in the bynary slot CV can be distributed in the four patterns CV, CxV, CV^x, C^xV^x; the model thus presupposes some restrictions.

a). Only one feature is considered, so CV is excluded because it would mean introduction of new binary feature - "$\bar{x}$",

b). The feature can be used only once; C^xV^x, $H\bar{V}$, $C^h\bar{V}$, C^hVH, HVH are all excluded (Grassmann's law).

3. A more complex model has been developed in order to explain the phonology of the Indoeuropean Protolanguage disintegration. The transformation of a syllabomorphemic language into a family of what one would call word-and-phoneme languages is accepted as the diachronic axis for this model. It comprises two unvariable deep-structure merismata - $\mathfrak{F}$ ("force") and $\mathfrak{L}$ ("laryngality"), which appear as different surface distinctors on three phonological levels: the phonemic level (Φ), the suprasegmental level (Σ) and the word-prosody level (Λ).

" Features" and phonological levels		Deep-structure merismata	
		$\mathfrak{F}$	$\mathfrak{L}$
"surface" distinctors on levels	supra-seg-mental Σ	˥ the high register	˧˥˧ the broken contour
	phonemic Φ	ə - "force" (suffocancy &c)	ʰ - aspiration
	word prosody Λ	distinctors of stress paradigms	distinctors of contact

Then the following consonant changes show how the protolanguage developed into two main filiations

I. D → T°/ ˥   II. D → DH/ ˧˥˧
D → D / ˧   D → D/ ˧

The consonant sets correspond in the following way:

Language branches		Proto-indo-european	Arian	Armeno-germanic	Balto-slavic
Typological formulae		$\Sigma^{\mathfrak{L}\mathfrak{F}}$ (Φ) (Λ)	(Σ) $\Phi^{\mathfrak{L}}$ $\Lambda^{\mathfrak{F}}$	(Σ) $\Phi^{\mathfrak{F}}$ $\Lambda^{\mathfrak{L}}$	$\Sigma^{\mathfrak{L}}$ (Φ) $\Lambda^{\mathfrak{F}}$
Sound "laws"	I	(D) ˥	D	$T^{ə}$	DṼ
	II	(D) ˧˥˧	DH		DV̋
	III	(D) ˩˥		D	DV́

The "experiences" of the merism $\mathfrak{L}$ had been investigated by G.S. Klychkov /5,6/, those of the merism $\mathfrak{F}$ - by V.A. Dybo /8/. L. Herzenberg has revealed the merismata mutations in the prehistory of Anatolian, Greek, Italic and other branches of Indoeuropean /7/. The reconstruction of Indoeuropean phonological diachrony thus requires a theory with two kinds of "features":

a) unvariable "deep-structure" merismata,

b) surface distinctors being "incarnations" and "reincarnations" of the merismata ; their mutual transformations and their transgressions from one phonological level to another are mutations which seem to be determined by typological language change tendencies.

The models introduced above presuppose dynamic realizations of linguistic units with the remaining constant character of their inner regenerational patterns. The relation "dynamic realization - constant structural pattern" is the main feature of all processes determinating the unity of linguistic families.

4. The next step in complicating the model is the introduction of semantic features. Relations between semantic and phonological features are supposed to lead to flexible non-discreet merismatic structures. Linguistic material is represented in the model as large dynamic and semantic sets. Processing of material necessarily becomes computerized. Three classes of Indo-European etymologies: words for "water", natural phenomena and an open class of random etymologies were described in terms of 35 semantic features and 35 phonological distinctors. Correlation coefficients between the classes on semantic features and phonological distinctors were calculated and the agree criterion x^2 was computerized twice. First it was phonological coefficients that were taken as theoretical data with semantical coefficients as semantical data, then vice versa. The main result obtained in the experiment lies in the realization of the fact that semantic and phonological relations in all the classes are orthogonal.

It becomes evident that both in speech production and speech perception the principle of shuttle movement is dominant. The focus of actualization moves incessantly between the vocal and consonant, between phonemic, suprasegmental and word-prosody components, between merismata and "files", between the phonological and semantic spheres.

References

/1/ H.M. Hoenigswald. The principal step in comparative grammar.- Language, Vol. 35, 1959.

/2/ Джаукян Г.Б.К вопросу о происхождении консонантизма армянских диалектов.- ВЯ, I960, № 6.

/3/ V. Pisani. Parenté linguistique.- Lingua, 3, 1, 1959.

/4/ H. Dahlstedt. Phonetic motivation as driving force in the formation and propagation of neologisms.- In : From sound to words. Umea, 1938, p. 27.

/5/ G.S. Klychkow. The theory of phonological features and definition of phoneme.- In: Abstracts of the Tenth International Congress of Phonetic Sciences.- Dordrecht, 1983, p. 685).

/6/ Клычков Г.С. Консонантизм и структура слова.- М., I967.

/7/ Герценберг Л.Г. Вопросы реконструкции индоевропейской просодики.- Л., I98I.

/8/ Дыбо В.А. Славянская акцентология.- М., I98I.

ACOUSTIC CHARACTERISTICS OF THE GLOTTAL STOP IN KAYABI

Helga Elisabeth Weiss

Summer Institute of Linguistics, Brazil

ABSTRACT

Spectrograms of the recorded speech of one male and one female speaker of the Kayabí language of Brazil demonstrate voiced variants of the normally voiceless glottal stop.
The glottal stop is realized by several variants which are reflected in the spectrograms. These variants range from a period of complete closure, through various kinds of creak, to a more slowing down of vocal cord vibrations, or a combination of these.

INTRODUCTION

This presentation describes the Kayabí glottal stop in terms of their acoustic cues and cue patterns as observed in spectrograms. Some degreee of glottal stricture from glottal trill, flap to glottal stop is perceived and identified by the listener as a glottal stop.
Kayabí belongs to the Tupí-Guaraní language family of Brazil as classified by Rodrigues (Rodrigues 1958). The Kayabí language is spoken by about 400 Indians living in Central Brazil, were the material on which this study is based was gathered.

1. METHOD

Data used for this research were mainly lists of isolated words of varying syllable types and lengths taperecorded by one male and one female Kayabí speaker, using an Uher 4000 Report 'S' taperecorder with a Sennheiser microphone.
Spectrograms were produced with a Kay Digital Sona-Graph 7800 at the University of Edinburgh.

2. RESULTS

The glottal stop in Kayabí is phonologically an articulation type, being in contrast with other stops. It occurs in syllable initial position and functions as a consonant. Examples of contrast are:

/taˀɨt/	'offspring'	/tajtɨ/	'cloth'
/aˀu/	'he/she works'	/katu/	'good'
/aˀɨ/	'a sloth'	/aɨ/	'it hurts'
/ɔˀɔ/	'meat, flesh'	/ɔɔ/	'he/she goes'
/ɨˀa/	'a gourd'	/ɨat/	'canoe'

Preglottalized consonants also occur syllable initial at morpheme boundaries. They are mostly the result of metathesis, as seen in:

/ipiraŋ/ 'red' plus suffix /-ˀi/ 'diminutive'
⟶ /ipiraˀŋi/ 'a little red'

2.1 Intervocalic glottal stop

The acoustic cues reflecting the intervocalic glottal stop articulation are seen on spectrograms as:
a. a gap of shorter or longer duration, reflecting a momentary and voiceless articulation:

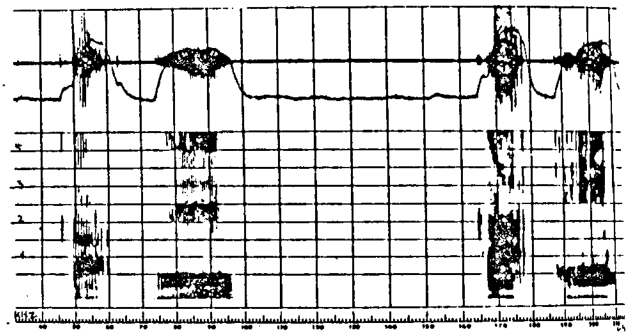


Figure 1: /kaˀi/ 'monkey'
left male speaker, right female speaker

b. a series of short and irregular gaps between stronger glottal pulses, with or without slowing down of vibrations before and after these gaps:

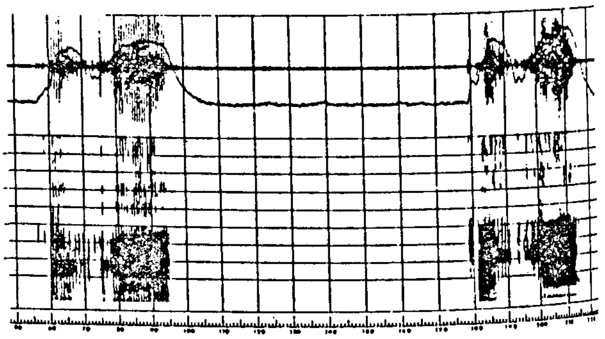


Figure 2: /kaˀa/ 'jungle, tree'
left male speaker, right female speaker

c. a rapid lowering and raising of F_0:

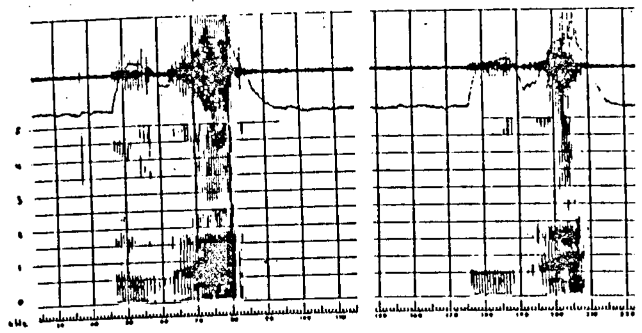


Figure 3: /ɨˀa/ 'gourd'
left male speaker, right female speaker

The various types of glottal activities are accompanied by a drop in intensity, which is maximum for the complete glottal closure.
The duration of the intervocalic glottal stop ranges from 80-160 msec, which is shorter than the duration of other stops.
The auditory effect is always that of a glottal stop, whatever the variation in degree and length of closure. Free variation of the acoustic cues for a glottal stop have been observed in the same utterance spoken on separate occasions by the same speaker.

2.2 Preglottalized consonants

A glottal stop preceding the consonants /m/ /n/ /ŋ/ /w/ /j/ /r/ /g/ can have the same variants as when occurring intervocally:

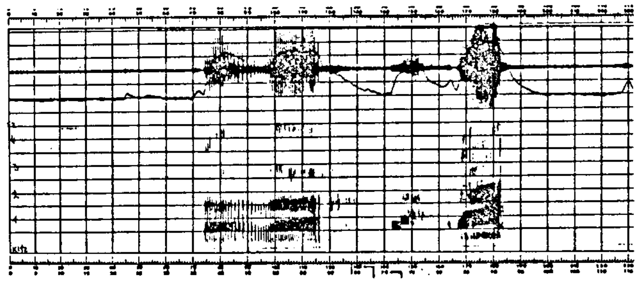


Figure 4: /kaˀra/ 'a root vegetable'
left male speaker, right female speaker

The duration of this glottal activity is about 70-150 msec, followed by a consonant of 20-60 msec. This consonant duration is shorter than the duration of the same consonant intervocalically. The auditory effect of the variants of the glottal stop preceding consonants is always that of a pre-stopped consonant.

3. FACTORS INFLUENTIAL IN THE CHOICE OF VARIANTS OF THE GLOTTAL STOP

Kayabí speech demonstrates a tendency towards a more lax articulation, especially in the speech of male adults. This results in incomplete or lax closures especially of /r/ and /ˀ/, and a more open aproximation for fricatives. Women tend to use a more tense and precise articulation with tighter closures and narrower constrictions, and often a longer duration of segments.
The creak variants are more prevalent in male speaker with a lower F_0. Female speakers do manifest creak, but show a tendency toward complete closure.
The creak variant is more common in open vowels than in close ones.
The glottal closure is longer the heavier the stress, with less closure or just creak with weaker stress. In faster speech and longer utterances the variants creak to slow vibrations are the more common.

4. CONCLUSIONS AND DISCUSSION

In Kayabí the target for the glottal stop articulation is of the category stop or closure, realized by a scale of glottal stricture from complete and prolonged closure, several short closures, through creak to tense voice.
The unit of perception is composed of acoustic cues and cue patterns as seen in the spectrograms, which reflect glottal activity of varying degrees, the totality of which is perceived and distinguished as the phoneme 'glottal stop'.

REFERENCES

Fant,G.(1960). Acoustic Theory of Speech Production. Mouton, The Hague.

Fischer-Jørgensen,E.(1954). Acoustic Analysis of Stop Consonants. Miscellanea Phonetica, **2**:42-49. Reprinted in Lehiste (1967).

Ladefoged,P.(1971). Preliminaries to Linguistic Phonetics. Chicago: University of Chicago Press.

Laver,J.(1980). The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

Rodrigues,Aryón D.(1958). Classification of Tupí-Guaraní. International Journal of American Linguistics, **24**:231-244.

Shoup,J. and Pfeifer,L.(1976). Acoustic Characteristics of Speech Sounds. In Lass,N.(1976) pp. 171-220.

GLOTTALIC STOPS IN GITKSAN: AN ACOUSTIC ANALYSIS

J. INGRAM * AND B. RIGSBY **

* Department of English
** Department of Anthropology & Sociology
University of Queensland
St Lucia 4067
Australia

ABSTRACT

A time series analysis of Gitksan ejectives provides additional evidence for a typological distinction between fortis and lenis glottalic stops.

INTRODUCTION

In the light of recent evidence of inter-language and inter-speaker variation [1, 2], it is apparent that the classical account of the glottalic airstream mechanism for ejectives [3] is in need of revision. This paper reports an acoustic investigation of plain and glottalized stops in Gitksan, a Tsimshianic language spoken in the Skeena River valley of British Columbia.

Glottalized stops in Gitksan are notable for their lenis character [4, 5], in contrast to the unmistakably ejective nature of glottalized stops in other Pacific Northwest Ameridian languages, such as Sahaptin or Kiksht (Upper Chinookan). For non-native listeners, Gitksan glottalized stops may, in certain instances, be perceptually confused with plain voiced stops, with which they are actually in phonemic contrast. Hoard [4] suggested that glottalized stops in Gitksan utilize an implosive airstream mechanism (in prevocalic position) and proposed a revision of the then current Chomsky & Halle [6] scheme for laryngeal features. Gitksan stops, because of their transitional status, provide an interesting testing ground for models of laryngeal features.

Hoard's conclusions were derived not from instrumental, but impressionistic phonetic observations, supported by then-known properties of glottalic consonants [7] and inferences based on the classical model of the glottalic airstream mechanism. We find no evidence to support Hoard's claim that Gitksan glottalized stops are implosive, but this negative result merely raises the question of precisely what the underlying articulatory mechanism may be. The question is significant for a model of glottalic features in general.

More recent instrumental studies of glottalic obstruents [1, 2] have revealed a greater range of cross-language and cross-speaker variation than was hitherto envisioned. In addition to the widely accepted distinction between ejective and ingressive mechanisms within glottalic consonants, it seems necessary to draw an additional typological distinction between fortis and lenis varieties of glottalic consonant.

Lindau [1] compared implosive and ejective glottalic stops acoustically in a number of languages, including Hausa, which has a labial implosive and a velar ejective as part of the same series of glottalic stops. She found greater speaker variation in Hausa glottalic stops than in the other languages examined (Degema, Kalabari, Orika, Bumo, Navajo). Comparison of Navajo and Hausa ejectives indicated substantial differences in manner of production, which we associate with a fortis-lenis typological distinction among glottalic stops.

Kingston [2], developing a theory of tonogensis for Athabaskan languages, distinguished between tense and lax ejectives, claiming that the following phonetic features of ejectives in Tigrinya, a Semitic language of Ethiopia, and Quiche, a Mayan language, exemplify the differences between the two types:

TABLE I
CHARACTERISTICS OF TENSE AND LAX EJECTIVES

Type of ejective:	Tense	Lax
F_0 of the following vowel:	raised	lowered
Voice onset time:	long	short
Intensity of release:	high	low
Vowel onset:	abrupt	gradual

(after Kingston, 1985)

The speakers: Two informants provided two tokens each of a word list elicited by one of the authors (BR) in 1985. One of the Gitksan speakers, LH a male in his late 30's, was the informant for Hoard's 1978 paper. The other Gitksan speaker, SH, is LH's mother. SH is the more conservative of the two speakers with respect to Gitksan norms of usage and also the more fluent. Both speakers are bilingual in Gitksan and English, but LH clearly favours English in his everyday speech.

Two examples from Chipewyan, an Athabaskan language with typical fortis ejectives, reported here for purposes of comparison, come from the speaker in Hogan's [8] study.

The Gitksan items were tape recorded on a Marantz cassette recorder (CP430) and then digitized at a sampling rate of 20 KHz for time domain and spectral analysis using the ILS signal processing package.

Acoustic analysis: Plain and glottalized stops in word initial pretonic position were examined from waveform displays comprising the whole word and a windowed frame of the first 358 msecs (see figure 1). The following acoustic features of the signal were examined by a combination of quantitative measurement and qualitative visual inspection of waveform characteristics:

1. The amplitude envelope of the oral release burst.
2. The voice onset time.
3. The amplitude envelope of the vowel onset.
4. The presence of aperiodicity and period by period fundamental frequency changes in the vowel onset.

These acoustic features are illustrated in Figure 1. Where qualitative judgements based on inspection of waveforms were used, rater reliability was checked by repeating the observations one month after the originals were made. Reliability rates, expressed as percentage agreement scores, ranged between 94% and 98%.

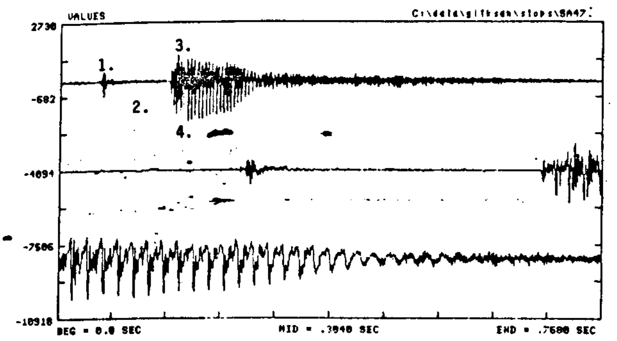


FIG. 1. Waveform of Gitksan ejective

Plosive release characteristics: Fortis and lenis varieties of glottalized stops may be expected to differ in the amplitude of the noise burst associated with the oral release gesture. It would be reasonable to infer that the amplitude of the release burst is monotonically related to intra-oral air pressure and the strength of the compression gesture just prior to release. In the case of fortis and clearly ejective glottalized stops, the release burst is highly damped and followed by a period of silence (approximately 100 msec.) before the onset of voicing. This contrasts with the release characteristics of voiceless aspirated stops, which typically also have a substantial voice onset time, but where the noise burst is relatively undamped and continues up to the onset of the vowel.

The contrasting release characteristics of the ejective and aspirated stops are attributable to two factors: a) higher intra-oral air pressure during the compression phase of the ejective as the larynx is raised, b) the open configuration of the glottis for aspirated stops, which permits sustained turbulent oral airflow up to the vowel onset.

In the case of (English type) plain voiced stops, where oral release occurs more or less simultaneously with voice onset, obviously no independent release burst is observable. In the case of prevoiced (Spanish type) or imploded stops, low amplitude voicing is observable in the waveform prior to oral release.

Measurements were made of the maximum amplitude of the oral release burst where it could be observed independently of the vowel onset. This was possible in all cases for the glottalized stops but generally not for the plain stops, except for the velars and uvulars. Figure 2 shows the observed distribution of release burst amplitude measurements for the Gitksan glottalized stops as well as for two reference tokens from Chipewyan. The amplitude measurements have been expressed as ratios of the maximum vowel amplitude for their respective tokens, so as to normalize the data for arbitrary variations in absolute signal strength.

Figure 2 illustrates the lenis character of the release burst in Gitksan glottalized stops in comparison with those of Chipewyan. It also indicates substantial variation in the relative strength of the release gestures.

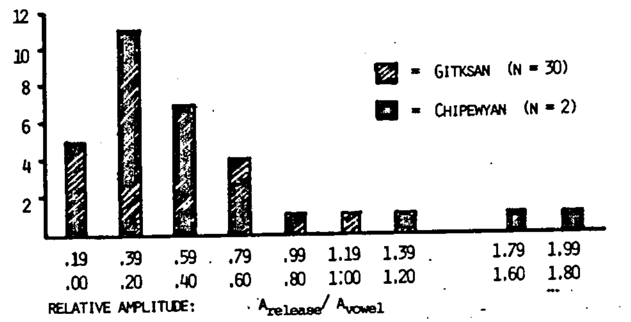


FIG. 2. Amplitude of release burst for Gitksan & Chipewyan ejectives

The envelope of each release burst was also classified by visual inspection of the waveform into one of four categories:

C (checked): a damped noise burst
A (aspirated): an undamped noise burst
V (voiced): noise burst coincides with voice onset
P (prevoiced): voice onset prior to oral release.

With only two exceptions all glottalized stops were judged to have a 'checked' noise burst, although the amplitude was very low in some cases, but still audible. The distribution of release types for the plain stops is shown in Table II.

TABLE II
Classification of plain stops by type of release burst and place of articulation.

	Prevoiced	**Voiced**	**Aspirated**	**Checked**
Alveolar	1	7	0	0
Velar	0	8	4	0
Uvular	0	3	0	3

It is typical of dorsally articulated stops to have some aspiration. The damped appearance of some uvular release bursts should not be taken as indicative of an ejective airstream mechanism, but it may explain why there is a tendency to mishear plain uvular stops as glottalized. Only one instance of a prevoiced stop was observed in the data set (SH, on second elicitation of /taw/, [daw], 'ice'). No examples of implosion were encountered, on phonologically glottalized stops or otherwise.

Voice Onset Time: Hogan [8] reports a mean Voice Onset Time of 114 milliseconds for Chipewyan ejectives in single word utterances elicited under comparable conditions to the present study. Voice onset times for the Gitksan glottalized stops were somewhat shorter with quite a high variance ($\bar{X}$ = 89.2, SD = 31.3 msec). Voice onset times for plain Gitksan stops fell within the range of English voiced stops ($\bar{X}$ = 11.1, SD = 36.1 msec). Figure 3 shows the distribution of voice onset times for all items. There is clear separation of plain and glottalized stops on the VOT continuum for SH but some overlap for LH whose VOT's are generally shorter for the glottalized series.

Amplitude envelope of vowel onset: For ejectives produced with a tense glottal configuration an abrupt vocalic onset may be expected, whereas with a more lenis configuration, the onset may be more gradual. A simple but adequate index of the abruptness of the vowel onset was provided by the peak amplitude of the third glottal pulse as a proportion of the maximum amplitude attained by the vowel. Figure 3 shows the distribution of this index for all tokens on the y axis of the graph. It is clear that the A3/Amax index does not distinguish glottalized from plain stops. It does however indicate a cross speaker difference. SH's vowel onsets are more abrupt than LH's.

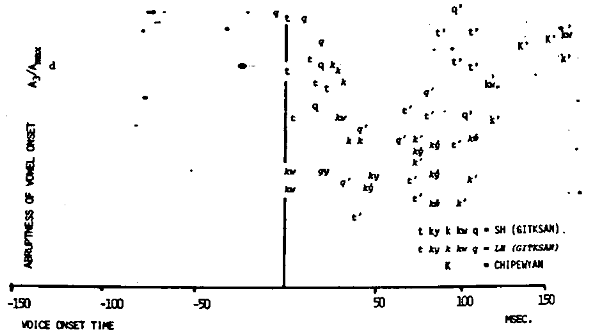


FIG. 3. Voice Onset Time and Abruptness of Vowel Onset in plain and glottalized stops.

Aperiodicity and frequency of vowel onset: Frequency characteristics of the vowel onset carry information about the glottal configuration. Kingston [2] distinguishes between tense and creaky voice onsets which follow tense and lax glottalized stops respectively. Tense voice is associated with stiff vocal folds, a high degree of general laryngeal constriction, and a higher than normal transglottal pressure differential to sustain phonation. Creaky voice, on the other hand, is associated with shortened but lax vocal folds, moderate medial compression of the vocal folds with a lax laryngeal configuration, and a lower than normal phonatory transglottal pressure differential [9].

There is some uncertainty about the acoustic features that distinguish tense and creaky voice. The most prominent difference lies in the fundamental frequency of phonation, which is very low in creaky voice and somewhat raised for tense voice. A greater degree of aperiodicity of vocal fold vibration may be expected for tense and creaky voice than in modal voice, though its time series and spectrographic expression may be different in the two non-modal voice qualities.

In creaky voice there is gross variation in the period of vocal fold vibration, pssibly due to insufficient airflow or subglottal pressure to sustain regular pulsation. In tense voice, the higher degree of stiffness in the vocal folds and surrounding laryngeal musculature, combined with higher levels of subglottal pressure produce a phonatory cycle that has a relatively longer closed phase than in modal voice and a spectrum characterized by increases in the amplitudes of higher harmonics. Frequency or amplitude variation in the higher vocal harmonics may result from inherent instabilities of the laryngeal configuratión for tense voice.

Aperiodicity during the first 40 milliseconds or so of the vowel is evident from the whole waveform display in Figure 1, but this was atypical. Only 26% of the Gitksan glottalized stops in pretonic position showed obvious aperiodicity in the first 40 milliseconds or so of the following vowel. However, there are possibly significant speaker differences on this parameter. For SH, six of ten vowels following glottalized stops showed aperiodic onset, compared with two out of twenty for LH.

Measurements were also made of the periods of the first eight glottal pulses. Figure 4 summarizes the results.

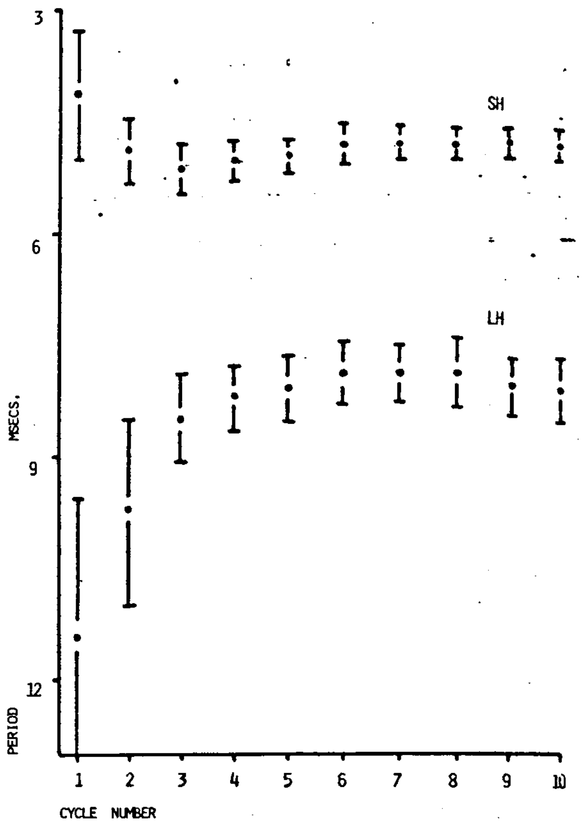


FIG. 4. Period by period frequency changes in vowel onset

There are notable speaker differences in the frequency contours of vowels following glottalized stops. For LH, whose glottalized stops are particularly lenis, the frequency at onset is low, well within the range of laryngealized voice, and rises nonlinearly to modal voice within the first five or six glottal cycles. SH's vocal frequency generally begins slightly high and drops to normal value within the first three cycles.

LH's vowel onset frequency contour follows that of lenis glottalized stops as observed for Hausa by Lindau [1] and Quiche by Kingston [2]. SH's frequency contour more closely approximates that of fortis Chipewyan stops (this paper) or Tigrinya [2]. A common feature of both speakers vowel onsets is the large and diminishing variance in the periods of the first two or three glottal pulses.

Summary and conclusions: To summarize, Gitksan glottalized stops in pretonic position are characterized acoustically by:

1) A relatively weak but damped release burst, consistent with a lenis ejective airstream mechanism.
2) A shorter VOT than is typically observed for (fortis) ejective stops.
3) A gradual rather than abrupt vowel onset in the majority of tokens, though this feature varied with the speaker and was correlated with:
4) An absence of visible aperiodity in the waveform of the following vowel-onset for the majority of tokens (contra illustration in Figure 1 above), but nevertheless:
5) Substantial, but declining, pitch period perturbation (jitter) over the first few glottal cycles of the vowel onset, with the f_0 contour rising (in the case of speaker LH) or steady-falling (SH).
6) Significant speaker variation on all of the above features, with LH consistently demonstrating a more lenis pattern of articulation.

It is possible that the observed speaker differences are attributable, not to inherent variation in Gitksan glottalized stops, but to the differential effects of language shift. As mentioned earlier, LH is less fluent than his mother, who maintains full native-speaker productive control over the language. Alternatively, the speaker differences may be at least in part attributable to stylistic variation. SH is a more conservative speaker, and her pronunciation may reflect the use of a more formal speech style. Regardless of the source of these speaker differences, it is possible to draw certain inferences about the underlying articulatory mechanism of Gitksan glottalized stops and to hazard some speculative comments as to their featural representation.

Analysis of the Gitksan data, taken in context of a growing body of data from other languages, suggests that a language typological distinction between fortis and lenis ejectives is warranted, where this term is understood in the traditional sense of the degree of vigour of the complex laryngeal and articulatory components which comprise the whole gesture [6, 10, 11]. Reduced upward movement of the larynx would produce a weaker and shorter compression phase. This is consistent with the lower observed amplitude release burst and shorter VOT's of Gitksan glottalized stops. A weaker medial compression, with lower overall muscular tension in the larynx will result in a non-abrupt vowel onset, but one which is more likely to begin with a creaky or laryngealized mode of vocal fold vibration. (While fortis ejectives have a glottal attack, lenis ones, begin with a laryngealized voice and hence both have a 'glottalized' voice quality.)

Of the variously competing laryngeal feature systems, that of Ladefoged [9] seems most naturally to accommodate the emerging picture of cross language variation in glottalic stops. The fortis-lenis contrast applied to ejectives is a **typological** rather than a **distinctive** phonetic feature. In Languages such as Hausa whose glottalic stops fall on the lenis end of the continuum, the contrast between ejective and implosive airstream mechanisms is less apparent, both perceptually and physiologically. Reduction of the laryngeal constriction and movement in the vertical plane will tend to result in half-way state that yields a brief period of laryngealized vocalization superimposed upon a weak plosive gesture that may be ejective in some environments and ingressive in others.

No evidence was found of an implosive airstream mechanism in Gitksan glottalic stops, but rather, a weakly ejective mechanism seems to be used. However, Hoard's important insight that a feature 'glottalized' is required, in order to adequately capture phonetic processes to which lenis glottalic stops may be particularly prone, still stands.

REFERENCES

[1] M. Lindau, Phonetic differences in glottalic consonants. Journal of Phonetics (1984) 12, 147-155.
[2] J. Kingston, The phonetics and phonology of Athabaskan tonogenesis. M.S. University of Texas, December 1985.
[3] J.C. Catford, Fundamental problems in Phonetics. Edinburgh University Press, 1977.
[4] J.E. Hoard, Obstruent voicing in Gitksan: some implications for distinctive feature theory. In E.D. Cook and J. Kaye (eds) Linguistic studies of native Canada. University of British Columbia Press, 1978.
[5] B. Rigsby, Gitksan Grammar. M.S. British Columbia Provincial Museum, 1986.
[6] N. Chomsky & M. Halle, The sound pattern of English. Harper & Row, 1968.
[7] J.H. Greenberg, Some generalizations concerning glottalic consonants, particularly implosives. International Journal of American Linguistics (1970) 36, 123-145.
[8] J.T. Hogan, An analysis of temporal features of ejective consonants. Phoneticia (1976) 33, 275-284.
[9] P. Ladefoged, The features of the larynx. Journal of Phonetics (1973) 1, 73-83.
[10] C.W. Kim, On the autonomy of the tensity feature in stop classification. Word (1965) 21, 339-359.
[11] K.J. Kohler, Phonetic explanation in phonology: the feature fortis/lenis. Phonetica (1984) 41, 150-174.

AFFRICATIZATION OF /t'/, /d'/ IN THE MODERN LITERARY PRONUNCIATION OF MOSCOVITES

S.B. VORONINA

Dept. of Philology
Institute of Information on Social Sciences
Academy of Sciences of the USSR
Moscow, USSR 113095

ABSTRACT

The research aims at describing the latest development of a new process in the articulation of moscovites - affricatization of /t'/, /d'/. The instrumental analysis shows that /t'/ and /d'/ develop fricative stages, as a result the occlusive /t'/, /d'/ may be replaced by the affricates /t͡s'/ and /d͡z'/.

The latest decades have shown a considerable development of a peculiar phonetic effect - affricatization of /t'/, /d'/. In terms of articulatory phonetics the effect can be described as follows. The occlusive stages of palatalised /t'/, /d'/ are followed by rather long perceivable whistling fricative stages.

The main aim of this study is to look into the problem of the quality of the palatalised /t'/ and /d'/, including such aspects as sex and age of informants. 87 native speakers took part in the phonetic experiment, conducted by the author of this paper. The informants were chosen at random according to the following criteria: 1. he or she must be moscovite by birth; 2. he or she must have no flans in articulation. The informants read a specially-prepared list of words /1/. Their reading was recorded on tape. The tape was subjected to an auditory analysis with the aim to establish the extent to which the phenomenon of affricatisation of /t'/. /d'/ is used by speakers oi different sex and age. Nine informants formed the so called "main group", These were permanent residents of Moscow, brought up in families, two generations of which had also lived in Moscow. The pronunciation of the main group was subjected to a more profound analysis with the use of an oscillographic, palatographic, linguographic and spectrographic methods of analysis, as well as auditory analysis (with the help of a segmentator of speech sounds with a rotating magnet head) and auditory analysis, performed by a group of linguists and non-linguists.

All the informants were divided into four groups according to their age: 1. schoolchildren (from 11 to 16 years); 2. students and post-graduates (from 20 to 30 years); 3. middle-aged people (40-50 years old); 4. the older generation (60-80 years old). The third and fourth groups included people with a higher education. Both the auditory and the instrumental analyses have shown, that out of the 75 informants of the first three groups 48 people (64% of all) had affricatisation. But this phenomenon was not typical of the older generation. The speech of the informants of this group represented a very weak fricative element after the occlusive stage of /t'/, /d'/. This element could be detected only by a trained phonetic ear. The fact that different age-groups of the speakers have different

degrees of affricatisation of /t'/,/d'/ accounts for the existing diversity of opinions on the problem.

The comparative analysis of the results of auditory and oscillographic experiments aimed at establishing degrees of affricatisation points to three degrees of this phenomenon. The average affricatisation (2) of /t'/ at 45% and the average affricatisation of /d'/ at 35% accounts for a weak degree of affricatisation. The average affricatisation of /t'/ at not more than 55% and the average affricatisation of /d'/ at 45% accounts for a moderate degree of affricatisation. The average affricatisation at 55% and mode and /d'/ - at 50% and more accounts for a very high degree of the phenomenon.

The auditory analysis performed by Russian phoneticians showed that the average affricatisation of /t'/, /d'/, accordingly 35% and 30% , must not be regarded as affricatisation phenomenon. It has already been proved in earlier experiments conducted by L.V.Bondarko, L.R.Zinder, L.A.Verbitskaya /3/ that a weak quality of affricatisation is an inherent part of palatalised /t'/ and /d'/. As a matter of fact this phenomenon has always existed in the literary pronunciation of moscovites, and now we can speak only about it modification. The quality of the palatalised /t'/, /d'/ is not stable. Because of this the degree of affricatisation of /t'/ or /d'/ in certain position, pronounced by different speakers, always will be different. On the other hand, the idiolect shows more stability. That is why this phenomenon can serve as a reliable means of identification of a speaker.The present phenomenon has some characteristic features. It must be pointed out that the degree of the affricatisation of /t'/ in all positions is greater than that of /d'/. The same fact is true for the Belorussian language and Russian dialects where affricatisation is a dialectal feature. This tendency is universal and the reasons for it lie in the acoustic and articulatory peculiarities of /t'/. Probably, a greater degree of articulatory tenseness and of the volume of air in /t'/, in comparison with /d'/, shows in the lengthening and intensification of the fricative stage of /t'/. Analysing the frequency of this phenomenon, it must be pointed out that the frequency of occurence of obvious affricatisation in the speech of men and women is different. In the speech of women of the second and third age-groups there is a tendency to use the moderate and the strong degrees of affricatisation of /t'/, /d'/, while in the pronunciation of men the moderate and the weak degrees of affricatisation are mostly used. Men's speech is characterised by a much shorter absolute and relative length of the fricative stage of the sound in comparison with the women's one. Thus, the average affricatisation of /d'/ in men's speech may be at 21-34% , for women's speech it's not peculiar. Another discrepancy characterising the speech of men and women is as follows. In women's speech the degree of affricatisation of /t'/ and /d'/ is balanced. In men's speech the balance is tipped in favour of /d'/, i.e. it is very low for /d'/.

Another feature is the intensity of the fricative stage, which in men's pronunciation is fairly low. Due to these features it is possible to identify the sex of the speaker.

All the above-mentioned experimental results point out the necessity of distinguishing male and female manners of modern Russian literary pronunciation /4/.

A very high degree of affricatisation of /t'/, /d'/ in women's pronunciation must be taken into consideration in speech synthesis, automatic recognition of speech and in teaching Russian to foreigners. To prove this it is worth describing some facts from our teaching experience. The students from Palestine and Equador, who were studying Russian here in Moscow, believed that they should write /t͡s/ in the suffixes of the infinitive. They made the following conclusions due to the impresions of their perception. These examples show that in the speech of modern women-moscovites the border between palatalised occlusive /t'/ and affricate /t͡s/ is not steady, this may lead to their perceptual mixing up.

To clear up the mode of production of the sounds /t'/, /d'/, pronounced with a high degree of affricatisation, the author has carried out an auditory analysis, in the course of which Ukrainians and Belorussians, whose native languages have affricates /t͡s'/, /d͡z'/, defined the quality of the Russian /t'/, /d'/. Then the results of the auditory and acoustic /5/ analyses were compared.

The results of the analyses have shown that in the case of high degree of affricatisation the majority of the /t'/, /d'/ sounds were identified as /t͡s'/ and /d͡z'/; in the case of moderate degree of affricatisation approximately 1/3 of the sounds was evaluated as affricates, the rest of the sounds were identified as palatalised occlusive consonants with an affricatisation of a different degree. In the case of the weak degree of affricatisation, only few sounds were identified as affricates. This phenomenon confirms the absence of a steady relationship in the system of the language and gives ground for its further development.

A comparative analysis of spectrogramms and palatogramms of the Russian, Ukrainian and Belorussian /t͡s'/, /dz'/ has shown that the affricates which have developed in the Russian language have got articulatory characteristics of their own and are not identical with the analogous sounds of the Ukrainian and Belorussian languages. In comparison with Belorussian and Ukrainian sounds the Russian sounds are more front. The Belorussian /t͡s'/ and /d͡z'/ occupy an intermediate position between the Russian and the Ukrainian sounds.

As regards the opposition of hardness/palatalisation it must be noted, that some new Russian affricates were identified by the Ukrainians and Belorussians as hard or not palatalised enough. An analysis of the phonetic context of these sounds showed that their new quality was conditioned by the context.

The phonetic context plays an important role in determining the degree of the affricatisation of /t'/, /d'/. It was revealed that the degree of the affricatisation is increased in the intervocal stressed position (the stressed vowel either precedes or follows /t'/,/d'/). This increase becomes possible owing to the phenomenon of spirantisation always taking place in this position. The strong degree of affricatisation is also typical for the word-final /t'/ (here the affricatisation of /t'/ is supported by the aspiration and, besides, the off-glide of /t'/ is not restricted by the following onset of the next sound). The degree of the affricatisation is also increased in the following context: before front-high vowels or /j/, in stressed position in comparison with unstressed one, in logically stressed positions, in front of /э̂/ in a closed (not open)syllable (compare: от/д͡з'э̂л'/ный and раз/д'э́л/). In the speech of the oldest group of moscovites the degree of the affricatisation may be sometimes increased in the position before /j/ and also before /v'/, /m'/ (if the palatalisation of /t /, /d / caused by assimilation). For example, /д͡з'в'э/, /т͡с'в'э̂р'/, /д͡з'м'и́/трий and so on.

Thus, the degree of the affricatisation can be put down to the influence of the

phonetic context. But the question arises what has triggered this process in the phonological system of the Russian language while the system itself has not been and is not being influenced by any other phonological system of a different language? If we try interpret this in terms of phonology only, then it must be said that the development of this phenomenon (i.e. the ongoing transformation of occlusives into affricates) is made possible by the state of phonological "permissibility". From the point of view of phonetics it must be mentioned that in comparison with the articulation of /t'/, /d'/ by old moscovites, the younger generations articulate these sounds in a more front part of the mouth. The advanced position of the whole body of the tongue, accompanied by the dorsal articulatory position, which characterises all affricated /t'/, /d'/, makes the obstruction of these consonants weak enough. Thus it breaks and a long intensive fricative stage develops. That is why high, front vowels (/i/ in the first turn) enhance this process because they do not prevent the movement of the bulk of the tongue forward. These sounds are made so front, that they loose partly the acoustic effect of palatalisation, which they would have being in the intermediate position, and that is why they may be identified as almost hard sounds.

The development of the affricatisation of /t'/, /d'/ influences the whole articulatory base of the Russian language and causes changes in the homorganic with /t'/, /d'/ sounds. The speaker, whose idiolect is characterised by affricatisation of /t'/, /d'/, increases the intensity and length of /s /, /s'/, /z/,/z'/ ; lengthen the off-glide of /t͡s/ (as a result tne listener may identify /t͡s/ as a biphonemic entity - /ts/). It becomes possible to replace the old affricate /t͡s/ by the new /t͡s'/ (such phenomenon was fixed in the words "дьяконица́", "вкра́тце", "меццо").

An all-round study of the process of affricatisation in Moscow may help in the solution of some problems of historical phonetics and account for similar processes in Polish and Belorussian. It may, in a way, throw some light on the development of the language system.

/1/ .Affricatisation of /t'/,/d'/was analysed in the following positions ; 1 .word-final position (for example, мать);
2. stressed position (тётя, сде́лать);
3. unstressed position (тяжёлый, деревя́нный); 4. stressed intervocal position (утя́та, тётя, иди́, Ва́дя); 5. unstressed intervocal position (сетево́й, ледяно́й).

/2/. The term "average affricatisation" means that the length of the fricative stage in the sounds /t'/ and /d'/ is expressed in per cent. The data were obtained from the oscillographic analysis as a ratio of the average relative duration of the fricative stage of /t'/ or /d'/ and the whole sound /t'/ or /d'/.

/3/. L.R.Zinder, L.V.Bondarko, L.A.Verbitskaya. Akusticheskaya kharacteristika razlichiya tvyordykh i myagkikh soglasnykh v russkom yazyke. - Uchyonyye zapiski LGY imeni A.A.Zhdanova. - L., 1964. - Vyp.69, N 325. Seriya philologicheskikh nauk. - P.28-36.

/4/. In linguistic literature concerning Russian dialects one can find indications of a similar kind. Thus, early in the XX century N.M.Karinskiy revealed the same phenomenon in the dialects of Bronnitskiy district. Nowadays a similar fact was described by P.A.Rastorguev in the dialects of the Smolenskaya district (N.M.Karinskiy. O gcvorakh vostochnoy poloviny Bronnitskogo uezda.-Spb., 1903; P.A.Rastorguev. Govoru na territorii Smolenshchiny. - Moscow, 1960).

/5/. In this case duration and intensity of the fricative stages are meant.

PHARYNGEAL FEATURES IN THE DAGHESTAN LANGUAGES

S.V. KODZASOV

Department of General, Comparative and Applied Linguistics,
Philological Faculty, Moscow University
Moscow, USSR, 119899

ABSTRACT

The phonetic data of the Daghestan languages give evidence of the inadequacy of the present views on articulatory possibilities of the pharynx. Apart from the articulations produced in the pharynx by the movement of the tongue body (the uvulars X, R, q) and of the epiglottis (the epiglottals ħ, ʕ, ʢ) there exist articulations produced by sphincteric narrowing of the pharynx itself. Such is the mechanism of producing a secondary feature called "pharyngealization" and of pharyngeals proper X̣, Ṛ. Structural regularities concerning pharyngêal^features in the Daghestan languages are examined in this paper.

PHARYNGEAL ARTICULATIONS IN DAGHESTAN

The Daghestan languages are highly active in using pharyngeal articulations and their data prove the incompleteness of the traditional nomenclature for this region of the vocal tract |1|, |2|. Three series of consonants are found in the Daghestan languages: the uvulars, epiglottals and pharyngeals proper (s. Figure 1). In addition, they possess two secondary features: epiglottalization and pharyngealization.

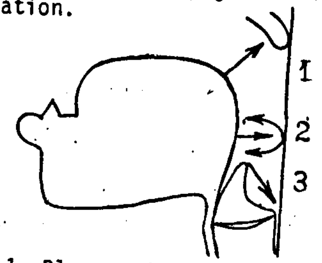


Figure 1. Places of articulation in pharynx:
1 - uvular, 2 - pharyngeal, 3 - epiglottal.

The term "uvulars" indicates the localization of a stricture in the upper-pharyngeal region, but not an active articular. The articulations of q-, X- and R- like sounds are produced by the backward - upward movement of the tongue-body. The uvula does not play an active role - it is either pressed to the posterior pharyngeal wall (plosives) |3|, or may optionally vibrate (spirants). The Daghestan uvulars are similar to the corresponding consonants of Arabic and of some other languages |4|. There are slight differences among languages in the backness of the localization of the stricture but the basic mechanism of its production remains the same.
The Daghestan ħ- and ʕ- like consonants do not differ from the corresponding Arabic sounds auditorily. Our fiberoptic data |5| show that the Daghestan ħ and ʕ are produced in a manner identical with that found by Laufer and Condax |6| for the ħ and ʕ of Hebrew: we observed the backward-downward movement of upper edge of the epiglottis to the posterior pharyngeal wall. Laufer and Condax explained the displacement of the epiglottis by the contraction of the aryepiglotticus and thyroepiglotticus. At the same time, El-Halees |7| considers that the movement of the epiglottis is a mechanical consequence of the larynx rising (this articulatory parameter for ħ, ʕ was assumed already by Troubetzkoy |8|). In any case, the term "epiglottals" seems the most appropriate for these sounds: it is used already by Soviet Arabists |9| and is conceptually close to Troubetzkoy's term 'emphatic laryngeals'. It should be mentioned that, contrary to the widespread opinion, the tongue does not participate in the production of ħ and ʕ: these sounds may be just as well pronounced with the tongue put out of the mouth.
There is a number of allophonic variants of the epiglottal phonemes in the Daghestan languages, these allophones form two series which differ in the extent of the epiglottis displacement from the neutral position (epiglottalization)(see Table 1).

Table I. Epiglottal consonants

1. Moderate epiglottalization	ʡ	ʿ	ħ/ɦ
2. Strong epiglottalization	ʢ	ʕ	H

The sounds of the first series are formed by the super-imposition of moderate epiglottalization on glottal postures for the "plain" laryngeals: ʡ - epiglottalized glottal stop, ħ/ɦ - epiglottalized aspiration (unvoiced/voiced), ʿ - epiglottal approximant (the vocal cords are in the position of "spontaneous voicing").
The basic articulatory component for the sounds of the second series is strong tilting of the epiglottis, the function of glottal articulations being identical with the function of phonation for oral consonants: ʢ - epiglottal stop (the closure between the epiglottis and the pharyngeal wall or the apex of the arytenoids, accompanied with the glottal closure), ʕ- voiced spirant, H - unvoiced spirant.
In the majority of the Daghestan languages we find two epiglottal phonemes, in "broad" transcription we use for them the signs ħ and ʕ accepted by the IPA. Here are examples of their allophonic realization: /ħ/=[H], /ʕ/=[ʕ] (Awar); /ħ/=[H],/ʕ/=|ʿ|/|ʡ| (Archi); /ħ/=[ħ]/[ɦ], /ʕ/=[ʡ] (Dargi); /ħ/=|ħ|, /ʕ/=[ɦ] (Lezgi). We have found similar variants in our pilot-study of the epiglottals ح and ع in Arabic dialects (cf. |10|). But there exist languages possessing three epiglottal phonemes (some Agul dialects and Budukh), in these languages the third member is the epiglottal stop /ʢ/.
Epiglottalization may function not only as the basic component of the epiglottal phonemes but also as a secondary (usually prosodic) feature. Troubetzkoy |8| described this feature as "emphatic palatalization". This designation is connected with the fact that the epiglottis displacement is usually accompanied by the larynx raising and the forward movement of the whole tongue. But the main parameter of the feature (about the relation of the notions "feature" and "parameter" see |11|, |12|) is the narrowing of the lower-pharyngeal passage.
A number of the Daghestan languages know another secondary feature resembling epiglottalization but easily distinguishable from it perceptually. This feature is usually called "pharyngealization" by the scholars studying these languages. Our preliminary investigation of articulatory correlates of this feature by means of fiber-optic technique has yielded the following results. We observed the inward movement of the posterior and lateral pharyngeal walls accompanied by the backward movement of the tongue root together with the epiglottis which led to the narrowing of the pharyngeal passage (see Figure 2). This articulation is probably caused by the contraction of the pharyngeal sphincters (the middle and inferior constrictors).

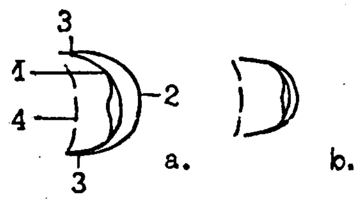


Figure 2. Arbiculatory mechanisms of pharyngealisation
a. Non-pharyngealized [a], b. Pharyngealized [a|]. 1 - epiglottis, 2 - posterior pharyngeal wall, 3 - lateral pharyngeal walls, 4 - tongue root.

The radiographic unvestigations of pharyngealization |3, 13| have revealed only a backward movement of the tongue, yet we believe that the main articulatory parameter is the circular narrowing of the pharyngeal tube invisible on radiographic tracings. For this feature the term "pharyngealization" seems to be appropriate: it indicates that the pharynx proper is in this case an active articulator.
Pharyngealization is usually a prosody, but in some languages it should apparently be considered as a consnantal feature (see below). There exist also the pharyngeal spirants X̣, Ṛ, they relate to the epiglottals ħ, ʕ in the same way as pharyngealisation relates to epiglottalization.
The backward movement of the tongue body seems to be the invariable component of Daghestan pharyngealization but the precise direction of this movement (back or down-back) and the configuration of the tongue blade are unknown. Cross-linguistic differences in timbre colouring of pharyngealization are indicative for the nonidentity of the parameters mentioned. In the majority of the languages (Tabasaran, Tsakhur, etc.) pharyngealization has the palatal timbre, but it is not the usual palatalization. It seems to be caused by a sort of deformation of the tongue: the bulk of the tongue moves back and down, whereas the tongue blade moves back and up (to the palatum). However, there are languages that combine pharyngealization with velarization (Archi |14|, Udi). This kind of pharyngealization resembles auditorily retroflexivization (particularly, ɹ- coloured sounds of American English).
Epiglottalization and pharyngealization do not contrast, they are supplementarily distributed among languages (or dialects) or sometimes among different consonants of the same language. The common origin of both features is apparent, the initial form being epiglottalization. Futher on for the this family of features a cover-term "pharyngeal stricture"(PS) will be used.
In some Daghestan languages (Lack, Dargi) both epiglottalization and pharyngealization participate in the production of the pharyngeal stricture equally. For this variety of PS the term "epiglottopharyngealization" is used here. Epiglottopharyngealization shares with eppiglottalization the palatalizing influence on consonants (especially, on velars) as well as on back vowels. The widening influence of the epiglottal component on narrow vowels is also typical: [i] approximates to [e] and [u] approximates to [ö]. The PS features also contain a slight nasal component.
We mark the PS features by a vertical bar (a|, q|), yet in the "narrow" transcription epiglottalization is marked by a crossed vertical bar (a┼, q┼).

LINGUISTIC FUNCTIONING OF PHARYNGEAL FEATURES

1. All twenty six languages of Daghestan have uvulars as a part of their phonological systems and all basic phonemic contrasts are found in the consonants of this series: all phonation types (voiced/unvoiced/aspirated/ejective), strength, labialization, palatalization.
The epiglottals ħ and ʕ are also found in the majority of the Daghestan languages, in some languages they have the labialised pairs. There is a close connection between the type of PS and the history of the epiglottals in a language given. The old epiglottals are lost if PS has a form of "pure" pharyngealization - they are replaced with the corresponding plain laryngeals, the epiglottal component being reflected by pharyngealization (Archi) or by umlaut of the adjacent vowels (Tsakhur, Tabasaran). The new epiglottals in loan-words may remain (Archi) or be again deepiglottalized. For instance, in Tsakhur ħ → h| and ʕ → ʔ|: h|ɨ|ǯ|a|t *dispute*, ʔ|a|r|a|b|a| *araba*.
In the languages which possess the mixed (epiglottopharyngeal) form of PS (Lack, Rutul) [ħ] and [ʡ] represent /h/ and /ʔ/ in the PS context. In the south dialects of Dargi such sounds may be found in the words both with and without PS. Spirant allophones of ħ and ʕ are typical for the languages which have epiglottalization and the languages without the PS features.
Whereas the old epiglottals are usually (but not always) lost in the languages which have pharyngealization, additional epiglottal spirants arise in

in the languages which have epiglottalization: X| → ħ, R| → ʕ (some Tsez and Agul dialects). The transition q'| → ʕ| is much more seldom (some Agul dialects).

Finally, there are Agul dialects(the dialects of this language demonstrate surprising diversity in the pharyngeal features development) in which the pharyngealized uvulars have changed into the pharyngeales proper: X| → X̂, R| → R̂. There are three series of the post-velar spirants in these dialects: the uvulars (X,R), epiglottals (ħ, ʕ) and pharyngeals (X̂, R̂). Here are some examples from the Richa dialect: X̂al *house* - ħač *apple* - X̂aw *udder*, Rad *hammer* - ʕak$_w$ *light* - R̂an *belly*. In addition, there are new pharyngealized uvulars in this dialect: X|a|w *rut*, R|a|b *stack*. It may be the most abundant system of postvelar spirants in the world languages.

2. Now let us consider linguistic behaviour of the PS features. In the majority of the Daghestan languages they reveal apparent prosodic properties: they tend to spread about the whole word. Here are examples from Archi: b|a|k'|o|n *rope*, k'|e|h|u| *hawk*; an example from North Tabasaran: g|a|r|a|w|R|u|nu|za| *growled*. As a rule the degree of PS diminish from the beginning to the end of the word. Consonants of the different local series differ as to their ability to join the PS features and to pass them onto next segments. The uvulars and laryngeals are the best to coarticulate with PS. They show a tendency to have the highest degree of this feature and may disturb the typical descending pattern. Some languages (e.c. Rutul) have PS only in the syllables which contain uvulars or laryngeals. This feature is naturally treated there as consonantal.

On the contrary, the dentals do not coarticulate with PS and may prevent the spreading of this feature from the preceding to the next vowel: x̄|u|m|u|s̄a *liquid* (Lack). As to the labials, hushing sibilants, laterals and velars, they are "transparent": they easily join PS and pass it to the next segments.

All the forms of PS can easily combine with the labialized consonants. Here are examples from Tsez: R$_w$|a|j *dog*, q$_w$|a|λ̌i *ankle*; examples from Archi: R$_w$|a|lq|i| *smog*, s$_w$|a|s| *last year*.

3. A concluding remark: Arabic "pharyngealization" is not identical with any form of PS and, probably, should be treated as a variety of velarization (cf. Troubetzkoy's term "emphatic velarization" |8|). The "emphatic" ṭ, ḍ, ṣ, ẓ, ḷ of Arabic are pronounced with the tongue displaced into the pharyngeal cavity |10|. However, the perceptual results of this movement are not identical to epiglottalization or pharyngealization of the Daghestan languages. It is a significant fact that just the same sounds (dentals) which participate in the "emphaticalness" contrast in Arabic do not coarticulate with the PS features in the Daghestan languages.

REFERENCES

|1| Ladefoged P. Preliminaries to linguistic phonetics. Chicago and London, 1971.

|2| Catford I.C. Fundamental problems in phonetics. Indiana University Press, Bloomington and London, 1977.

|3| Гаприндашвили Ш.Г. Фонетика даргинского языка. Тбилиси, 1966.

|4| Delattre P. Pharyngeal features in the consonants of Arabic, German, Spanish, French and American English. - Phonetica, v. 23, No. 2, 1971.

|5| Кодзасов С.В., Кулиев Р.Х.-М. Эндоскопическое наблюдение фарингальных и ларингальных артикуляций. - "Проблемы фонетики и фонологии", Москва, 1986.

|6| Laufer A., Condax I.D. The epiglottis as an articulator. - UCLA Working Papers in Phonetics, No. 45, 1979.

|7| El-Halees Y.A. A fiberoptic and xeroradiographic study of emphasis in Arabic. In "Abstracts of the X Int. Congress of Phonetic Sciences", 1983. Foris Publ. Dodrecht, Cinnaminson.

|8| Трубецкой Н.С. Основы фонологии. Москва, 1960.

|9| Гранде Б.М. Введение в сравнительное изучение семитских языков. Москва, "Наука", 1972.

|10| Al-Ani S.H. Arabic phonology: an acoustical and phisiological investigation. The Hague, Mouton, 1970.

|11| Ladefoged P. What are linguistic sounds made of? UCLA Working Papers in Phonetics, 45, 1979.

|12| Кодзасов С.В. Об универсальном наборе фонетических признаков. - "Экспериментальные исследования в психолингвистике", Москва, 1982.

|13| Джейранишвили Е.Ф. Фарингализованные гласные в цахуро-рутульском и удинском языках. - Иберийско-кавказское языкознание, т. XI, 1959.

|14| Кодзасов С.В. Фонетика арчинского языка - Опыт структурного описания арчинского языка, т. 1. Москва, 1977.

IN DEFENSE OF THE PHONETIC ADEQUACY OF THE TRADITIONAL TERM "VOICED ASPIRATED"

R. Prakash Dixit

Department of Linguistics, University of California
Los Angeles, California 90024-1543, USA

ABSTRACT

This paper defends phonetic adequacy of the traditional term "voiced aspirated" as a descriptor and classifier of a fourth category of homorganic stops such as /bh/, which has been questioned in recent phonetic literature.

INTRODUCTION

In those languages that possess four manner categories of homorganic stops such as /p/, /ph/, /b/ and /bh/, the fourth category has been, traditionally, described and classified as "voiced aspirated". This description and classification of the fourth category of stops has long been considered adequate phonologically as well as phonetically. Recently, however, the phonetic adequacy of the term "voiced aspirated" as a description and classifier of the fourth category of stops has been questioned. According to Ladefoged [7] "when one uses a term such as voiced aspirated, one is using neither the term voiced nor the term aspirated in the same way as in the description of the other stops." That is, unlike the voiced unaspirated stops which are produced with normal closure voicing, the closure voicing during the voiced aspirated stops is not normal [3,9]. Moreover, the voiced aspirated stops are also not aspirated either, since during their production the release of the oral closure is not followed by a period of voicelessness. They are thus unlike the voicless aspirated stops where the release of the oral closure is immediately followed by a period of voicelessness [1,3,7,9].

There has been some confusion in the phonetic literature as to what aspiration really is. Part of this confusion can be, perhaps, attributed to Lisker and Abramson's [11] work on voice onset time associated with stop consonant production, although to no fault of theirs, as they did not consider voiced aspirates. Their findings on voice onset time led them to regard the "noise feature of aspiration"..."simply as an automatic concomitant of a large delay in voice onset". Unfortunately, "the noise feature of aspiraton," which provided the phonetic basis for the description and classification of the voiced aspirated stops as aspirated, was forgotten and the "large delay in voice onset" or "voicing lag" became the equivalent of aspiration. From then on most phoneticians and linguists used these terms in the sense of aspiration. Thus, the voiced aspirated stops were considered phonetically neither voiced nor aspirated and a few new terms such as "whispery voiced", "breathy voiced", "murmured", "murmured aspirated" and "voiced phonoaspirated" were suggested as phonetically more adequate replacements of the term "voiced aspirated".

The purpose of this paper is to examine and discuss the phonetic adequacy or inadequacy of the various terms mentioned above in the light of glottographic, aerodynamic and spectrographic data from Hindi (a four-category Indo-Aryan language) and to show that the term "voiced aspirated" is a better phonetic descriptor and classifier of the fourth category of stops than its suggested replacements.

EXPERIMENTAL RESULTS

The results presented here are based on the analysis of a large body of data. Although only a few illustrations are given here, they may be taken as typical of the data as a whole. Glottographic, aerodynamic and spectrographic data from one speaker of Hindi are presented in Figures 1, 2 and 3, respectively. These figures display records obtained during the nonsense words /pipi/, /phiphi/, /bibi/ and /bhibhi/ which were produced in a frame sentence /didi -- bolije/ 'elder sister -- (please) say'. We will not consider the data on the voiceless unaspirated stop /p/ in detail in the present study; we will simply note that Photo-Electric Glottograms (PEG) in Figure 1 show that the glottis is slightly apart during the initial /p/. The figure also shows that /b/ is produced with an approximated glottis and vibrating vocal folds; while /bh/ and /ph/ are produced with a moderately and a widely open glottis, respectively. The glottal opening during /bh/ begins appreciably before the oral release, peaks around the middle of the noise interval and terminates during the initial part of the following vowel. However, during /ph/ the glottal opening starts at or slightly prior to the articulatory closure, peaks at or near the articulatory release and terminates during the early portion of the following vowel. Notice that the glottal opening during /ph/ is approximately double that during /bh/.

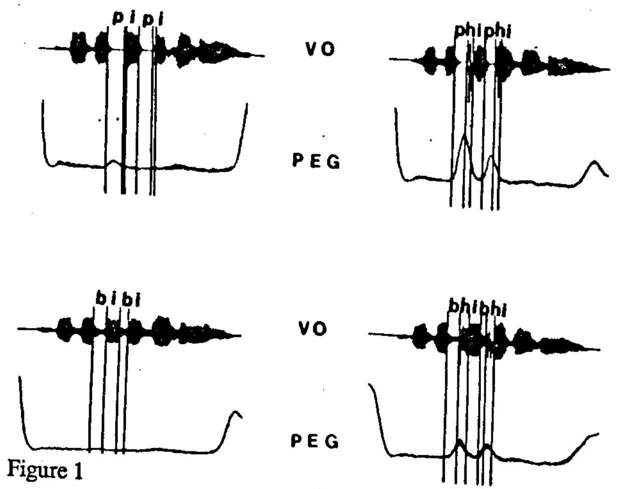


Figure 1

Oral air pressure (Po) and oral air flow (Uo) curves in Figure 2 show that the pressure profiles and the magnitudes of pressure during the articulatory closure for /b/ and /bh/ are about the same, but the magnitude of flow after the articulatory release is much greater for /bh/ than for /b/. However, the pressure profile as well as the magnitude of pressure and flow for /ph/ are different than those for /b/ and /bh/. For /ph/ the pressure rise is rapid, the pressure build up is higher and the flow rate is greater than each of these for either /b/ or /bh/.

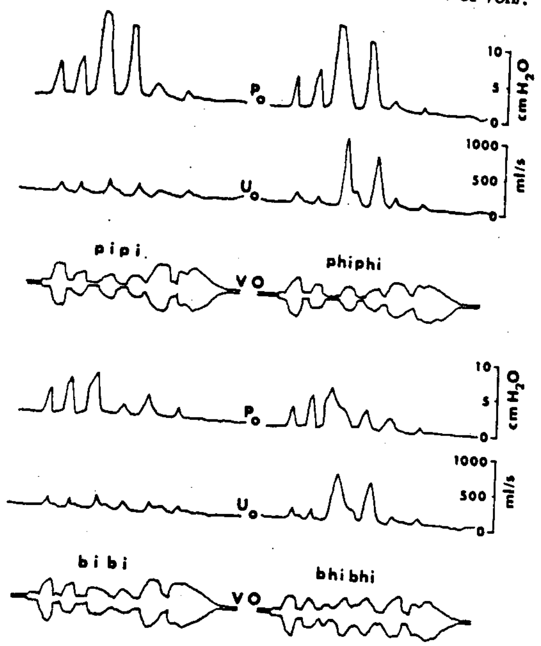


Figure 2

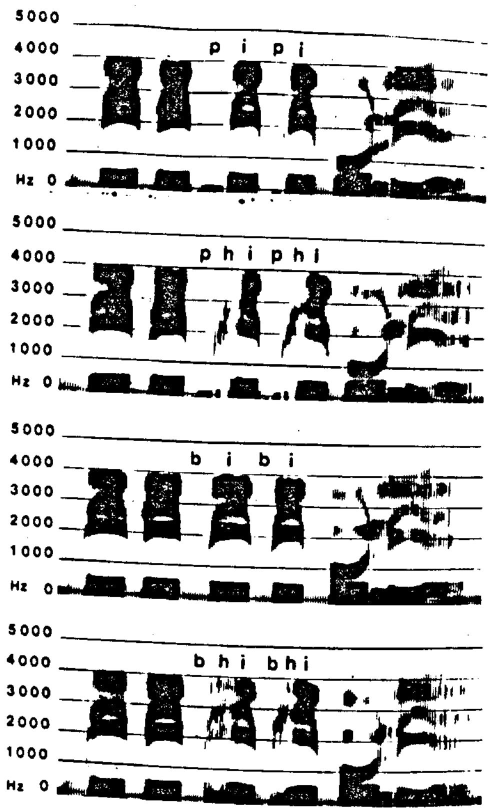


Figure 3

The spectrograms in Figure 3 show that the closure interval of /ph/ is mostly voiceless, except for a few vertical striations indicating vocal fold vibration continuing from the preceding voiced environment. The closure intervals of /b/ and /bh/, on the other hand, are fully voiced. The acoustic patterns of closure voicing for /b/ and /bh/ appear to be virtually identical. Notice that a period of voicelessness occurs between the articulatory release of /ph/ and the onset of the following vowel. However, such a period during /bh/ is not voiceless; it is occupied by a fuzzy acoustic pattern of vertical striations, which appear to be quite different from the one observed during closure interval of /bh/ or /b/. On the other hand, during this period acoustic noise can be observed for /bh/ and /ph/ alike, in about the same frequency regions as the resonances of the vowel following these stops.

DISCUSSION

The spectrographic data presented above clearly demonstrate that the voiced aspirated stops of Hindi are both voiced and aspirated. They are voiced because they are produced with regular closure voicing like the voiced unaspirated stops and they are aspirated because, like the voiceless aspirated stops, they are produced with glottal noise - in other words aspiration - following the release of oral

closure. However, the period following the release of oral closure, during which aspiration occurs, is voiceless in the voiceless aspirated stops but not in the voiced aspirated stops; in the latter category of stops the vocal folds continue to vibrate through this period. In the voiced aspirated stops the mode of vocal fold vibration during the period of aspiration is apparently different from that during the closure interval. This is reflected in the fuzzy pattern of vertical striation during the aspiration vis-a-vis the clear pattern of vertical striations during the closure interval (Figure 3). The difference in the mode of vocal fold vibration within the voiced aspirated stops is further reflected in an approximated glottis and an extremely low flow rate during the closure interval versus a moderately open glottis and a high flow rate during the noise or period of aspiration (Figures 1 and 2).

During the aspiratory interval of these plosives the glottis is only moderately open and the vocal folds are relatively slack; consequently they can vibrate in the absence of an articulatory obstruction to the airflow. But they do not touch one another while vibrating. On the other hand, during the aspiratory interval of the voiceless aspirated stops the glottis is widely open and the vocal folds are relatively tense; they simply cannot vibrate even in the presence of high flow rate. Thus, the aspiratory interval of the voiceless stops is voiceless but that of the voiced stops is not. Further, the spectrograms in Figure 3 show that the acoustic noise during the voiceless aspirated as well as voiced aspirated stops is found in about the same frequency regions as resonances of the following vowel, which indicated that the noise source is located at the glottis. If the source were located elsewhere a different noise pattern would have resulted. Glottal noise is commonly called aspiration. Thus, the phonetic description of the voiced aspirated stops as aspirated cannot be reasonably rejected [4]. Likewise, their phonetic description as voiced is also unquestionable. In anticipation of the forthcoming aspiratory phase the glottis begins to open appreciably before the release of the oral closure; but the vocal folds continue to vibrate and remain fairly close together almost until the articulatory release, as attested by fiberoptic observations [2,6]. Those sounds in which the vocal folds "form a closure or near closure duing successive periods of the oscillation' are said to be regularly voiced [12]. Thus, the traditional term "voiced aspirated" is an adequate phonetic descriptor and classifier of the fourth category of homorganic stops. Further "voiced aspirated" is a better term since it produces a symmetrical matrix of classificatory terms, and is capable of capturing phonological generalization, while its suggested replacements produce an asymmetrical and counterintutive matrix and create problems in the description of such sound changes as Grassman's Law, as shown by Halle [5].

On the basis of the definitions of phonetic terms given in Petrerson and Shoup [12], Benguerel and Bhatia [2] have proposed the term "voiced phonaspirated" as a phonetically more adequate descriptor than the term "voiced aspirated" for the fourth category of homorganic stops. There is no problem with the term "voiced" which adequately describes what happens during oral closure. However, the term "phonoaspirated" which describes what happens after the release of oral closure is problematic. It appears that "phono" in "phonoaspirated" was prefixed to "aspirated" to indicate the particular mode of vocal fold vibration which occurs during the aspiratory period and which is different from the one that occurs during the oral closure. However, "phono" is also prefixed to "constricted" in the term "phonoconstricted" where it indicate a very different mode of vocal fold vibration from the one that occurs during the aspiratory interval. As "phono" describes two entirely different modes of vocal fold vibration, it renders the term "phonoaspirated" phonetically inadequate, and thus unacceptable.

The discussion of phonation types in Catford [3] suggests the term "whispery voiced" for the voiced aspirated stops. He says that "the fact is that in such sounds as [bh] there is whispery voice rather than voice during the stop and for a certain period after its release." The glottal stricture used in the production of whispery voice is described by Catford as "narrowed vibrating." The degree of opening according to him is less than 25% of maximal glottal opening, while during voiceless stops it is from 60 to 95% of maximal glottal opening. Let us assume that the opening for the voiceless aspirated is 95% of maximal glottal opening. Now recall that the degree of glottal opening for the voiced aspirated stops in the data presented here was about half of that for the voiceless aspirated stops. Thus, the voiced aspirated stops were produced with more than 45% rather than less than 25% of maximal glottal opening. Obviously, they were not produced with whispery voice, since the glottal stricture was inappropriate for the generation of whispery voice ,being twice as wide as required for such a phonation. Moreover, Catford's assumption that the glottis is in the phonatory posture for whispery voice during the stop phase is also not borne out by the data. But even if such a posture were present during the stop phase it could not generate whispery voice in the presence of a supraglottal obstrucion. Thus, the term "whispery voiced" instead of "voiced aspirated" is phonetically inapproopriate.

Ladefoged has suggested the terms "murmured" [7,8,9] and "murmured aspirated" [10] in place of the term "voiced aspirated" for the phonetic description of the fourth category of homorganic stops. According to him [9] "murmured sounds are sometimes made...with the glottis fairly open at one end. They can also be made with a narrower opening extending over nearly the whole length of the vocal cords." Ladefoged has called both these physiological possibilities the "murmur" state of the glottis and has assumed that such a state occurs during the oral closure as well as after the release of the closure in the so-called murmured or murmured aspirated stops [8,10]. Thus the vocal fold vibrations that occur during the oral closure are assumed to be of "the kind that would be expected from a small volume of air flowing through the glottis while it is in the position for a murmured sound" [8], that is the vocal fold vibrations during articulatory closure are said to be different from those in normal voice vibrations. However, these assumptions do not find support in the glottographic and acoustic data presented here. The phonation that is generated after the release of a closure was earlier [7,8,9] described as murmur or breathy voice. Later [10] in relation to the somewhat different voiced aspirated stops of Owerri Igbo it was surprisingly though unjustifiably described as aspiration, (surprisingly since "aspiration" for Ladefoged is " a period of voicelessness during and immediately after the release of an articulatory stricture" [7]). This was the result of redefining aspiration in an attempt to accomodate the voiced aspirated and the unvoiced aspirated stops under the same phonetic category of "aspiration". The attempt, however, did not succeed since the closure voicing in the voiced aspirated stops of Owerri Igbo was still considered to be murmur, which is almost certainly contrafactual.

Lately, Ladefoged has changed his position. In the second edition of his book *A Course in Phonetics* he states that "voicing during the vowel and the closure are, as usual, the result of air flowing between the vocal cords while they are held loosely, fairly close together". In other words the closure voicing in the voiced aspirated stops is normal regular voicing. This is strongly supported by the glottographic and acoustic data presented here. Further, the vibrations that occur after the articulatory release of these stops are described as "murmured (breathy) vibrations". That is, after the articulatory release of the voiced aspirated stops "murmur" or breathy voice" occurs. Breathy voice may as well be called "voicy aspiration".

As we have seen, in the voiced aspirated stops of Hindi aspiration is accompanied by glottal vibration which noticeably changes its quality. It sounds more like breathy voice. However, it would be inappropriate to call the voiced aspirated stops "breathy voiced," because this term describes only the state of the glottis after the release of the closure in these stops. On the other hand, the term "voiced breathy voiced" sounds strange. If the term "murmur" could be strictly limited to the meaning of "breathy voice," then perhaps a term like "voiced murmured" could be suggested to describe the fourth category of homorganic stops. The seeds of this term were already present in Lisker and Abramson's [1] and Benguerel and Bhatia's [2] work, but for some reason they did not suggest it. If this term is accepted then a diaeresis [..] should not be used under the stop part of the consonant, since it will give the wrong impression that the closure voicing is murmur rather than regular voicing.

To conclude, the term "voiced murmured," although phonetically adequate, will produce an asymmetrical matrix of classification that will fail tó capture phonological generalizations. However, it may turn out to be a useful term in speech synthesis. On the other hand, the term "voiced aspirated" is not only phonetically adequate but also produces a symmetrical matrix of classificatory values and is capable of capturing phonological generalizations. It will thus be more attractive to the linguist. The other terms which really are both phonetically and phonologically inadequate should be discarded.

REFERENCES

[1] D. Abercrombie, Elements of Genral Phonetics. Edinburgh University Press, 1967

[2] A-P. Benguerel, T.K. Bhatia, Hindi stop consonants: an Acoustic and Fiberscopic study, Phonetica 37: 134-148, 1980

[3] J. C. Catford, Fundamental Problems in Phonetics. Indiana University Press, 1980

[4] R.P. Dixit, On Defining Aspiration, Proceedings of the XIIIth International Congress of Linguists, CIPL, 1983

[5] M. Halle, Review of P. Ladefoged, Preliminaries to Linguistic Phonetics, Language 49: 926-933, 1973

[6] R. Kagaya, H. Hirose, Fiberoptic, Electrogmyographic and Acoustic Analysis of Hindi Stop Consonants, Ann. Bull. RILP 9: 27-46, 1975

[7] P. Ladefoged, Preliminaries to Linguistic Phonetics. University of Chicago Press, 1971

[8] P. Ladefoged, The features of the Larynx, J. Phonetics, 73-83, 1873

[9] P. Ladefoged, A Course in Phonetics. Harcourt Brace Jovanovich, 1975

[10] P. Ladefoged, K. Williamson, B. Elugbe, A. A. Uwalaka, Stops of Owerri Igbo, Studies in African Linguistics, Supplement 6:147-163, 1976

[11] L. Lisker, A. Abramson, A Cross-Language Study of Voicing in Initial Stops: Acoustic Measurements, Word 20: 384-422, 1964

[12] G. E. Peterson, J. E. Shoup, A Physiological Theory of Phonetics, J. Speech Hear. Res. 9:5-68, 1966

MAJOR DETERMINANTS OF SPEECH RHYTHM: A PRELIMINARY MODEL AND SOME DATA

EVA STRANGERT

Department of Phonetics
Umeå University
S-901 87 Umeå, Sweden

ABSTRACT

In the speech signal rhythm manifests itself in the temporal structure of stressed and unstressed syllables. This structure differs between languages and seems to be the basis for perceived rhythmic differences. At the same time there is evidence of temporal adjustments towards regularity which seem to occur irrespective of the language spoken.

It is assumed that these characteristics of the speech signal reflect two major determinants of rhythm - language structure and speech production constraints, respectively.

Some predictions based on this model are tested on three rhythmically different languages, Swedish, Spanish, and Finnish.

INTRODUCTION

That something is rhythmic means that it is temporally constrained. The impression of rhythm seems to depend on the impression of temporal regularity. In speech this regularity concerns syllables, stressed and unstressed.

While temporal regularity seems to be at the base of rhythm, different languages seem to have different kinds of temporal regularities, that is, they often sound rhythmically different. To capture such differences Pike /1/ introduced the stress-timing/syllable-timing dichotomy implying two different rhythmic principles. In a language with stress-timing, then, the regularity concerned the stressed syllables, while in a language with syllable-timing the regularity concerned all syllables, stressed and unstressed alike. Temporal regularity also implied the strongest possible temporal constraints, isochrony. Thus, in a stress-timed language stressed syllables were assumed to recur at equal intervals irrespective of the number of intervening unstressed syllables, and in a syllable-timed language all syllables, stressed and unstressed alike, were assumed to have equal duration.

Isochrony seems to be an important aspect of the perception of speech. For example, Pike /1/ based his distinction between stress-timed and syllable-timed languages entirely on the listener's impression of temporal regularity. Observations of the production of speech, on the other hand, do not support any strict regularity in the sense implied by either stress-timing or syllable-timing. Intervals between stressed syllables, and syllable durations, seem to differ within fairly wide ranges in both allegedly stress-timed and syllable-timed languages.

However, in measurements of the speech signal tendencies to temporal regularities have been found. The duration of segments and syllables seem to be inversely related to the number of unstressed syllables between stressed ones, implying a weak tendency to stress-timing . Most of these observations have been based on English but also, to a certain extent, on other languages including so-called syllable-timed languages (see /2/, p. 3-5, for a survey). There are, on the other hand, several studies in which any tendencies to temporal regularities are denied. One example is a study by Lehtonen /3/ examining the temporal structure of Finnish.

All these aspects of rhythm have to be accounted for within a general model of speech rhythm. As first step to such a model I will outline a conceptual frame for studying rhythm in speech.

A CONCEPTUAL FRAME FOR STUDYING SPEECH RHYTHM

Three basic concepts all contribute to the complex of rhythm in speech as well as in other types of rhythmic behavior: (a) grouping, (b) alternation, and (c) temporal regularity.

Grouping is the most fundamental concept. All kinds of activities seem to be organized by grouping the elements of which they are made up. Grouping occurs in both production and perception, as shown in experiments by Fraisse /4/ and Woodrow /5/. In complex activities there may be several levels of organization. One group at a higher level may contain two or more groups at a lower level. Such hierarchical grouping is very obvious in music but it seems to be a characteristic also of speech and other kinds of human activities. Thus grouping may be seen as a general means for structuring information, and therefore what we perceive as rhythm may be a consequence of a natural way of handling information.

Alternation often characterizes a sequence of elements. Normally some elements in a sequence are marked from the others, for example by being longer or more intense. The marked elements will then alternate with the unmarked ones. Alternation is an important basis for grouping as groups are delimited from one marked element to another. However, grouping also occurs when there is no alternation at all in a sequence of elements. In this case grouping may be achieved by marking some of the originally unmarked elements.

Grouping may lead to temporal regularity. Grouping elements together means that there are temporal constraints on how they are processed. Related elements have to be kept together and will accordingly constitute a unit in the temporal domain, also. The basis for this temporal unity of groups might be a cyclic and regular processing of information. This means that in groups with many elements there has to be a temporal compression of these elements, while in groups with few elements no such compression will be needed. Such compression in longer groups has been reported in several studies. It even seems to be a tendency to adjust differently-sized groups towards an intermediate or average duration /4/.

Grouping, alternation, and temporal regularity all contribute to the complex of speech rhythm as outlined in the following simple input-output model:

MESSAGE STRUCTURE → ARTICULATORY PLANNING → RHYTHMIC STRUCTURE

For a more detailed description, see /2/, p. 19-27.

The message structure

By message structure I refer to all structural information which is needed for uttering a short sentence or a phrase: phonologic, prosodic, syntactic, semantic, as well as pragmatic specifications. Thus, it is a situationally coloured language structure.

The message structure differs widely between languages. The most important differences as far as rhythm is concerned. include stress, quantity distinctions, and syllable and word structure. These characteristics all have to be preserved throughout the production process in order to signal the intended message.

The characteristics of the message structure form the basis for grouping. Stressed alternating with unstressed syllables would be such a basis giving groups of one stressed and a number of unstressed syllables. Such a stress group is commonly defined as one stressed syllable and all following unstressed ones up to the next stressed syllable irrespective of word and syntactic boundaries.

However, both word boundaries and syntactic boundaries might be alternative bases for grouping. Thus, there may be words or word groups beside stress groups. Support for such alternatives may be found in /6/. Other characteristics of the message structure may be used too, and it seems most reasonable to suppose that different languages use different kinds of bases for groupings. Also, in each specific language there may be a certain optionality in the choice of what to base the grouping on. Different alternatives may "compete" with each other. What determines the specific kind of grouping may be the situation as a whole and the specific intentions of the speaker.

Articulatory planning

In the articulatory planning grouping is a means of structuring information. Grouping is assumed to occur at two levels at least. At the first one the input string is restricted to contain units about the size of a short sentence or a phrase constituting the message structure as described above. The basis for this may be intonation characteristics coinciding with syntactic boundaries and delimiting semantically coherent units.

At the next level this string is scanned for elements to base further grouping on, for example stressed syllables or different kinds of boundaries as suggested above. Thus there will be stress groups or possibly words or word groups.

If the groups contain more than just a few elements there may be further subgrouping. Most reasonably, this would be the basis for rhythmic alternation of unstressed syllables. In this case grouping seems to be achieved by strengthening some in a string of several unstressed syllables /7/.

The planning system converts the elements in the input string into articulatory coded units. These units are then converted into commands to the motor execution system and eventually transformed into acoustic events in the speech signal.

The rhythmic structure

By rhythmic structure, the output of articulatory planning, I refer to those aspects of the temporal structure of the speech signal which are the basis for the impression of rhythm. Rhythmic structure is a result of both the articulatory planning and the message structure, as effects of both will be mixed in the speech signal.

However, within the conceptual frame as given above, together with careful analysis of the specific data, the two effects may be separated. Furthermore the strength of each may be predicted in each specific case.

TESTING SOME PREDICTIONS OF THE MODEL

Data

I will present some data from Swedish, Spanish, and Finnish, chosen so as to represent rhythmically different languages. Referring to the frequently-used rhythmic dichotomy Swedish would be a stress-timed and Spanish a syllable-timed language. Finnish, though difficult to categorize in rhythmic terms, was chosen for the complexity of its quantity system. This would make it possible to test the interplay between temporal constraints of the planning process and the constraints of the input structure.

The material consisted of sentencs which were syntactically and semantically similar in all three languages. They all had an invariant frame in which test words with different numbers of syllables were inserted. The sentences were read in a neutral manner without giving prominence to any specific word.

The data are more thoroughly accounted for in /2/, p. 117-146:

The predictions against data

1 There will be similar temporal adjustments to regularity irrespective of the language spoken.

If grouping is a natural means of structuring information and tendencies to temporal adjustments are a consequence of grouping, then temporal adjustments should occur in languages in general. Also, as grouping occurs hierarchically, there should be adjustments on several levels,for example, (a) the phrase and (b) the stress group.

The effects of articulatory planning, then, will be temporal adjustments of segments and syllables so that they are more compressed the more elements there are in a unit.

However, there are no claims regarding isochrony and in effect, no timing rules at all are implied.The temporal adjustments are seen simply as a consequence of the assumed tendency to cyclic and regular processing as outlined above.

The data support the prediction. There are similar temporal adjustments decreasing the temporal differences between stress groups with different numbers of syllables in all three languages. In stress groups with one, two, and three syllables the duration of the first (stressed) syllable decreased successively upon the addition of the second and third syllable. Figure 1 gives an example from Spanish. Thus, temporal adjustments associated with stress-timing seem to occur also in languages assumed to be syllable-timed. Assuming that the temporal adjustments are related to articulatory planning, it seems that the stress-timing/syllable-timing distinction does not reflect different planning strategies.

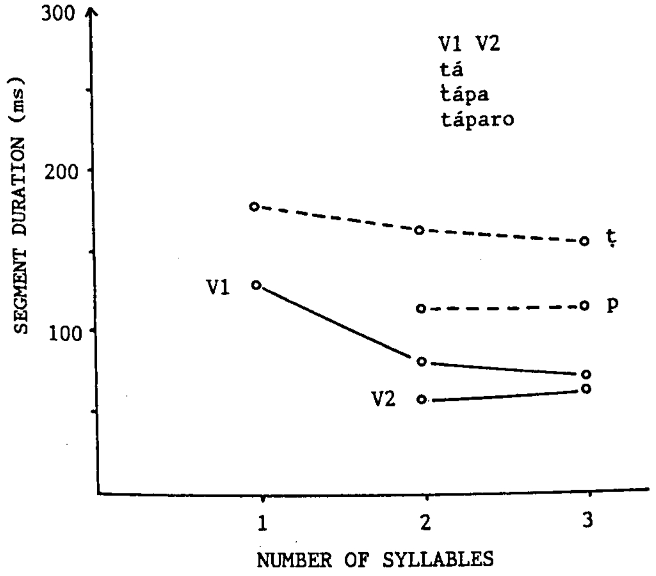


Figure 1. Duration of vowels and consonants in three Spanish test words as a function of the number of syllables (1-3) in the test word. The words were inserted in a sentence frame. N=6. From /2/.

2 Differences in rhythmic structure between languages is a consequence of structural differences.

The message structure, without doubt, is an important determinant of rhythmic structure. The reason naturally is that many of the characteristics of the input structure carry the burden of functional distinctions which have to be preserved in order to convey the intended message to the listener. Especially important characteristics are (a) stress, (b) quantity distinctions and (c) syllable structure.

It seems most reasonable to assume that such differences form the basis for the stress-timing/syllable-timing dichotomy. So-called stress-timed languages, for example, seem to have a clear distinction between stressed and unstressed syllables. In general, the stressed syllables have a more complex structure than the unstressed syllables. In so-called syllable-timed languages, on the other hand, stressed and unstressed syllables are structurally more alike /8/.

There is empirical support also for the second prediction. Language-specific structural characteristics and their temporal manifestations differ widely in the three languages. The greater similarity, structural and temporal, of stressed and unstressed syllables in Spanish as compared to Swedish contributes to making rhythmic structure quite different in the two languages. And the elaborate quantity system in Finnish contributes to the characteristics of rhythmic structure in this specific language.

3 Temporal adjustments to regularity will only occur insofar as functionally important structural features are not destroyed.

The planning mechanism is sensitive to the specifications in the message structure. Therefore, general characteristics of articulatory planning will be temporally reflected only when the temporal aspects of planning and these specifications do not conflict. When in conflict, the constraints of the message structure take precedence over, or simply obscure, temporal constraints of planning. Such conflicts may arise more often in elaborated than in neutral renditions of speech. The maintenance of certain structural distinctions may also produce such conflicts. They may occur, for example, in languages with elaborated quantity systems.

An analysis of the Finnish data point to a complex control of articulatory planning. Obviously there is complex interplay between the constraints of the input structure and the planning mechanism. This interplay seems to be conditioned by the demands of the quantity distinctions in Finnish. As is well known, both vowels and consonants are phonologically either long or short in Finnish, and this distinction has to be made in both stressed and unstressed syllables. In the speech output the phonological distinction is reflected in segments of longer or shorter duration. What happens in Finnish is that there are temporal adjustments to a certain regularity of stressed syllables in some cases but not in others. Phonological length seems to be an important conditioning factor, as only long vowels and consonants are compressed to any significant degree. A second conditioning factor seems to be whether the temporal adjustments will obscure important quantity relations or not. Thus, compression only occurs insofar as it will not affect the quantity relation between the first and second syllable in a word. Figure 2 showing two different cases, one with (a) and the other without (b) temporal adjustments, illustrates this conditionality.

This may be the reason why Lehtonen /4/ found no compression effects in Finnish. His study was based mainly on such quantity patterns in which compression would be very restricted.

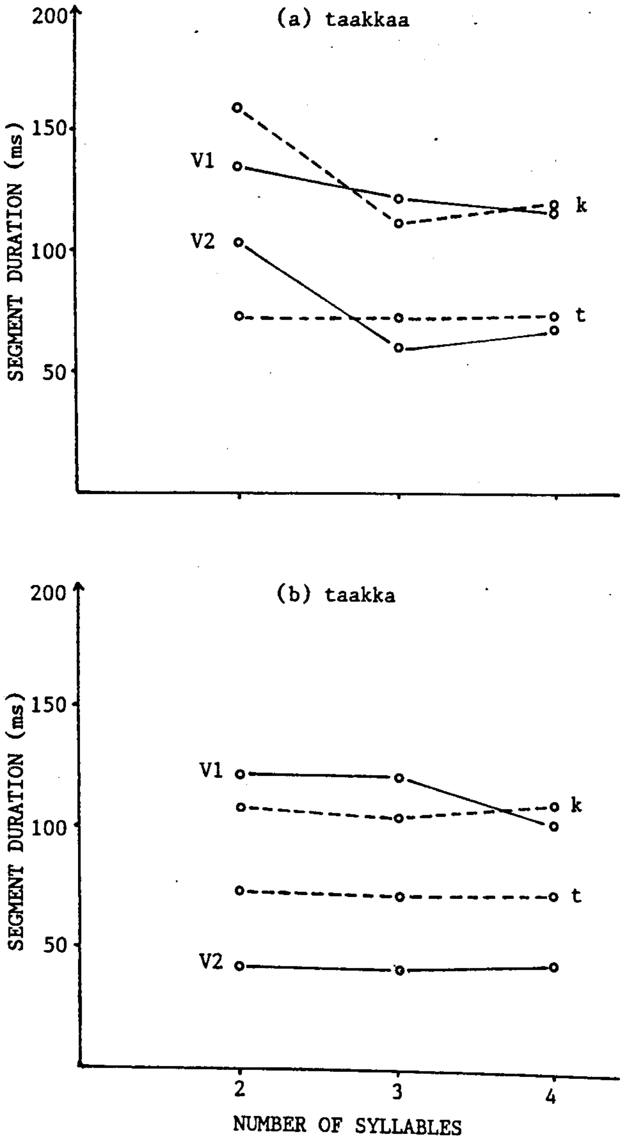


Figure 2. Duration of vowels and consonants in Finnish two-syllable sequences as a function of the number of following syllables. The test words contained 2-4 syllables: (a) taakkaa and (b) taakka followed by nsa and nsa + han in words with three and four syllables, respectively. The words were inserted in a sentence frame. N=6. From /2/.

CONCLUDING REMARKS

The conceptual frame as outlined fits well to the observations in the three languages which were chosen so as to represent different kinds of rhythm. The data reveal the expected differences as well as the similarities between the three languages. Thus the frame may be used as a starting-point for further research on speech rhythm.

A more detailed account of the contents of this paper is given in /9/.

REFERENCES

/1/ Pike, K.L. 1946. The Intonation of American English. Ann Arbor: University of Michigan Press.

/2/ Strangert, E. 1985. Swedish Speech Rhythm in a Cross-Language Perspective. Umeå Studies in the Humanities 69. Stockholm: Almqvist & Wiksell International.

/3/ Lehtonen, J. 1974. Word length and sound durations. Virittäjä, 78, 160. (English summary of a paper in Finnish).

/4/ Fraisse, P. 1982. Rhythm and tempo. In Deutsch D. (Ed.) The Psychology of Music, 149-180. New York: Academic Press.

/5/ Woodrow, H. 1951. Time perception. In Stevens, S.S. (Ed.) Handbook of Experimental Psychology, 1224-1236. New York: Wiley.

/6/ Fischer-Jørgensen, E. 1982. Segment duration in Danish words in dependency on higher level phonological units. Annual Report of the Institute of Phonetics, University of Copenhagen, 16, 137-189.

/7/ Bruce, G. 1984. Rhythmic alternation in Swedish. In Elert, C.-C., Johansson, I. & Strangert, E. (Eds.) Nordic Prosody III, 31-41. Umeå Studies in the Humanities 59. Stockholm: Almqvist & Wiksell International.

/8/ Dauer, R.M. 1983. Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51-62.

/9/ Strangert, E. 1987. Speech rhythm: Data and preliminaries to a model. Forthcoming.

ENGLISH STOP ALLOPHONES IN METRICAL THEORY

John T. Jensen

Department of Linguistics
University of Ottawa
Ottawa, Ontario K1N 6N5 Canada

ABSTRACT

The foot, a prosodic unit containing one stressed syllable, is the domain for determining the allophones of stops in English. Aspiration is restricted to foot-initial position. Consonants are laxed within a foot after a nonconsonantal segment and lax voiceless stops are glottalized in syllable codas; lax alveolar stops are flapped syllable initially. Some revisions to the rules establishing feet are proposed. Because the metrical grid provides no constituents, it is not adequate for predicting the distribution of stop allophones in English.

INTRODUCTION

In contemporary phonology there is general agreement that representations need to be enriched with prosodic organization, including such units as the syllable and the foot. This view contrasts sharply with the practice of early generative phonology [1], where phonological representations consisted entirely of strings of segments and boundaries. The original motivation for metrical theory was to offer a more natural account of stress systems [11], but it soon became apparent that prosodic organization also allows for the correct description of certain segmental processes as well. Aspiration of voiceless stops in English, for example, occurs in a variety of disparate environments. Selkirk [15] lists word-initial position *(Toronto)*, before a stressed vowel unless [s] precedes *(hotel* vs *astonish)*, before a sonorant plus a stressed vowel unless [s] precedes or [t] is followed by [l] *(apply* vs *display, Atlantic)*. Such a process is difficult to describe in purely segmental terms, and indeed no systematic account of stop allophony appears in *The Sound Pattern of English* [1]. Selkirk observes correctly that aspiration occurs only in syllable-initial position, which partially accounts for these observations. In order to account for the nonaspiration of the underlined stops in words like *happy, hefty,* Selkirk proposes language-particular resyllabification rules that attract consonants leftward out of stressless syllables, giving *happ.y, heft.y,* thereby removing these stops from the domain of aspiration. While I find these resyllabifications counterintuitive, there is an empirical argument against this analysis. Selkirk's resyllabification rules are subject to a structure-preservation principle that requires derived syllables to conform to the canonical syllable patterns of the language. In a word like *At.kins*, resyllabification to **Atk.ins* is impossible, since English syllables never end in *-tk.* This predicts that [k] of *Atkins* should aspirate, which it does not, any more than the [t] of *actor,* where *act.or* would be a possible resyllabification.

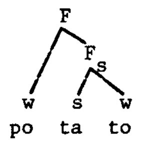
A second syllable-based approach to English stop allophones is that of Kahn [8], which is couched in terms of autosegmental phonology. Instead of resyllabification, Kahn allows consonants to be ambisyllabic, i.e., part of both the preceding and following syllables. This would be the case of [p] in *happy* and [t] in *hefty,* for example. Kahn's rule aspirates voiceless stops that are syllable initial but not syllable final (i.e., not ambisyllabic) and thus achieves the same effect as Selkirk, and runs into the same difficulty with *Atkins*. Since [k] here can't be ambisyllabic, he wrongly predicts that it should aspirate. (In fact, he claims that it does aspirate in slow speech, but I find this possible only in very careful speech where both syllables are stressed.)

Kiparsky [9] was the first to propose that the stress feet of Liberman and Prince [11] could also be considered the domain of certain segmental processes. Instead of resyllabification or ambisyllabification, Kiparsky proposed rule (1) (modified).

(1) $C \rightarrow [-\text{tense}]\ /\ \ldots[-\text{cons}]\_\_\_\_$ within a foot

Kiparsky restricts aspiration to tense voiceless stops at the beginning of a syllable, thus accounting for *happy*. But Kiparsky predicts aspiration on the second syllables of *hefty, Atkins*, where the stops [t] and [k] are unaffected by rule (1), since they are preceded by [+consonantal] [f] and [t] respectively. Hammond [4] notices such problems with the foot-based analysis, and advocates a return to Kahn's ambisyllabic approach. I propose to retain the foot-based approach, but to restrict aspiration to foot-initial position. Some modification of Kiparsky's system is needed anyway. Working within the original metrical framework [11], Kiparsky retained the feature [±stress] and with it the possibility of stressless feet. He analyzes *potato* as two feet, the first unstressed, $[_{F}\text{po}][_{F}\text{tato}]$, predicting aspiration on the foot-initial [p] and [t] and flapping (via laxing) of the second [t]. Since then, metrical theory has rejected the feature [stress], holding that stress is the property of being the strongest syllable in a foot [14]. If [po] of *potato* is not a foot, and aspiration is limited to foot-initial position, how does the [p] come to be aspirated? Hayes [5, 6] proposes that stray syllables (i.e. those not associated with any foot) are adjoined to an adjacent foot. If we assume that adjunction creates nested feet, we get the representation in (2) (where w=weak, s=strong, F=foot; s and w must always appear as sisters to each other.

(2)



In (2), both [p] and the first [t] are foot initial, and so get aspirated, while the second [t] laxes and flaps, as in Kiparsky's treatment. This captures the essence of Kiparsky's proposal, and resolves Kiparsky's problem with *Atkins* and *hefty*.

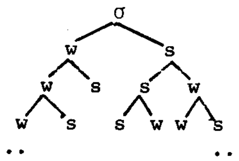
Subsequent studies have confirmed the role of the foot in segmental phonology as well as stress systems. Prince [12] states rules for gradation and overlength in Estonian partly in terms of foot structure. Similarly, Hayes [7] discusses certain segmental processes in Yidin^y, an Australian language, in terms of foot conditioning, thereby obviating the necessity for phonological rules to refer to the odd-numbered syllables in a word. Even for stress systems, constituency is necessary. Halle and Vergnaud [3] cite a number of studies showing that deletion of a (potentially) stressed vowel in many languages results in a stress shift to an adjacent syllable within the foot.

In the remainder of this paper, I will first review the properties of syllables, propose some modifications to Hayes's rules of foot construction, then show the role of the foot in English stop allophones, making crucial reference to rule (1).

SYLLABLES

The acoustic record provides no direct evidence of syllables and their boundaries. The syllable is an abstract unit which makes it possible to provide a more insightful statement of certain phonological processes. Among competing approaches, we assume the metrical representation of Kiparsky [9], in which the syllable has the same type of s-w labelling as the foot, as in (3).

(3)



In this representation, sister constituents are required to observe the sonority hierarchy, according to which segments are ordered (from weakest to strongest) as stops, fricatives, nasals, l, r, glides, vowels. In addition, English imposes language-specific constraints on onsets and rimes. For example, a syllable cannot begin with a sequence of two stops (including nasals), and the rime is limited to the sequence V([+sonorant])(C)([+coronal]), where the 'coronal' position may exceptionally contain [st] or [sθ] as in *next, sixth*. An additional [s], [z], [t], or [d] may follow if it is inflectional, e.g. *sixths*. Even though contrary to the sonority hierarchy, [s] plus voiceless stop can occur in the onset, and also the sequence [s] plus voiceless stop plus liquid (but not **stl-)*.

The question of dividing between syllables is a more difficult one. There is no difficulty with *At.kins*, which can only be syllabified as shown. Kiparsky proposed that, in English, the onset is maximized when two or more divisions are possible at the boundary between two syllables. In this respect English contrasts with Finnish and Estonian, where the coda is maximized, and where, in general, the onset is limited to a single consonant. This raises an interesting question: what happens to VsT(R)V clusters (where T=voiceless stop, R=liquid or glide)? The sonority hierarchy predicts Vs.T(R)V; the onset maximization principle predicts V.sT(R)V.

Davidsen-Nielsen [2] investigated this question experimentally. He measured the degree of aspiration in words like *despise* and compared it to that of words like *pin* (aspirated) and *spin* (unaspirated). Measurements revealed that the stops in words like *despise* are normally unaspirated, supporting the syllabification V.sT(R)V. The only exceptions were in words that contained "a prefix with *-s* followed by an intuitively transparent morpheme boundary, e.g. *miscalculate, discourteous,"* where the stops are aspirated, thus supporting a syllable division coinciding with the morpheme boundary, i.e. Vs.T(R)V in these words.

In a sense, or course, this argument is circular. The syllable boundary is inferred from the degree of aspiration on the stop, while the rule for aspiration is assumed to affect only syllable-initial stops. (The stops in question are also foot initial, and so consistent with our hypothesis also.) However, this conclusion is independently supported by the stress pattern of these words. The prefixes *mis-* and *dis-* exhibit secondary stress, and we might expect to find similar effects from a preceding stressed syllable, even if it doesn't constitute a morpheme. In Davidsen-Nielsen's material, *gèstátion* and *fàstídious* have a somewhat greater average degree of aspiration (3.0 and 3.09 csec respectively) than *bestow* and *establish* (2.5 and 2.41 csec, respectively. To test this further, I recorded two speakers of North American English in words containing *s*-stop clusters, both with and without stress on the preceding syllable. Results were analyzed using a Mingograf 804 connected to a Kay Elemetrics Visipitch 6087 and also with a Kay Elemetrics Sonagraph 7800. The results are given in Table 1.

		JJ	ML
a.	ìnfèstátion	2	2.5
	èlàstícity	5	3
	plàstícity	3	2
	pèstíferous	2	2
	òsténsible	2	2
	MEAN	2.8	2.3
b.	askánce	1	2
	órchestra	2	2
	astónish	1.5	1
	pédestal	1	2
	sustáin	1	1.5
	MEAN	1.3	1.7

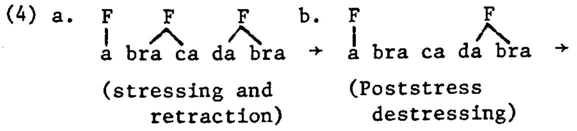
Table 1. Duration of release stage (in csec.) of medial stops (underlined) in words with (a) and without (b) secondary stress on the preceding syllable.

While these results are not conclusive, there is somewhat more aspiration in words where a secondary stress precedes the cluster in question than when the preceding syllable is unstressed. This supports the syllabification Vs.T(R)V for the words of Table 1(b).

FEET AND ASPIRATION

For the core system of English stress, Hayes [5] proposed left-dominant maximally binary feet constructed right to left across a word. Ternary feet (i.e. with three syllables) can arise only by adjoining a stray syllable to a binary foot. Syllables become stray either by being made extrametrical or as a result of destressing. Before foot assignment, the final consonant of a word, the final suffix of an adjective, and the final syllable of a noun are extrametrical. The rightmost foot may be binary only if its second syllable ends in a short (lax) vowel. Conversely, a monosyllabic foot must contain a long vowel, a diphthong, or at least one final consonant. This accounts for the familiar observation that a monosyllabic (stressed) word cannot end in a 'checked' vowel; i.e. *bee* [bi:] and *bit* [bɩt] are possible (and actual) words, while **bi* [bɩ] is an impossible word.. Prince [12] claims that Estonian is subject to the same constraint on possible feet. Other languages, e.g. French and Hungarian, are not restricted in this way. In English, the only exception to this generalization is words with an initial monosyllabic foot of the proscribed form followed by a well-formed monosyllabic foot, such as *essay* [ˈɛ,sej], *Hanoi* [,hæˈnɔj]. Hayes proposed several destressing rules, but we will be concerned with only one: Poststress destressing, which removes a binary foot whose first syllable is open and which immediately follows a monosyllabic foot. Hayes appeals to this rule in his derivation of words like *abracadabra* (4).

(4)



(stressing and retraction)

(Poststress destressing)

(stray syllable adjunction)

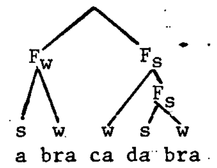
Speakers find the division (4c) counterintuitive. To test this, I asked a group of 28 native English-speaking first-year undergraduate linguistics students to divide words of this type "into two parts, according to the pronunciation." None of the subjects knew the purpose of the test beforehand. Control words were inserted into the list to prevent extraneous strategies from being used. Subjects had a printed list of words and were asked to indicate a single division between letters in each as they were pronounced by the author, with only a short interval between tokens. The results are shown in Table 2.

With two unstressed syllables flanked by two stressed syllables, the preferred pattern seems to be to join the first unstressed syllable to the preceding foot and the second to the following foot, as long as the second unstressed syllable is open. This produces the intuitively correct structure (5) from (4b).

	1/		2/		3/		other	
abracadabra	1	*3.6%*	23	*82.1%*	3	*10.7%*	1	*3.6%*
Navratilova	5	*17.8%*	22	*78.6%*	1	*3.6%*	0	
Winnepesaukee	1	*3.6%*	14	*50.0%*	11	*39.3%*	2	*7.1%*
Tippecanoe	4	*14.3%*	24	*85.7%*	0		0	
Luxipallila	3	*10.7%*	22	*78.6%*	1	*3.6%*	2	*7.1%*
Nebuchadnezzar	2	*7.1%*	5	*17.8%*	17	*60.7%*	4	*14.3%*
Kilimanjaro	1	*3.6%*	17	*60.7%*	10	*35.7%*	0	

Table 2. Division of words into two parts. Number *(and percentage)* of responses. Column headings: 1/ indicates a division after the first syllable, 2/ after the second, 3/ after the third. Other includes no division or more than one division.

(5)



If the second unfooted syllable is closed, it joins the foot on the left: 60.7% of subjects preferred the division [Nebuchad][nezzar]. Prince suggests that this is because *chadnezzar* (with the first syllable unstressed) is not a possible word type: in our terms, a possible foot [13]. On the other hand, *Kilimanjaro* may go both ways, [Kili][manjaro] or [Kiliman][jaro], since both divisions into two parts give two possible phonological words, or feet.

Prince uses facts such as these to argue against the foot as a legitimate phonological unit. Because the metrical theory of stress uses only a small fraction of the types of tree structure that the theory allows in principle, he proposes eliminating the trees and displaying relative prominence in terms of a grid, in which column height correlates with greater prominence. Such a representation has no constituents, and thus no way of capturing the segmental processes that we have seen depend on these constituents. Prince notes the virtually obligatory aspiration of the [t] of *Navratilova*, unexpected if it is metrically structured as in (4c). However, rather than discard foot theory, the answer lies in modifying it so that it will produce structures like (5), where aspiration of [k] *(abracadabra)* and [t] *(Navratilova)* is expected, under the hypothesis that aspiration of tense voiceless stops occurs only in foot-initial position.

As with syllabification, we sought instrumental verification of the proposed division into feet (5). Table 3 gives the duration of the release stage of the stops at the beginning of the third syllables of the words of Table 2 (except for *Kilimanjaro*, which has no stop in that position). Speakers and equipment are the same as in Table 1.

	JJ	ML
abracadabra	4	4
Navratilova	7	6
Winnepesaukee	5	2
Tippecanoe	5	5
Luxipallila	2	5
Nebuchadnezzar	6	4
MEAN	4.83	4.33

Table 3. Duration of release stage (in csec.) of medial stops (underlined) in potential foot-initial position

These results are consistent with the hypothesis that (5) represents the correct foot structure, on the assumption that only foot-initial voiceless stops are aspirated.

GLOTTALIZATION

Glottalization of stops is manifested differently in various English dialects. Cockney is notorious for the extent to which glottalization appears between vowels. In RP and North American dialects, glottalization is restricted to voiceless stops in syllable codas laxed by rule (1). Examples are o*c*tave, a*t*las, a*t* Lynne's. The only case where voiceless stops are glottalized in syllable-initial position is before syllabic [n], as in *kitten*. Nonrhotic speakers (e. g. RP) can also have glottalized [t] in words like *pattern*, where *r*-loss makes the [n] syllabic; North American speakers, with syllabic [r] in such words, have the expected flap. It is notable that Cockney speakers use glottalized stops (or [ʔ]) where North American speakers have the flap. In my analysis, this results from the lack of the flapping rule in British dialects, coupled with the extension of the glottalization rule to lax voiceless stops in all positions, and is especially noteworthy when it affects labial and velar stops, as in [pəjʔə] 'paper'.

In Selkirk's account, both flapped alveolars and glottalized stops are in syllable-final position as a result of her resyllabification rule. She therefore resorts to a feature [±release], claiming that alveolar stops are flapped in syllable-final position when they are released, generally before a vowel. Unreleased voiceless stops are glottalized. This runs into two difficulties, only one of which she discusses. Phrases like *get off* can only be pronounced [gɛɾɔf] by her account, with a flap. Kahn notes an alternate pronunciation [gɛtʼɔf] or [gɛʔɔf], both impossible under Selkirk's analysis, since stops are obligatorily released before vowels and thus never glottalized there. She proposes that [ʔ] is inserted before certain initial vowels under emphasis. This makes [t] unreleased, since it is followed by a nonvowel. While this works for the North American dialects she is discussing, it won't account for the Cockney facts just mentioned. The medial stop in *paper* is followed by a vowel, and there is no possibility of inserting [ʔ] under emphasis. In any case, nonrelease is not generally associated with glottalization, as many languages have phonemic released glottalized stops (e.g. Georgian). We conclude that it is more natural to describe the difference between glottalized and flapped allophones in English in terms of syllable position and dispense with the feature [release].

FLAPPING

In North American English, but not in most forms of British English, alveolar stops [t], [d], [n] are flapped within words before stressless vowels, and often between words regardless of stress. So, the second [t] of *potato* is flapped, as is the [t] in *met Ann*, although this [t] can also be glottalized. The difference depends on the syllabic status of [t] here. Kiparsky proposes a rule that flaps alveolar stops in syllable-initial position if they are lax (by rule (1)). Since [t] in *met Ann* is lax, it will flap only if it is resyllabified with the following vowel; otherwise it is glottalized. We assume that resyllabification is optional at word boundaries. Notice that, in phrases, it doesn't matter that the following vowel is stressed. What matters is that the [t] of *met* is laxed within its foot before it is syntactically concatenated with *Ann*.

SUMMARY

Laxing (1): Consonants become lax after a nonconsonantal segment within a foot.

Aspiration: Tense voiceless stops are aspirated at the beginning of a foot.

Glottalization: Lax voiceless stops are glottalized in the syllable coda. (Generalized in Cockney to all positions).

Flapping: Lax alveolar stops (including [n] are flapped in the syllable onset (North American only).

ACKNOWLEDGEMENTS

I wish to thank Melanie Lukach, Régis Foré, Margaret Stong-Jensen, Helen Goodluck, and Phil Hauptman and his Introduction to Language class for help in preparing this paper. All errors are my own.

REFERENCES

[1] N. Chomsky, M. Halle, *The Sound Pattern of English*, Harper and Row, 1968.

[2] N. Davidsen-Nielsen, "Syllabification in English words with medial *sp, st, sk,*" *Journal of Phonetics* 2, 15-45, 1974.

[3] M. Halle, J.-R. Vergnaud, "Stress and the cycle," *Linguistic Inquiry* 18, 45-84, 1987.

[4] M. Hammond, "Foot domain rules and metrical locality," *West Coast Conference on Formal Linguistics* 1, 207-218, 1982.

[5] B. Hayes, *A Metrical Theory of Stress Rules*, MIT dissertation (IULC), 1981.

[6] B. Hayes, "Extrametricality and English stress," *Linguistic Inquiry* 13, 227-276, 1982.

[7] B. Hayes, "Metrical structure as the organizing principle of Yidin^y phonology," *The Structure of Phonological Representations (Part I)* 97-110, Foris, 1982.

[8] D. Kahn, *Syllable-based Generalizations in English Phonology*, MIT dissertation (IULC), 1976.

[9] P. Kiparsky, "Metrical structure assignment is cyclic," *Linguistic Inquiry* 10, 421-441, 1979.

[10] P. Kiparsky, "From cyclic phonology to lexical phonology," *The Structure of Phonological Representations (Part I)* 131-175, Foris, 1982.

[11] M. Liberman, A. Prince, "On stress and linguistic rhythm," *Linguistic Inquiry* 8, 249-336, 1977.

[12] A. Prince, "A metrical theory for Estonian quantity," *Linguistic Inquiry* 11, 511-562, 1980.

[13] A. Prince, "Relating to the grid," *Linguistic Inquiry* 14, 19-100, 1983.

[14] E. Selkirk, "The role of prosodic categories in English word stress," *Linguistic Inquiry* 11, 563-605, 1980.

[15] E. Selkirk, "The syllable," *The Structure of Phonological Representations (Part II)*, 337-383, 1982.

LENGTH AND SYLLABIFICATION IN ICELANDIC

Margaret Stong-Jensen

Department of Linguistics
University of Ottawa
Ottawa, Ontario K1N 6N5 Canada

ABSTRACT

The domain of length in Modern Icelandic is the syllable Rhyme. Length in stressed syllables is realized by either a branching Nucleus or a branching Coda. The consonant at the end of a word is extrametrical. Arguments are presented against an analysis in which the domain of length is the syllable Nucleus. The analysis takes into account the lengthening of preconsonantal consonants in stressed syllables observed by traditional scholars. The resulting analysis predicts length by a single lengthening rule, and avoids syllable restructuring and vowel shortening rules.

INTRODUCTION

Modern Icelandic exemplifies the close relation between stress and quantity that has been observed in many languages. In Modern Icelandic, length of syllables is predictable from stress: stressed syllables are long, and unstressed syllables are short. Icelandic thus contrasts with English, in which stress is at least partly predictable from syllable length. Quantity in Icelandic has been approached from both a prosodic and a segmental point of view. For Haugen [11], quantity belongs to syllables; in particular to the Nucleus, which Haugen claims is complex in long syllables. Anderson [1] and Árnason [3] refine Haugen's proposal by saying that the Nucleus is branching in long syllables and non-branching in short syllables. Benediktsson [4] adopts a segmental approach to length, arguing that quantity can be represented at a phonemic level by the contrast between long and short consonants, with vowel length predicted by allophonic rules. I will argue for the prosodic approach to Icelandic quantity, using an autosegmental framework. I will claim that length is inherent in the syllable Rhyme rather than in the Nucleus alone.

In Icelandic, primary word stress falls on the initial syllable of a word, and secondary stresses occur in alternating patterns, with morphologically determined variations [2]. Syllables are long under both primary and secondary stress, although some shortening occurs under secondary stress [6]. In stressed syllables, long vowels and long consonants are in complementary distribution, as in (1).

(1) a.	VC:	menn	'men'(nom.pl.)	[mɛn:]
b.	V:C	men	'necklace'	[mɛ:n]
c.	V:	bú	'household'	[bu:]

A syllable with a V or VC Rhyme is not long, and a V:C: Rhyme is not permitted. Icelandic is thus unlike English, in which a VC syllable may be long, and unlike Estonian, in which a long syllable may be V:C: [14].

Ófeigsson [16] and Einarsson [5,6] have noted in addition that preconsonantal consonants are lengthened under stress, as in hestur 'horse' and iðja 'industry,' which Einarsson [6] transcribes as [hɛs·tYr̥] and [ɩð·ja]. This consonant lengthening is most apparent under contrastive stress [3] and in words used as citation forms. Ófeigsson and Einarsson transcribe the lengthened consonants as half-long, assuming a degree of length between long and short. Liberman [15] did not find any significant difference in duration between the s in last 'blame' and the s in gulast 'most yellow,' which should be short since it is in an unstressed syllable. But Liberman notes that the phonetic correlates of quantity are as yet ill-determined and may involve intensity as well as duration. Liberman concludes that consonants such as the n in mynd 'picture' and the s in last carry the "quantitative peak" and leaves the phonetic realization indeterminate. In this paper, I shall adopt Haugen's proposal [11] that preconsonantal consonants have full length phonologically under stress. I shall not address the question of their phonetic value, but I shall assume with Haugen that they may be reduced by reduction·processes operating in consonant clusters. Lengthening applies to continuants, sonorants, and voiced stops, as in (2).

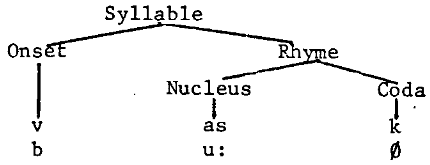
(2) a.	hafði	[hav·ðɩ]	'had'
b.	lax	[lax·s]	'salmon'
c.	sagði	[saɣ·ðɩ]	'said'
d.	sandur	[san·d̥Yr̥]	'sand'
e.	harður	[har·ðYr̥]	'hard'
f.	sigla	[sɩg̊·la]	'to sail'

Voiceless stops are preaspirated in this position [19], a topic I cannot explore here.

SYLLABIC ANALYSES

Árnason [3] proposes a syllabic account of quantity in Icelandic which incorporates lengthening of preconsonantal consonants under stress. He assumes with Haugen [11] that quantity is localized in the Nucleus and that the lengthened consonant is part of the Nucleus. Árnason represents vask [vas·k] 'sink' (acc.sg.) and bú as in (3).

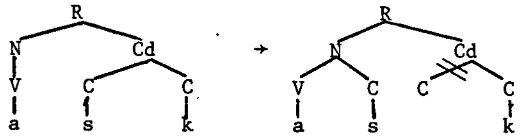
(3)



Árnason speaks of elements in the Nucleus as being "stretchable." In more formal terms, we can say that the second element in the Nucleus is lengthened, giving a long s.

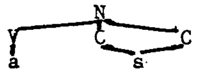
Anderson [1] gives a more formalized syllabic analysis. He too assumes that length is localized in the Nucleus. However, he does not consider the consonant lengthening. He defines stressed syllables as those that have branching nuclei, which need not be binary branching in underlying form, but are reduced to binary branching on the surface. A stressed syllable with a short vowel fills in the Nucleus by moving the consonant in the Coda into the Nucleus, as in (4), which represents the Rhyme of vask. Here, the association of the consonant to the Coda is broken and the consonant is reassociated to the Nucleus. Cd stands for Coda.

(4)



We can formalize the lengthening of the s by a rule that adds a C slot to the Nucleus, as in (5).

(5)



Ternary nuclei must be limited to this structure, since Icelandic does not have overlong vowels or overlong syllabic consonants. The need for a lengthening rule shows that Anderson's movement rule is not sufficient to account for the data. By giving up the movement rule and lengthening the consonant in its base generated position in the Coda, we will achieve the same empirical result and at the same time simplify the grammar. A grammar without the movement rule must also give up the requirement that long syllables have branching nuclei, since the Nucleus in vask will be non-branching.

Another problematic aspect of length is the difference between monosyllables and polysyllables. In monosyllables, a vowel or diphthong is long if it ends the word, as in skó [skou:] 'shoe' and bú [bu:], or if it is followed by just one consonant, as in skip [sk$_j$ɩ:p^h] 'ship' and hár [hau:r] 'hair.' (It is debatable which part of the diphthong is lengthened, Haugen [11] claiming it to be the off-glide. I will not pursue the matter here.) In monosyllables ending in two consonants, the consonant immediately following the vowel is lengthened, as in skips [sk$_j$ɩf·s] 'ship' (gen.sg.) and báls [baul·s] 'fire' (gen.sg.). In disyllables, the stressed vowel is lengthened if it ends the syllable, as in hö$fuð [hö:vYð] 'head.' The $ here is not part of the spelling but marks the syllable division. If the syllable ends in one consonant, that consonant is lengthened, as in haf$ði [hav·ðɩ] 'had.' These patterns are schematized in (6).

(6) a. V:(C)# c. V:$ (# = word boundary;
b. VC:C# d. VC:$ $ = syllable boundary)

If we reduce (6a and b) to (6c and d) by ignoring the word-final consonant, we obtain the generalization that the last segment in the syllable is long. I will formalize the notion of ignoring the word-final consonant by adopting Kiparsky's proposal [13] that in Icelandic the last consonant in a word is extrametrical, that is, is not visible to rules applying to metrical structure. This constraint is based on work by Hayes [10] and Harris [9] which shows that a unit at the edge of a constituent may be ignored in prosodic systems such as stress and syllable structure. Consonant extrametricality in Icelandic is specified as in (7).

(7) C → [+extrametrical]/ ___#

The extrametrical consonant is adjoined to the syllable after metrical rules have applied.

Stressed syllables ending in consonant clusters (excluding extrametrical consonants) are transcribed by Einarsson [6] with no length either on the vowel or on the postvocalic consonants, as in (8).

(8) a.	sigl$di	[sɩɣlg̊ɩ]	'sailed'
b.	vins$tri	[vɩnstrɩ]	'left'
c.	efl$di	[ɛvlg̊ɩ]	'strengthened'
d.	eflt	[ɛfl̥t]	'strengthened'(p.p.)
e.	vasks	[vasks]	'sink' (gen.sg.)

These examples show that a sequence of two consonants in the Rhyme is sufficient to make the syllable long (disregarding extrametrical consonants in (d) and (e)). Comparing (6) and (8) and abstracting from extrametrical consonants, we represent the Rhyme of a long (stressed) syllable as in (9).

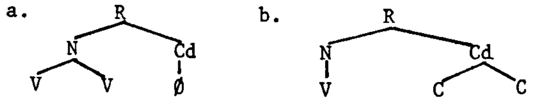
(9) a. V: b. VC: c. VC_1C_2

Adopting the terminology of autosegmental phonology, in which feature complexes are linked to timing slots in a CV skeleton, and assuming that geminates are linked to two timing slots, we summarize (9) as:

(10) A long (stressed) syllable in Icelandic is one that has either two V slots or two C slots (but not both) at the end of the Rhyme.

(10) is diagrammed in (11).

(11)



Note how closely (10) and (11) approximate the traditional definition of a long syllable in Icelandic as one with a long vowel or a long consonant in complementary distribution.

SYLLABIFICATION

I assume that syllable divisions in Icelandic are governed by the conventions in (12)-(14).

(12) Syllable onsets conform to the sonority hierarchy in (i) (from Kiparsky [12]).

Vowels-	Glides-	r-	l-	Nasals-	Fricatives-	Stops
1	2	3	4	5	6	7

(i) Vowels-Glides-r-l-Nasals-Fricatives-Stops

Numbers 1 to 7 are in order of decreasing sonority. Segments in an Onset must be in order of increasing sonority. An Onset may not contain two segments of the same sonority. Violations of (12) may occur word-initially, as in the s + stop or fricative in stiga 'to ascend,' spara 'to save,' skattur 'tax,' svartur 'black.'

(13) Maximize the Coda of a syllable. However, clusters consisting of a voiceless stop (p,t,k) or s followed by a segment of sonority level 3 or higher (glide j or r), or a voiceless stop followed by v are in the onset.

(14) A non-null onset is preferred.

Rules (13) and (14) interact to give the syllable divisions hestur, hafði, and sigl$di. If there is only one intervocalic consonant, it belongs to the Onset of the next syllable, as in ha$fa, showing that (14) has priority over (13). (13) applies crucially in of$run [of·rYn], [ov·rYn] 'lifting' and el$ja [ɛl·ja] 'skill,' where (12) would allow both intervocalic consonants in the Onset.

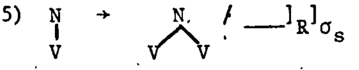
The exception statement in (13) accommodates the well-known cases where a stressed vowel in a disyllabic word is long preceding a sequence of p,t,k, or s followed by j,v, or r. These sequences regularly form onsets, as in le$pja 'to lap up,' snu$pra 'to rebuke,' vi$tja 'to visit,' ve$kja 'to awaken,' vö$kva 'to water,' ti$tra 'to shiver,' a$krar 'fields,' E$sja (name of a mountain), and há$sra 'hoarse' (gen.pl.). The first vowel is invariably long in these examples, e.g. [lɛ:p^{h}ja], [snY:p^{h}ra]. Onsets in pv, which are not attested, seem to be an accidental gap. An onset beginning with sv should not occur, since it violates the sonority hierarchy (12i), s and v being both of sonority level 6. We find that this is in fact the correct prediction. The standard example of an sv onset is tvisvar [t^{h}vɩ:svar] 'twice' [6]. However, Orešnik and Pétursson [17] point out that this is really a compound consisting of tvi + svar (cf. þrisvar 'thrice'), whose syllabification before sv is due to the presence of a word boundary before svar. As evidence that sv is not an Onset, they cite the following examples, which have a short vowel before sv: hösvir [hös·vɩr̥] 'wolf,' hösvast [hös·vast] 'potter about,' hösvan [hös·van] 'grey'(acc.sg.m.). By eliminating sv as a possible Onset, we avoid Garnes' claim [8] that v is a glide, which is meant to make sv conform to the sonority hierarchy.

The exception statement in (13) results in the syllabification vins$tri 'left' and aum$kva 'to pity.' Einarsson's transcriptions [6] of these as [vɩnstrɩ] and [öym·k^{h}va] suggest that this is the right analysis.

LENGTHENING

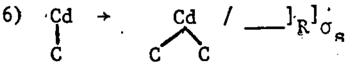
I propose (15) and (16) to derive (11).

(15)



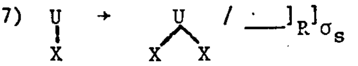
(σ = Syllable, s = strong.)

(16)



(15) and (16) are combined as (17).

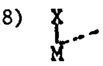
(17)



(U = Nucleus or Coda; X = V or C)

I assume with Kiparsky [13] and others that vowel length is not distinctive in Icelandic. I represent this by assigning only one V slot to the underlying Nucleus. (17) inserts a V slot in the Nucleus of a stressed syllable with an empty Coda, and inserts a C slot into the Coda of a stressed syllable with a Coda containing just one C slot. (17) applies after the final metrical structure of the word is determined, assuming Árnason's analysis [2] of stressed syllables as strong. (17) applies postlexically, after operations that affect syllable structure, such as inflectional and derivational suffixation, attachment of clitics, and phonological rules such as u-Epenthesis, j-Deletion, and Syncope, as proposed by Kiparsky [13]. I assume the constraints on autosegmental representations given in Pulleyblank [18], namely that features are associated to timing units from left-to-right in a one-to-one relation, and that association lines do not cross. I also assume with Pulleyblank that spreading is not automatic. In Icelandic, spreading is rightward, as specified in (18).

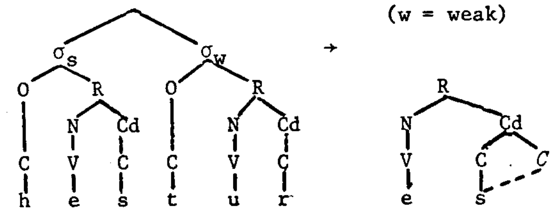
(18)



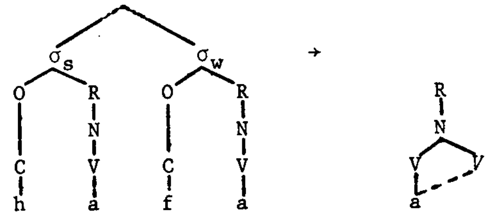
(X = a skeletal slot V or C; M = a melody unit.)

Applying Lengthening (17) and Spreading (18), I derive hestur and hafa in (19) and (20).

(19) (w = weak)

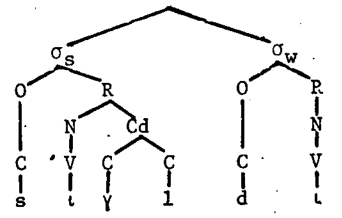


(20)



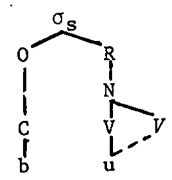
The structural change in (19), (20), and the derivations below shows only the affected Rhyme of the first syllable. Inserted slots are in italics. Lengthening (17) does not apply to sigldi (21), so there is no change.

(21)

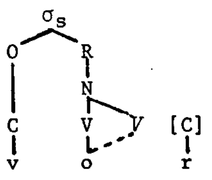


The monosyllables bú, vask, and vor [vɔ:r] 'spring' are derived in (22)-(24). Extrametrical consonants are enclosed in brackets; inserted V or C is in italics. I am simplifying Árnason's representations by omitting the weak empty syllable that completes the foot of a monosyllable.

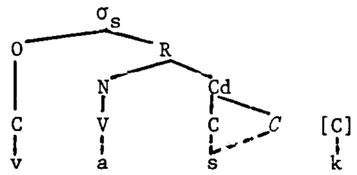
(22)



(23)

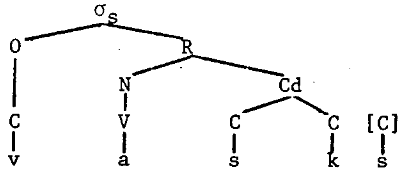


(24)



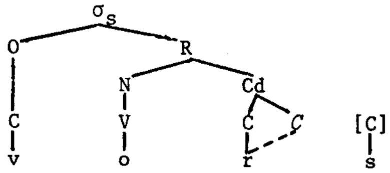
The last monosyllabic type is vasks (25). Here, Lengthening (17) does not apply.

(25)



This analysis needs no additional rules to account for the alternation in length of the s in vask [vas·k] and vasks [vasks], which follows from the application of Lengthening (17) to the input structures in (24) and (25). The alternation in vowel length shown in vor [vɔ:r] and vors [vɔr̥·s] 'spring' (gen.sg.) is likewise handled by Lengthening (17), which applies to vors in (26) to derive the short-vowel form. Compare (23), in which (17) derives a long vowel.

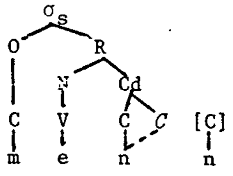
(26)



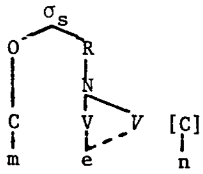
This alternation in vowel length occurs regularly when a consonantal suffix is added to a monosyllable ending in a single consonant, such as skip [skjɩ̥:pʰ] 'ship' and skips [skjɩf·s] (gen.sg.); bát [bau:tʰ] 'boat' (acc.sg.) and báts [baus:] 'boat' (gen.sg.). My analysis accounts for the alternation in vowel length, as well as the alternation in consonant length, by the general lengthening rule (17), without needing recourse to the additional shortening rule needed in Anderson's analysis [1]. (Consonant assimilations in these examples are due to other rules.)

Finally, the minimal pair menn and men (1a, 1b) are derived in (27) and (28).

(27) menn 'men'



(28) men 'necklace'



The final nasal cluster in menn (27) is reduced by later rules to obtain the approximately equal duration found by Garnès [7] for menn and men.

To summarize, the rule of Lengthening (17), with the syllabification conventions (12)-(14), accounts for Icelandic syllable quantity in a simpler and more empirically adequate way than does the analysis proposed by Anderson [1], which appeals to additional syllable restructuring and shortening rules.

ACKNOWLEDGEMENTS

I would like to thank John T. Jensen for his helpful criticisms and suggestions. All errors are my own.

REFERENCES

[1] S. Anderson, "A metrical interpretation of some traditional claims about quantity and stress," in M. Aronoff, R. Oehrle, eds., *Language Sound Structure*, The MIT Press, 1984, 83-106.

[2] K. Árnason, "Icelandic word stress and metrical phonology," *Studia Linguistica* 39, 93-129, 1985

[3] K. Árnason, "The segmental and suprasegmental status of preaspiration in Modern Icelandic," *Nordic Journal of Linguistics* 9, 1-23, 1986.

[4] H. Benediktsson, "The non-uniqueness of phonemic solutions: quantity and stress in Icelandic," *Phonetica* 10, 133-153, 1963.

[5] S. Einarsson, *Beiträge zur Phonetik der isländischen Sprache*, Brøggers, Oslo, 1927.

[6]. S. Einarsson, *Icelandic*, Johns Hopkins Press, 1945.

[7] S. Garnes, "Suprasegmental aspects of Icelandic vowel quality," *Working Papers in Linguistics 17*, Ohio State University, 144-159, 1974.

[8] S. Garnes, "Perception, production and language change," *Chicago Linguistic Society Parasession on Functionalism*, 156-169, 1975.

[9] J. Harris, *Syllable Structure and Stress in Spanish: a Non-Linear Analysis*, MIT Press 1983.

[10] B. Hayes, "Extrametricality and English stress," *Linguistic Inquiry* 13, 227-276, 1982.

[11] E. Haugen, "The phonemics of Modern Icelandic," *Language* 34, 55-88, 1958.

[12] P. Kiparsky, "Metrical structure assignment is cyclic," *Linguistic Inquiry* 10, 421-441, 1979.

[13] P. Kiparsky, "On the lexical phonology of Icelandic," in C. Elert, I. Johansson, E. Strangert, eds., *Nordic Prosody III*, Almqvist & Wiksell, 135-164, 1984.

[14] I. Lehiste, "The function of quantity in Finnish and Estonian," *Language* 41, 447-456, 1965.

[15] A.S. Liberman, "Stress in Icelandic," *Phonetica* 31, 125-143, 1975.

[16] J. Ófeigsson, "Træk of moderne islandsk Lydlære," in S. Blöndal, ed., Islensk-dönsk orðabók, Gutenberg, Reykjavík, 1920, XIV-XXVII.

[17] J. Orešnik, M. Pétursson, "Quantity in Modern Icelandic," *Arkiv* 92, 155-171, 1977.

[18] D. Pulleyblank, *Tone in Lexical Phonology*, Reidel 1986.

[19] H. Thráinsson, "On the phonology of Icelandic preaspiration," *Nordic Journal of Linguistics* 1, 3-54, 1978.

IS GERMAN STRESS-TIMED? A STUDY ON VOWEL COMPRESSION

BERND POMPINO-MARSCHALL WOLFGANG GROSSER KARL HUBMAYER WILFRIED WIEDEN

Institut für Phonetik und
Sprachliche Kommunikation
Universität München, FRG

Institut für Anglistik und
Amerikanistik
Universität Salzburg, Austria

ABSTRACT

Vowel and syllable compression due to syllabic composition of stress feet is shown to be relativlely weak in German. The effect rather works at the word level as proposed in the model of Lindblom & Rapp [5].

INTRODUCTION

The languages of the world can be divided into different types depending on what units tend to be equally spaced in the time course of an utterance [1, 6]: "stress-timed" if this unit is the stress foot, "syllable-" or "mora-timed" if these respective units have a tendency towards isochrony. Although in quite a number of experimental investigation no clear isochrony could be found, there are several effects that can differentiate between the different types of languages [3]

Vowel and syllable compression due to the syllabic composition of rhythmic feet as reported for English can be taken as evidence for stress-timing. As German also is considered to be stress-timed, we constructed an experiment parallel to one of Fowler's ([2]; exp. 7), in which she demonstrated changes in the duration of the stressed vowel due to this factor (working within and across word boundaries) in sets of sentences like the following (relevant stress feet marked by underlining):

The fact started the argument
The factor started the argument
The fact restarted the argument
The factor restarted the argument
The factory started the argument
The fact has restarted the argument.

METHOD

Analogously, in our German material we varied the syllabic composition of the testword and the foot by introducing different prefixes and suffixes ("'Trakt", "'Traktor", "Ver'trackte", "Ver'trackteste"; stress position marked by an apostrophe) as well as two different verbs ("'gab" vs. "er'gab") the testword being used both in utterance-initial and utterance-final position (relevant stress feet marked by underlining):

Der Trakt gab den Ausschlag
Der Trakt ergab den Ausschlag
Der Traktor gab den Ausschlag
Der Traktor ergab den Ausschlag
Der Vertrackte gab den Ausschlag
Der Vertrackte ergab den Ausschlag
Der Vertrackteste gab den Ausschlag
Der Vertrackteste ergab den Ausschlag

Den Ausschlag gab der Trakt
Den Ausschlag ergab der Trakt
Den Ausschlag gab der Traktor
Den Ausschlag ergab der Traktor
Den Ausschlag gab der Vertrackte
Den Ausschlag ergab der Vertrackte
Den Ausschlag gab der Vertrackteste
Den Ausschlag ergab der Vertrackteste

The sentences were uttered twice in randomized order at a - individually chosen - normal rate of speech by five native German speakers (middle bavarian). The durations of the vowel /a/, the syllable /trak(t)/, the rhythmic feet and the entire utterances were measured on the oscillogram using an inkwriter output at a paper speed of 100mm/sec.

RESULTS

The results are show in the following table and, for the vowel measurements only, in Fig. 1.

In contrast to the English data the analyses of variance only showed a weak tendency of German towards stress-timing. In the following, the different effects are discussed individually.

Stress Foot Duration

Two-factorial analysis of variance shows a significant effect of the test-item ($F(7,144) = 32.7$; $p < .001$), of position ($F(1,144) = 15.4$; $p < .001$), and a significant interaction ($F(7,144) = 4.73$; $p < .001$) on the duration of the stress foot. The simple main effects show a tendency for compression only in those items where the syllabic variation takes place within the testword (the most complex one, "Vertrackteste", is not compressed). For the sentence-initial stress feet we get the following order in duration (shortest first; the items that are not significantly different in one line; $p < .05$):

Trakt, Vertrackte, Traktor
Trakt er_
Vertrackte er_, Vertrackteste, Traktor er_
Vertrackteste er_

Parallelly, in final position only "Vertrackteste" is significantly longer than the other items. As to be expected the duration of the stress foot correlates with the duration of the entire utterance (initial $r = .829$; final $r = .683$; in both cases $p < .001$), weaker in the final stress feet ($p < .05$), because the variation in the verb is independent of the stress foot.

Syllable Duration

Syllable duration also shows no effect across word boundaries. There is a significant effect of the testword in initial and final position ($Fs(2,54) = 4.52, 3.66$; $p < .05$): only "Vertrackteste" has shorter syllable durations.

Vowel Duration

Vowel duration (see Fig.1) is not affected by variation of stress foot duration beyond the boundaries of the testword either. There is a significant testword effect in initial and final feet ($Fs(3,72) = 7.43, 7.32$; $p < .001$), but due only to the vowel always being significantly longer for the testword "Trakt". Interestingly, in general the vowel duration was longer for the initial items ($F(1,144) = 4.$; $p < .05$).

DISCUSSION

In general, our data only show a weak compression effect, favouring the model of Lindblom & Rapp [5], where this effect is assumed to work at the word level.

Table I:
Mean dura+ions (and standard deviations - in brackets) with different testwords in different positions and contexts

	duration of vowel	syllable	foot	utterance
testitem				
"Trakt"				
_ gab	131.5	389.5	398.	1484.
initial	(17.6)	(54.8)	(46.2)	(178.9)
_ ergab	123.5	385.	476.5	1537.
	(13.1)	(38.6)	(69.3)	(177.4)
_ gab	128.	463.5	463.5	1436.
final	(20.6)	(41.3)	(41.3)	(187.2)
_ ergab	121.	484.	484.	1515.5
	(18.4)	(56.2)	(56.2)	(230.8)
"Traktor"				
_ gab	111.	298.5	447.5	1533.5
initial	(10.2)	(31.4)	(54.1)	(216.2)
_ ergab	115.	297.5	578.	1646.5
	(17.6)	(39.2)	(122.8)	(254.7)
_ gab	106.5	316.	565.5	1493.5
final	(13.3)	(48.3)	(52.4)	(182.5)
_ ergab	107.5	296.	544.	1602.5
	(15.7)	(40.9)	(62.9)	(196.6)
"Vertrackte"				
_ gab	112.	292.	424.5	1659.5
initial	(14.4)	(31.4)	(37.5)	(207.6)
_ ergab	110.	289.5	534.	1745.
	(15.8)	(44.3)	(77.3)	(243.4)
_ gab	104.5	305.	517.	1594.5
final	(12.6)	(41.8)	(45.2)	(199.6)
_ ergab	107.	304.	519.5	1699.
	(17.7)	(43.2)	(41.4)	(232.6)
"Vertrackteste"				
_ gab	106.	264.5	572.5	1780.5
initial	(11.3)	(34.4)	(62.8)	(224.1)
_ ergab	110.	268.	685.5	1878.5
	(11.5)	(26.7)	(103.8)	(271.2)
_ gab	103.	276.	678.5	1780.5
final	(18.)	(37.1)	(37.3)	(193.7)
_ ergab	103.	274.5	663.	1812.
	(10.3)	(29.5)	(46.2)	(202.1)

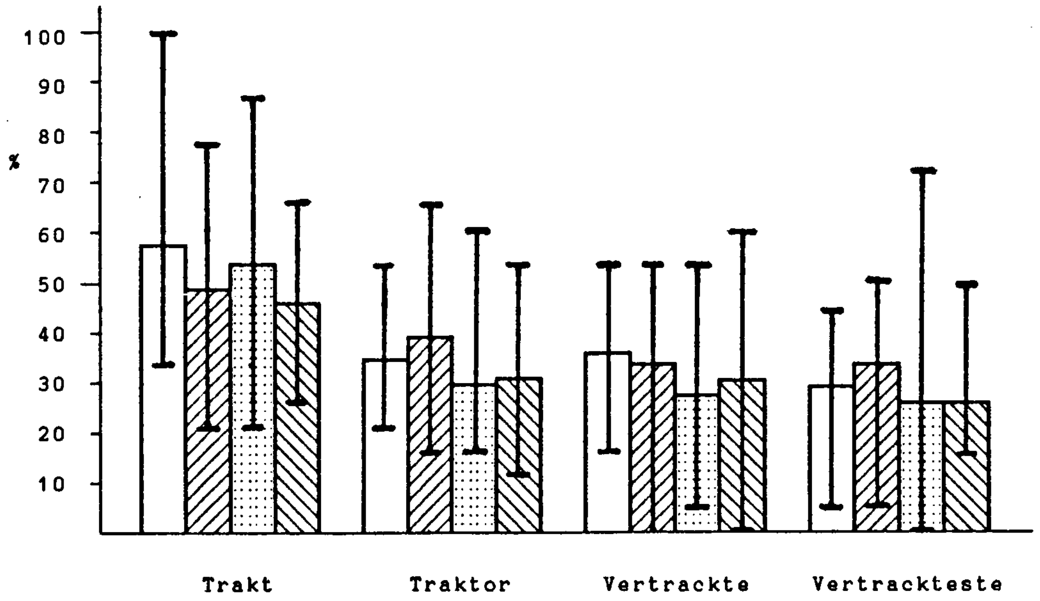


Fig.1: Mean vowel duration (and range) in percent total variation in vowel duration (100% = 170 msec; 0% = 85 msec) in the different testwords in different positions and contexts (open: initial _gab; upward hatch: initial _ergab; dotted: final _gab; downward hatch: final _ergab)

As Huggins [4] and Fowler [2] report that the compression effect is only seen at relatively fast rates of speech we reanalyzed our results, omitting the data of the one subject who produced the utterances at a noticeable slower rate of speech than the others. This reanalysis however showed exactly the same effects as before.

REFERENCES

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

[2] Fowler, C.A. 1977, Timing Control in Speech Production (unpublished doctoral dissertation, Bloomington, Indiana University Linguistics Club).

[3] Hoequist, C. 1983, Syllable duration in stress-, syllable- and mora-timed languages. Phonetica 40, 203-237.

[4] Huggins, W.A.F.1975, On isochrony and syntax. In: Fant, G./Tatham, M.A.A. (eds.), Auditory Analysis and Perception of Speech (London, Academic Press), 455-464.

[5] Lindblom, B./Rapp, K. 1973, Some temporal regularities of spoken Swedish. In: Papers from the Institute of Linguistics, University of Stockholm (PILUS) 21, 1-59.

[6] Pike, K.L. 1946, The Intonation of American English (AnnArbor, University of Michigan Press).

SYNTHESIS OF LOGICAL ACCENTUATION IN DIFFERENT LANGUAGES

Antti Sovijärvi & Reijo Aulanko

Department of Phonetics, University of Helsinki,
Vironkatu 1, SF-00170 Helsinki, Finland

ABSTRACT

The phonetic realization of logical emphasis in different languages (Finnish, Hungarian, Estonian, Swedish spoken in Finland, German and Italian) is studied with speech synthesis by creating short sentences where different words in different sentence positions are accented one by one, using the possibilities of our method for changing the intensity, fundamental frequency, and duration parameters. In our system, four quantity degrees can currently be produced with the prosodic rules, and the general pitch level can be lowered or raised from the neutral level, in addition to slowly rising or falling pitch contours and several types of local modifications in pitch. The synthesis rules for logical accentuation in these languages are presented, and the differences among them are considered.

INTRODUCTION

In a normal, neutral utterance there is generally thought to be one nuclear point, which has the main sentence stress. We use the term **logical accentuation** for utterances where one constituent that expresses a paradigmatic opposition with some implicit alternative(s) gets a special, additional stress. Other terms often used for this kind of accent are e.g. contrastive sentence stress and emphatic accentuation [4, 5]. A speaker has at his/her disposal certain phonetic means for actualizing the necessary logical contrast in the acoustic speech signal. The phenomenon is generally assumed to be universal in its phonetic realization, but different languages may, nevertheless, use different phonetic systems for the realization of this kind of logical accentuation. For instance, the quantity of a logically accented syllable is often lengthened, in some languages especially the vowel and in Finnish the syllable-final consonant.

Our task is to find out (for a number of languages) the pitch and stress contours (and their changes) of logically accented words in short sentences and to formulate these results as synthesis rules. One aim of the project is to improve the quality of synthetic speech by making the prosodic rules correspond more closely to the prosody of natural speech. The synthesis rules for producing logical accentuation could be implemented in various rule synthesis programs.

METHOD

Recordings

As a basis for synthesizing accented variants of a sentence, we have recorded examples of logical accentuation in six different languages: Finnish (speaker M.L.), Hungarian (I.S.), Estonian (A.K.), Swedish spoken in Finland (L.N.), German (M.R.), and Italian (G.G.). The affirmative test sentence (see the Appendix) was semantically identical in all languages, meaning approximately **'Quite soon, I shall be going to a conference abroad.'** The first variant was a sentence with neutral sentence stress distribution. In the other variants a word or word pair was logically accented. We have analyzed the prosody of the recorded sentences and then synthesized them. The (male) speakers were all barytones, with the exception of the Hungarian who was a bass and the Italian who was a tenor. The synthesis, however, is realized in the pitch range of a bass in all languages, and the pitch movements found in the infromants' utterances have been adapted for the synthesis accordingly.

Synthesis equipment

In our synthesis project we have been using the speech synthesizer OVE IIIb controlled by an HP 21 MX real-time computer [11]. Developed starting out from Sovijärvi's **beat phase theory** for word stress [7], our system is based on the use of **four-phase diphones** ("keys") stored in the computer's memory [8]. In our system each diphone has four matrix rows, and each matrix row contains 16 parameter values controlling the OVE IIIb synthesizer. Thus, each diphone with its four phases has a total of 64 parameter values. The rule synthesis system was designed primarily for synthesizing **Hungarian**, and the whole 'library' contains now of about 1330 keys, 80 of which represent sound combinations specially constructed for this paper, departing from the original 'library'.

In the diphones, the fundamental frequency (F0) of each unstressed and unmodified vowel varies within a range of one semitone (82-87-82 Hz = E-F-E). The contours of A0 (= amplitude of the first harmonic) are typically different for stressed and unstressed vowels [11], and they have been imitated in the sentence examples of all the six languages. In consonants the values of the parameters A0, AH, AN, FN, K0, K1 and K2 depend on the essential manner of articulation. The value of F0 is constantly 82 Hz in all unstressed, unmodified and unvoiced consonants [9].

Synthesizing Intonational Variants

For regulating the optional variations in intonation and/or stress contours we use special prosodic control symbols [9]. The slightly rising-falling basic pattern of the F0 contour of a vowel is always preserved, regardless of the application of any prosodic symbols. The prosodic control symbols are the following [10, 12]:

An overall **raised F0 level** can be produced with the symbols # (5 semitones) and ^ (2 semitones) in a sequence delimited by brackets (<>).

An overall **lowering of the F0 level** (by 2 semitones) is produced with the symbol =.

The minus sign (-) produces a **falling intonation:** F0 falls until the slash (/) or the first pause sign (_) gradually during the three next sounds (1st sound 2 semitones lower, 2nd sound 3 semitones lower, 3rd and subsequent sounds 4 semitones lower).

The plus sign (+) causes a **rising progredient intonation:** F0 increases until the slash or the first pause sign by 1 semitone for each sound until the fifth sound. In subsequent sounds F0 remains at the level of the fifth sound. (This symbol is not represented in the examples of this paper.)

A special **interrogative intonation** typical of Hungarian interrogative sentences that presuppose a *yes-no* answer and contain more than two syllables, is produced with the question mark (?): F0 rises by 7 semitones on one syllable. (This symbol is not used in the examples of this paper.)

The symbols " and ' are used for producing the **stronger and weaker degrees of stress.** The former symbol produces a 10% increase in duration, an increase in F0 by 4 semitones, and a rise in A0 by the dB values 3, 8, 6, and 2 in the successive subphases of the stressed vowel; furthermore, in the vowel of the following syllable, F0 increases by 2 semitones, and, in the consonant(s) appearing between the syllables, by 3 semitones above the level of 82 Hz. The latter symbol (i.e. weaker stress) produces an increase in F0 by 2 semitones and a slight rise in A0 by 0, 4, 3, and 1 dB in the successive subphases of the vowel after the symbol. That is, these two symbols are used to vary both the musical and dynamic aspects of intonation, whereas all the symbols mentioned previously only deal with the musical aspect.

The colon (:) is used for **lengthening** a sound: the duration parameter (DR) is increased by 150% in the third phase of a vowel and in the second phase of a consonant. An **overlong** degree of a sound can be produced by a double colon (::), which causes the DR to be increased by 300% in the corresponding phases.

The semicolon (;) is used for **shortening** a sound: the DR of the third phase of a vowel is reduced to 30% of the original duration.

The comma (,) is used for shortening a sound to a **half-long** degree: the duration of the third phase of a vowel is shortened to 70% of the original.

The **rate of speech** is controlled by a special coefficient placed before the input sequence. (In all the sentence examples of this paper the speed coefficient was set at 38.)

In order to investigate the quality of intonation in our four-phase synthesis system with material from the six languages analyzed, we constructed accentual variants of a sentence on the basis of the recordings described above, and experimented with several ways of synthesizing their intonation and stress contours. In our synthesis system we can change and vary very quickly the pitch, intensity, and/or duration of the accented words (or word pairs) in the written string, which has, at first, rather arbitrary symbols of prosody. Examples are listed in the Appendix in three different manners: (a) the normal orthographic form of the test sentence as translated into the six languages, (italics are used to denote the words with logical emphasis) (b) the input strings for synthesizing these sentences (with the diphone names and all the necessary prosodic control symbols), (c) the (highest) pitch levels of syllable nuclei displayed as semitones (zero meaning the rest level, negative numbers a lower pitch level, and positive numbers a higher pitch level).

RESULTS

About the Sentence Examples

Based on our completed 'library' of Hungarian speech-sounds we used in the sentence examples about the same vowel and consonant qualities [1-3, 6]. In the Finnish sentences (both neutral and logical variants), for instance, the symbols for synthesizing the short vowels [a] and [e] are written as **A2;** and **E2;**. However, for the relatively dark Finnish [s] (denoted with **S3** in the input strings) we had to construct new keys. The difference in duration between single and geminate tenuis plosives [p t k] is far greater in Finnish than it is in Hungarian, where the geminate plosives are relatively much shorter [2]. For this reason we have used the symbol :: in the word *piakkoin* 'quite soon'.

The Hungarian (and other) examples demand, according to our judgment, sometimes either the symbol ^ for the rise of two semitones or the symbol ' for the same rise and the weaker stress. Therefore we have synthesized the important words *hamarosan* 'quite soon' and *külföldi* 'abroad', changing these symbols.

The Estonian sentences presuppose the vowel quality [ə˔] in the words *sõidan* 'I travel' and *õige* 'quite' which have the corresponding marks of **O3**. The symbols **P3 T3 K3** correspond to the letters b d g.

In the sentences representing Swedish spoken in Finland we have very often used the slash (/) which acts as an end point for the pitch-lowering effect of the minus sign (-). The symbol **U2** represents the vowel [ʉ], which differs from the more [y]-like corresponding sound in Swedish spoken in Sweden.

In the German examples (for instance in the words *werde*, *einer*, and *ausländischen*) there are the special vowel symbols **E3 I3 U3** which correspond to the sounds [ə i˕ u˕], respectively, and the special symbol **R2** for the German variants of [ʀ ʁ ɐ].

In the Italian language we have to employ the half-long duration mainly in the cases of a stressed vowel and of a syllable ending in a consonant. For this reason we use the comma (,) for instance in the word *andrò* 'I shall be going': A2,NDRO2;.

The Synthetic Realization of Logical Accentuation

The logically accented nuclei have almost always the synthesized height of five semitones above the rest level [13]. In only one case in the examples, namely the Italian words *fra poco* 'quite soon', we find the higher intonation of 2 → 7 → 5 on the nucleus of the syllable *po-*; the string used to produce this is ^<FRA2;>#<P'02K02;>. The other alternatives for the peak height - 4 or 7 semitones - sound rather unnatural, according to our auditory experiments with various synthetic versions.

The difference of the highest tonal point between the usually and logically accented affirmative sentences is 3-5 semitones in all six languages. This means that the whole intonational contour in the neutral sentences is definitely narrower with a phonetic certainty in a given language than in the same sentences with any logically accented word or word pair (5-8 against 8-11 semitones, depending on the language).

We have not yet thoroughly investigated the interrogative sentences from the logical accentuation point of view. However, we have recently described [13] altogether three cases of logical accentuation, two sentences of which represent different questions (with the highest pitch of 4 or 5 semitones) and only one represents an affirmative sentence (with the corresponding height of 7 semitones).

At the present stage the controlling system does not allow the synthesis of intonational phenomena of those languages which are characterized by an alternating "winding" tonal patterning and often by its relatively variable durative dimensions. The languages of this kind include for instance American and British English, French, and Swedish spoken in Sweden. In contrast, the language type in which the intonational patterns are largely based on the fixed or nearly fixed patterning of quantity and stress in the concatenation of syllables and/or words is suitable for the synthesis system presented in this paper.

REFERENCES

[1] Bolla, K. "Magyar hangalbum (A magyar beszédhangok artikulációs és akusztikus sajátságai)." Magyar fonetikai füzetek (Hungarian Papers in Phonetics) 6, 1980, 1-167.

[2] Bolla, K. "A finn beszédhangok atlasza (A beszédhangok képzési és hangzási jellemzöi)." Magyar fonetikai füzetek (Hungarian Papers in Phonetics) 14, 1985, 1-249.

[3] Bolla, K. & Valaczkai, L. "Német beszédhangok atlasza (A beszédhangok képzési és hangzási sajátságai)." Magyar fonetikai füzetek (Hungarian Papers in Phonetics) 16, 1986, 1-210.

[4] Carlson, R. & Granström, B. "Word accent, emphatic stress, and syntax in a synthesis by rule scheme for Swedish. STL-QPSR 2-3, 1973, 31-36.

[5] Garde, P. "Contraste accentuel et contraste intonationnel." Word 23, 1967, 187-195.

[6] Olaszy, G. "A magyar beszéd leggyakoribb hangsorépitő elemeinek szerkezete és szintézise. A számitógépes beszédelőállítás néhány kérdése." Nyelvtudományi Értekezések 121. sz. Budapest: Akadémiai Kiadó, 1985.

[7] Sovijärvi, A. "Alustavia mittaushavaintoja suomen yleiskielen sanapainosta. (Vorläufige Messungsbeobachtungen über den Wortakzent der finnischen Hochsprache.)" Virittäjä 62, 1958, 351-365.

[8] Sovijärvi, A. "Examples of some synthesized Hungarian sentences." Sprache und Sprechen. Festschrift für Eberhard Zwirner zum 80. Geburtstag. Tübingen: Max Niemeyer, 1979, 113-120.

[9] Sovijärvi, A. "On the realization of Hungarian phonemic combinations and prosody in the light of experiments in speech synthesis." Congressus Quintus Internationalis Fenno-Ugristarum, Turku 20.-27.VIII.1980, Pars VI, 1981, 273-278.

[10] Sovijärvi, A. & Aulanko, R. "Unkarinkielisten puhunnosten sääntösynteesijärjestelmästä. (Introducing Speech Synthesis by Rule on Hungarian)." 11th Meeting of Finnish Phoneticians - Helsinki 1982 (ed. A. Iivonen & H. Kaskinen). Publications of the Department of Phonetics, University of Helsinki 35, 1982, 175-193.

[11] Sovijärvi, A. & Aulanko, R. "Rule synthesis of intonation and stress based on four-phase diphones." XIII Meeting of Finnish Phoneticians - Turku 1985 (ed. O. Aaltonen & T. Hulkko). Publications of the Department of Finnish and General Linguistics of the University of Turku 26, 1985, 217-231.

[12] Sovijärvi, A. & Aulanko, R. "Rule synthesis of variable intonation." Proceedings of Nordic Acoustical Meeting 20-22 August 1986 in Aalborg, Denmark (ed. H. Møller & P. Rubak), Aalborg University Press, 1986, 211-214.

[13] Sovijärvi, A. & Aulanko, R. "Die automatische Regelung der Intonations- und Betonungsgestalten beim Synthetisieren des Ungarischen." Gedächtniskolloquium für Eberhard Zwirner (ed. H. Bluhme). Publications of the Department Germaanse Filologie, Universität Antwerpen. (In press)

Appendix. Sentence examples

```
Finnish. Matkustan piakkoin erääseen ulkomaiseen kokoukseen.
=<_M'A2;TKUS3TA2;M>^<PI>A2;=<K::OI2N_>E2;RE:S3E2N_^<UL>KOMA2;I2S3E2N3=<K'OKO-UKS3E2N_>
     0-2-2           20-2            000           2000                0-2-4-6     [range: 8 semitones]
Matkustan piakkoin erääseen ulkomaiseen kokoukseen.
=<_MA2;TKUS3TA2;M>^<PI>A2;=<K::OI2N>E2;RE:S3E2N_#<UL>KO=<MA2;I2S3E2N3>KO=<KO-UKS3E2N_>
    -2-2-2          20-2           000           50-2-2              0-2-4-6       [11 st]
Matkustan piakkoin.erääseen ulkomaiseen kokoukseen.
=<_MA2;TKUS3TA2;M>#<PI>A2;=<K::OI2N_>^<E2;>=<RE:S3E2N_>^<UL>KO=<MA2;I2S3E2N3>=<KOKO-UKS3E2N_>
    -2-2-2          50-2              2-2-2            20-2-3                 -2-2-4-6   [11 st]
```

```
Hungarian. Hamarosan megyek egy külföldi összejövetelre.
^<_HA>MA=<ROS2AN>^<ME>=<G2EK_ET2>K'YL=<FQLDI>^<QS:>=<-EJQVETELRE_>
    20-2-2         2-2      -2    2-2-2      2-4-6-6-6-6        [8 st]
Hamarosan megyek egy külföldi összejövetelre.
^<_HA>MA=<ROS2AN>^<ME>=<G2EKET2>#<KYL>=<FQL-DI/Q-S:EJQVETELRE_>
    20-2-2         2-2      -2    5-2-5      -2-5-6-6-6-6       [11 st]
Hamarosan megyek egy külföldi összejövetelre.
#<_HA>MA=<ROS2ANMEG2EKET2>^<KYL>=<FQL-DI/-QS:EJQVETELRE_>
    50-2-2      -2-2 -2     2-2-5       -4-6-6-6-6-6            [11 st]
```

```
Estonian. Ma sõidan õige varsti välismaale ühele konverentsile.
_MA2;S'03I2T3-A2,N_=<03I2K3E2,>V'A2;RST-I_V'E=<LISMA2LE2;_>'YH2=<E2;LE2;K'ON:VE2;>-RE2;NCI=<LE2;_>
  0    2-2          -2-2        2-2       2-2-2-2         2-2-2         0-2-3-4-6        [8 st]
Ma sõidan õige varsti välismaale ühele konverentsile.
_MA2;S'03I2T3A2,N03I2K3E2,^<VA2;RS>TI_#<VE>^<LIS>MA2-LE2;_^<YH2>E2;LE2;^<KON:>=<-VE2;RE2;NCILE2;_>
  0    20       00          20          520-3            200           2-5-6-6-6        [11 st]
Ma sõidan õige varsti välismaale ühele konverentsile.
_MA2;S03I2T3A2,N_03I2K3E2,#<VA2;RS>^<TI_VE>LIS=<-MA2LE2;/>^<YH2>E2;LE2;^<KON:>VE2;-RE2;N=<CILE2;_>
  0   00        00          52          20-5-6           200           20-3-6-6         [11 st]
```

```
Swedish spoken in Finland. Jag skall resa ganska snart till en utländsk sammankomst.
_JA2;SKA2;R'E2SA2;GA2;NSK-A2;/SN'A2RT:IE2N_'U2:TLE2;NSKS'A2;M:A2;N3-KOMST_
  0    0   20     0-2           2    0 0  20           20-3              [5 st]
Jag skall resa ganska snart till en utländsk sammankomst.
_JA2;SKA2;R'E2^<SA2;>GA2;NSKA2;SN'A2RT:I-E2N_#<U2:T>LE2;NSK=<S'A2;M:A2;N3>-KO;MST_
  0    0   22        00         2    0 -2   50              0-2-3      [8 st]
Jag skall resa ganska snart till en utländsk sammankomst.
_JA2;SKA2;RE2S-A2;^<GA2;NS>KA2;#<SNA2RT:>IE2N_'U2:TLE2;NSK=<S'A2;M:A2;N3>-KO;MST_
  0    0  0-2       20            5    0 0  20             0-2-3       [8 st]
```

```
German. Ich werde ziemlich bald zu einer ausländischen Versammlung abreisen.
_IJ2=<VE2R2D-E3/ʒCI:MLIJ2^<PA2;L>C:UA2;I3-NE3R2_'A2;U3S-LE2;NDIS2:E3N_FE2;R2Z'A2;M-LUN3/=<A2;-PR2A2;I3ZE3N_>
 0     -2-4      00        2      0 0-3        2-3-4-4                02-3                -2-6-6    [8 st]
Ich werde ziemlich bald zu einer ausländischen Versammlung abreisen.
_IJ2V-E2R2DE3/CI:M-LIJ2/^<PA2;L>C:UA2;I3N-E3R2_#<A2;U3S>-LE2;NDIS2:E3N_=<FE2;R2Z'A2;MLUN3_-A2;PR2A2;I3ZE3N_>
 0   -2-4     0-3         2      0 0-2        5-3-4-4                  -20-2           -4-6-6  [11 st]
Ich werde ziemlich bald zu einer ausländischen Versammlung abreisen.
_IJ2=<VE2R2D-E3/CI:MLIJ2>#<PA2;L>C:UA2;I3N-E3R2_=<'A2;U3SLE2;NDIS2:E3N_FE2;R2Z'AMLUN3_-A2;PR2A2;I3ZE3N_>
 0    -2-4      -2-2       5      0 0-2          0-2-2-2             -20-2          -4-6-6      [11 st]
```

```
Italian. Fra poco andrò ad una conferenza all'estero.
_FRA2;P'02^<K02;_>A2,ND^<R02;>A2,D^<U:NA2;>K02,NFE2;R'E2,NCA2;^<A2,L:'E2,ST>-E2;R02;_
   0    22       02          0     22     0020                2    4-2-4              [8 st]
Fra poco andrò ad una conferenza all'estero.
^<_FRA2;>=<P'02>^<K02;_>A2,ND^<R02;>A2,DU:NA2;=<K02,NFE2;R'E2,N>^<CA2;>_=<A2,L:>#<E2,ST>-E2;R02;_
     2      02         02          0  00      -2-202                  -2     5-2-4  [9 st]
Fra poco andrò ad una conferenza all'estero.
^<_FRA2;>#<P'02K02;_>=<A2,NDR02;A2,D>U:NA2;=<K02,NFE2;R'E2,N>-CA2;_=<A2,L:>^<E2,ST>-E2;R02;_
     2      75        -2-2     -2   00     -2-20-3               -2     2-2-4       [11 st]
```

THE LINGUISTIC ASPECT OF MULTI-LANGUAGE SPEECH SYNTHESIS

YELENA KARNEVSKAYA

Minsk State Pedagogical Institute of Foreign Languages
Minsk, Byelorussia, USSR 220662

ABSTRACT

The present program adopts language-independent principles of modeling segmental and prosodic units of speech. Phonetic description of these units is based on experimental contrastive analyses of Russian, English, French and German phonetic characteristics and is carried out within a single normalized range of acoustic parameters.

The linguistic program includes rules for letter-to-phoneme conversion, phrasing and accent location rules, as well as algorithms for a prosodic contour choice and modification under varying contextual conditions.

INTRODUCTION

Rapid improvement of speech synthesis technology over the last two decades has resulted in the appearance of new programs permitting a wide range of user-specified applications. The general trend toward greater flexibility of synthesis systems is well seen through the growing interest in text-to-speech synthesis designs, and especially those handling a variety of languages /I/.

Multi-language systems apparently derive from programs suited to the needs of one particular language. Linguistically, this is justified by a universal, language-independent nature of phonetic categorization, which predetermines a largely universal character of speech synthesis as an analogue of natural spoken language. Thus any synthesis model will distinguish classes of phonemes and reflect coarticulation processes that sounds undergo in running speech; it is also bound to convey the polyparametric nature of speech prosody and take account of its linguistic uses relating in all languages to the communicative contents of an utterance.

Furthermore, articulatory and acoustic similarity of sound units belonging to identical classes in different languages - as a consequence of the phonological systems' typological similarities - suggests a possibility of applying a single descriptive apparatus for indentifying the phonetic features of the languages in creating the data bases for multi-language synthesis.

The above general prerequisites find ample explication in the current program which is based on a model, originally devised for the Russian speech synthesis. This model's applicability for multi-language purposes is due to the linguistically invariant principles underlying the design of its fundamental elements - portraits of phonemes and prosodemes /2/. Phonemes' acoustic portraits, in particular, incorporate a sufficient amount of parameters to convey exhaustive information about phonetically significant features of vowels and consonants of any language. The acoustic parameters, specifically, are presented as complexes of interacting targets and transitional functions whose values are determined by the unit's inherent properties, on the one hand, and the influence of its environment, on the other. Importantly, there are no constraints on a phoneme description either as regards parameter value modifications or the unit linear subsegmentation. It is clear that the portraits in question serve as a convenient tool for achieving allophonic output on the basis of phonemic input in conformity with the main principle of language units' actualization in speech.

In the same way, portraits of prosodemes are built in accordance with the assumed language-independent structure of an intonation-unit. They provide a sort of mould for embodying qualitative and quantitative properties of the selected patterns. Realization of the patterns is achieved by applying special rules for prosodic feature modifications depending on a number of previously defined variables.

However, multi-language orientation causes inevitable alterations of the original model, which essentially consist in adjusting the latter to the overall, broader program of which it becomes only a part. For the linguistic component this implies elaboration of a single classification matrix, which is optional in the sense that certain types or gradations may remain unoccupied in a concrete language, and at the level of acoustic analysis it suggests using absolute-relative characteristics, rather than purely relative, as preferable for the purposes being discussed. The main argument here is that an absolute-relative scale provides greater accuracy and precision in revealing interlanguages phonetic differences, especially in the case of typologically similar units.

Evidently, this kind of scale can be achieved only by normalizing data relating to different languages within a single range of parameter features: formant frequencies, amplitudes, pitch levels.

The idea of single acoustic space emerges in association with the processes of articulatory program shifting, commonly observed in the speech of bilinguals (multilinguals). It would seem that the conformity of the suggested approach to phenomena of natural speech givės ground for considering it valid.

I. Description of Phonetic Features

I.I. Material and Procedure

A normalized range of parameters has been obtained as a result of special contrastive studies in which Russian phonetic units were consistently compared to English, German and French ones.

Following the above assumptions we decided to have our test materials recorded by bilingual speakers in addition to having provided native speakers recordings.

The use of bilinguals, as an experimental method in a study of this kind, has got obvious advantages in that much of irrelevant acoustic variation is avoided, and search for interlanguage differences is thus facilitated. However, these strong points can hold only if the bilingual speaker's command of the second language pronunciation norms is really good, near to native, in fact, since phonetic interference otherwise pertinent may seriously invalidate the results.

With these requirements in mind test recordings were carried out by 3 bilingual speakers whose performance was assessed as normative (highly acceptable) by English, French and German native listeners, respectively.

The materials themselves were also constructed with a view of reducing, as much as possible, uncontrolled phonetic variability. The first set of utterances, for instance, were built in each pair of languages exclusively of the so-called interlanguage homophones, e.g. Klin(R.)- Clean(E.).

The words have been selected with a view of covering all permissible combinations of CV type in the languages being compared (as well as VC).

These words were grouped in three to produce nonsense utterances which were pronounced as declarative sentences of the type "John loved Mary".

I.2. Formant Characteristics

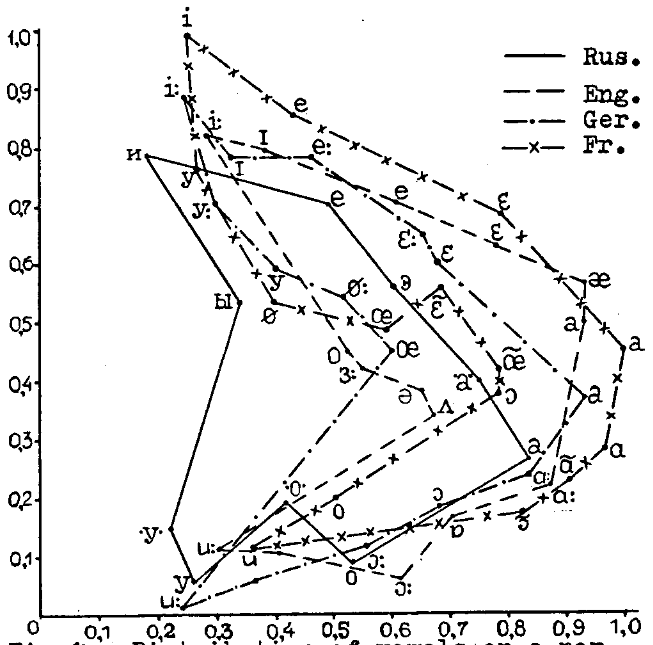


Fig.1. Distribution of vowels on a normalized FI/ FII plane.

Table 1. CV coarticulation (F II).

V / C	u			a			i		
	Eng.	Ger.	Fr.	Eng.	Ger.	Fr.	Eng	Ger.	Fr.
t	1,60	1,80	1,50	1,30	1,26	1,20	0,83	0,85	0,92
p	1,02	1,09	1,02	0,97	0,94	0,97	0,83	0,86	0,92
l	1,39	1,60	1,29	1,14	1,16	1,16	0,74	0,88	0,93

Through comparison of the distances between phonologically similar units in Russian and each of the other three languages frequency values of the first three formants were found for the vowels and sonorants, and the commonly recognized interlanguage differences in the degree of CV, VC and CC coarticulation were evaluated. Some results of this study are shown in Fig.I and Table I above.

I.3. Prosodic Characteristics

The present model is based on the concept of prosodic contour as a major operational unit of non-segmental organization of speech. In view of the complex parametric structure of the contour a componental approach has been offered here, as in most current work, both to its analysis and synthesis. Obviously enough, the components into which the overall model is split are the fundamental frequency, intensity and duration contours. These can be taken as representing the pitch, accent and timing perceptible patterns of speech only if due notice is given to their close interaction in producing (and consequently, modeling) the intonational effects ultimately aimed at. Therefore, interaction of the components must become one of the underlying principles of the model.

Realization of this principle includes several aspects. For one, connection between the contours is ensured by their compatibility due to co-extensiveness with one and the same segmental base and identical internal structure: contour of any type is constituted by one or more (typically 2-3) accent-groups with the nuclear group as an obligatory element.

While the nuclear accent-group plays a special part in the characterization of all contours, its predominant role stands out most clearly in the FO contour. The status of the latter in the overall prosodic model, in general, differs from that of the other two components and this functional inequality is another form of the contours' interaction.

Specifically, the total set of prosodic contours used in the program is determined by the number of tonal patterns that have been shown to be significantly contrasted. It needn't be argued that such a relationship is well in line with the widely accepted theories of utterance prosody.

The implications involved here are that the role of intensity and duration modifications is confined to that of accompanying features contributing to the tonal pattern ample realization. However, these modifications are only partially controlled by FO patterning. It is widely known that increases and decreases in intensity and duration of sounds are utilized by language in various other ways. Importantly, variations along these two parameters are rather more closely, by contrast with FO changes, associated with the intrinsic properties of segmental units and such aspects of prosodic organization of speech as rhythm and stress, whose functions are distinctly different from those performed by FO modulations.

Tha present program takes account of the observed peculiarities of prosodic parameters. It has been assumed that relevant information pertaining to the duration and intensity contours can be carried by segmental units, if their current temporal and dynamic characteristics are determined by special algorithms in which both segmental and prosodic variables are dealt with.

The FO model in its turn is not completely independent upon the characteristics of the segmental base upon which a contour is "superimposed", although intrinsic FO influence is outside the scope of the reported work. On the other hand, vowel length type and, to some extent, consonant manner class have been considered as capable of altering the shape of FO configurations, in English and German, in particular. Presence or absence of marginal syllables in an accent-group has also been taken into consideration. This limitation is but an exception from the general model which is independent of the given factor. Yet it cannot be ignored, e.g. in case of the Russian rising nuclear tone of the rising-falling configuration (∧): its falling element is accomplished on the post-nuclear syllables and in their absence the given configuration will take the shape of a steep wide rise.

The role of the above factors has been confirmed in a number of listening tests in which some synthetic realizations of FO contours displayed a markedly lower percentage of correct identification both of the communicative meaning of the speech unit (in terms of such dichotomies as complete vs.incomplete, interrogative vs.declarative, neutral, calm vs.categoric, expressive, etc.) and the phonetic type of the tonal pattern (in terms of pitch-change directional types and pitch-level gradations). Perceptual "deficiency" of these contours clearly stemmed from insufficient duration of their segmental bases - an effect noted in numerous earlier writings. This difficulty is overcome by supplying multiple acoustic correlates to a single functional type of contour.

There is also positional and combinatory variation of tonal patterns. The former is achieved by assigning lower values to one or more FO peaks of the contour, the pattern as such remaining unaltered due to a zonal nature of perceptible pitch categories.

The object of combinatory variation is to avoid monotony when two or more functionally identical contour types are demanded by the context. In this case the rules modulate the shape of the contour elements so that the resulting contour is slightly different both phonetically and semantically from the basic one and can be defined as its close synonym.

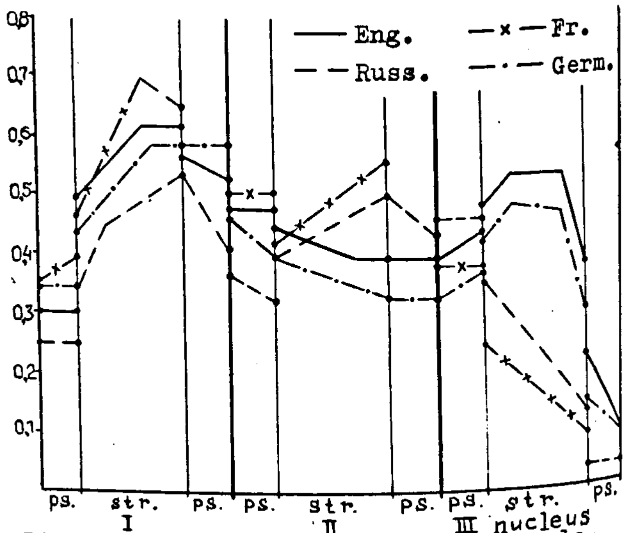


Fig.2. Phonetic realization of a declarative pitch contour.

The above rules are preceded in the program by phonetic description of the basic patterns, established as a result of experimental investigation /3/. Some peculiarities of pitch contours in the languages under study are shown in Fig.2.

A separate algorithm for a contour choice is designed on the basis of distributional tendencies and semantic properties of the contours, such as, e.g. preferable use (in English) of a falling-rising pattern in an initial parenthetical phrase, or a tendency toward using directionally similar contours in adjacent non-final syntagms, etc.

The temporal and dynamic algorithms start with establishing inventories of durational and intensity allophones, respectively, in accordance with the adopted principles of their classification. More detailed account will be given in this paper of the duration rules.

In the suggested classification all allophones (of vowels and consonants, alike) are characterised by a single set of parameters, each having several discrete gradations. As a result, allophones differ in a combination of parameter features, the number of distinctions ranging from 1 to 6. The maximal figure corresponds to the number of factors regarded as potentially relevant: type of juncture on the syllable's left and right; immediate segmental environment of the given sound; the pitch pattern (tone) of the accent-unit; degree of prominence of the given syllable. Only some of the possible feature combinations are selected, the larger part having been excluded apriori as insignificant. Each allophone in each phoneme class is assigned a coefficient which is a ratio to the phoneme intrinsic duration. The latter was identified with the duration of a sound in an initial stressed syllable of a word. In determining phonemic duration it was important to bring out quantitative peculiarities of phonemes within a class, on the one hand, and to display interclass and interlanguage differences, on the other.

Analysis of the "minimal pairs" of allophones (those differing in one parametric feature) yielded quantitative evaluation of the effect produced by each of the factors considered in the study. One of the findings here was that the ratio values changed within a fairly wide range, e.g. from 15% to 35%, or from 50% to 75%, depending on the concomitant factors. Thus, the shortening effect of a voiceless stop upon a stressed vowel was twice as high in a phrase final position as compared to an initial accent-group; the lengthening effect of prepausal position upon a stressed vowel in a monosyllable is the greatest for a falling-rising nuclear tone, and so on. The conclusion to be made is that positing coefficients of increase of sound duration under the influence of separate factors is invalid unless all the co-occurring segmental and prosodic conditions are taken into account. The content of an allophone in the suggested classification is just such a complex of co-occurring conditions, thought to be relevant for determining segment duration. It would seem that this approach captures the non-independent character of the factors significant for modifying segment duration /4/.

Prior to the choice and realization of contours is determination of an intonation-group boundary location and placement of accents inside this unit.

An attempt has been made to express the syntactical-semantic segmentation markers of utterance (previously singled out and classified for each of the languages) in terms of morphological features of constituent words, their position, environment and potential semantic weight.

CONCLUSIONS

The suggested language-independent classifications of parametric allophones and tonal contours as well as the absolute-relative methodology of phonetic description have proved applicable to multi-language synthesis. The present research has widely employed contrastive analysis of phonetic units as an indispensable stage towards speech synthesis. It must be stressed that analysis for synthesis is always analysis through synthesis as well, and this aspect is undoubtedly most interesting from the point of view of verifying the perceptual importance of the phonetic peculiarities revealed as a result of the analysis.

Further efforts are required to achieve greater formalization in the linguistic component of the program.

REFERENCES

/1/ Hertz S. "Multi-language Speech Synthesis - A Search for Synthesis Universals", J.Acoust.Soc.Am. 67, Suppl.1, 198Ø, p.s. 39

/2/ Lobanov B. The Phonemophone Text-to-Speech System. In this volume.

/3/ Karnevskaya E.B., Lobanov B.M. Modeli sinteza melodicheskogo kontura russkih i angliyskih fraz. In: ARSO-82.- Kiev, 1982.
Karnevskaya E.B. Tonalny kontur kak edinitsa prosodicheskoi organisatsii svyaznogo teksta. In: Sistemnye characteristiki ustnoi i pismennoi rechi. - Minsk, 1985

/4/ Klatt D. "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", J.Acoust.Soc.Am. 59 (5), 1976, pp. 1208 - 1221.

APPLYING THE TONETIC STRESS MARK SYSTEM TO THE SYNTHESIS OF BRITISH ENGLISH INTONATION

Briony J. Williams and Peter R Alderson

IBM UK Scientific Centre, St Clement Street, Winchester, SO23 9DR, U.K.

Abstract

The synthesis of intonation in a text-to-speech system has long been a neglected area. Recently, work by Pierrehumbert has developed a model for synthesising American English intonation which uses a string of 'pitch accents', assigned autosegmentally. On the other hand, the 'British school' of intonation analysis has developed a representation of intonation that has been used successfully in transcribing spoken (British) English and in teaching intonation to foreign learners of English.

The work reported here is an attempt to blend the two appraoches in the context of a text-to-speech synthesis system for British English. The input text contains a linguistic representation of intonation, using units such as 'nucleus' and 'head' These units are converted to a set of abstract 'target values', restricted to a scale of one to ten. These in turn are converted to frequency values by the superimposition of a declining frequency envelope, the parameters of which are dependent both on the speaker model used and on the higher-level declination currently in force. The frequency values are added to the segmental information, and the result is output as speech.

1 'British' school of intonation analysis

Most analyses of English intonation proposed by linguists may be placed in one of two major schools of thought: the 'American' and the 'British'. The 'American' approach sees pitch levels as phonemic for intonation, and pitch contours as simply the concatenation of levels. For linguists of the 'British' school, however, the pitch contour is the primary unit of analysis and there is no attempt to segment it into its constituent levels. This approach was developed partly as a pædagogical tool for the teaching of English as a foreign language, and also for the practical transcription of the intonation of real speech. An example is the work of O'Connor and Arnold [4] or Crystal [2], who split each intonational phrase or **word group** into constituent units. A word group contains one obligatory unit, the nucleus, which falls on the most prominent word of the group. Preceding accented syllables are referred to collectively as the **head**, and any unstressed syllables before these are known as the **prehead**.

1.1 Adapted 'British' system

The work reported below makes use of a model of intonation based on that of O'Connor and Arnold, with features from Crystal's analysis but differing in some respects from both. It has been formulated to avoid some inconsistencies found in the units proposed by O'Connor and Arnold, as detailed in [8]. The system as a whole closely parallels that found in the work of other 'British school' linguists. The basic units of the system are shown in Figure 1 below.

Tone-units: A major tone-unit boundary mainly occurs at a longer pause; a minor tone-unit boundary is mostly found at a shorter pause or *filled pause*, i.e. with lengthening of the final syllable of the minor tone-unit.

Accented syllables: Five types of pitch movement are recognised for accented syllables: fall, rise, fall-rise, rise-fall, and level. If the accented syllable is followed by one or more unaccented syllables, then the pitch configuration is spread over the accented syllable and the following unaccented syllables. The five accent types apply equally to the nucleus and the head, thus simplifying the analysis considerably. For O'Connor and Arnold, as for Crystal, the types of pitch pattern found in the head are phonemically distinct from those found in the nucleus. The analysis described here makes no such rigid division, thus allowing a generalisation to be stated. The accent types may be either *high* or *low* (represented by super- and subscript symbols respectively). These terms refer to the initial pitch of the accented syllable as compared to the pitch of the preceding syllable.

	Tone-unit boundaries		
Major:	||	Minor:	|
	Accented syllables		
Fall:	ˋs, ˎs	Rise:	ˊs, ˏs
Fall-rise:	ˇs, ˬs	Rise-fall:	^s, ‸s
Level:	¯s, _s		
	Unaccented syllables		
Booster:	↑s	Drop:	↓s
	Stressed:	·s	

Figure 1. Intonational units used

Unaccented syllables: Stressed but non-pitch-prominent syllables may occur at any point in the tone-unit. They are marked with a mid-high dot. Pitch-prominent but unstressed syllables are those syllables which deviate markedly from the pitch direction so far established. They may be either much higher or much lower than the immediately preceding syllable, and are marked by up-arrow and down-arrow respectively. Unstressed and non-pitch-prominent syllables form the majority of unaccented syllables, and are notationally unmarked.

1.2 Background to the model

A 'British school' system was chosen, rather than an 'American school' system, because the former type has proved its value in the transcription of real speech. Although O'Connor and Arnold originally used only carefully-constructed examples, for pædagogical purposes, the same type of system has been used successfully in the transcription of sizeable corpora of spoken English ([2], [6], and the corpus described below). The American school' type of model has not been as extensively used for this purpose. Therefore it was felt that the former type was more likely to reflect all and only the linguistically-significant pitch movements of (British) English.

2 Spoken English Corpus

The intonational model described above is being used for the prosodic analysis of a corpus of contemporary spoken British English that is currently being compiled by researchers at the University of Lancaster, U.K., and the IBM UK Scientific Centre. This involves the recording of programmes from the radio. These are non-spontaneous monologues dealing with such subjects as current affairs (both newsreading and live reporting), financial advice, Open University lectures, dramatic narrative, religious services, and general-interest lectures.

After the initial high-quality recording of a programme, a portion is transcribed prosodically using the system outlined above. The prosodic transcribing is divided between two phoneticians: Dr. Gerry Knowles of Lancaster University, and Dr. Briony Williams of the IBM UKSC. There seem to be no serious discrepancies between the two transcribers, and there is a high degree of agreement between them on the accent types and boundary locations used. To date, approximately 33,000 words have been transcribed prosodically. The finished corpus is expected to contain 50,000 words, all prosodically transcribed.

3 Synthesising from a prosodic transcription

A few sentences were chosen at random from texts included in the Spoken English Corpus, and the (manually-assigned) prosodic transcription of these sentences was used as the basis for synthesis of the intonation. The hypothesis was that the prosodic transcription, having been made by hand from the recording, was a full and sufficient description of the linguistically-relevant pitch variation in the utterance. If a version of the utterance synthesised from the prosodic transcription then proved to be essentially indistinguishable from the (resynthesised version of) the original, this would support the view that the linguistic units chosen for annotation were necessary and sufficient for the prosodic characterisation of that utterance. With this in mind, the following sentence was arbitrarily selected as an example:

Dada did not really attempt itself to offer a consistent solution. It was enough to expose the crisis in the relevance of art. However. Dada did put forward some positive proposals.

3.1 From the prosodic transcription to 'target values'

Using the (manually-assigned) prosodic transcription shown in Figure 2 as input, syllables were then assigned **target values**. These are integer values between 1 and 10, representing an abstract scale of linguistically-relevant pitch height. These target values are similar to those in [5]. Each accented syllable had a target value, while those carrying pitch glides had two or three as appropriate. In addition, each unaccented syllable at the end of an accent contour (i.e. just before the next tonetic stress mark, or before a tone-unit boundary) was given a target value.

↑¯Dada did not _really aˋttempt | itˋself | to ¯offer a con_sistent soˋlution || it was eˋnough | to ex¯pose the ˎcrisis | in the ˋrelevance of ˋart || ↓howˊever | ·Dada ¯did put _forward _some | ·positive proˋposals ||

Figure 2. Prosodic transcription made by hand from recording

The target values, under the proposed system, are assigned according to simple rules based on the accent types marked. For example, a high (superscript) fall is assigned an initial target value that is three greater than that of the immediately preceding syllable within the same minor tone-unit, while its final target value is six less than this initial value (with a minimum value of 1 and a maximum of 10). The final value applies to the end of the syllable, if the accent is monosyllabic: otherwise, it applies to the last of the following unaccented syllables, the F0 of the intervening ones being later interpolated.

3.2 From target values to Hz frequency values

These target values are then converted into frequency values in Hz. This is done using essentially the same method as in [5]: i.e. superimposing an overall pitch envelope that incorporates declination. In this case, unlike the original method used by Pierrehumbert, the baseline represents the lowest possible limit of the speaker's pitch range, and is constant. The topline, on the other hand, declines exponentially from start to end of a minor tone-unit. The topline declination is set on a global basis, by specifying its value at the beginning and end of the (first) minor tone-unit, and interpolating exponentially between those values. At the start of any following minor tone-unit within the same major tone-unit, the initial F0 value for the topline is reset, but at a point somewhat lower than that of the corresponding point in the preceding unit; and similarly by the same proportion for the value of the topline at the end of the minor tone-unit. Thus the effect is an exponential decline in topline reset values over the course of a major tone-unit. In addition, at the start of a new 'paragraph', the topline returns to its original value.

For the purposes of the present investigation, the values for the baseline, topline start, topline end, and drop in reset value of topline, were adjusted such that the closest possible match was obtained between the output Hz values for the vowels and those of the original utterance. The aim was to match the output to the original utterance in order to form an impression of the validity of the linguistic units used.

Having set the values for the overall pitch envelope as described above, the target values were then taken as specifying proportions of this overall envelope. The program superimposing the declination envelope converted each target value to a frequency value in Hz. The recorded utterance was digitised at 10 kHz using a 4.5 kHz low-pass filter. This digitised utterance was then analysed by linear predictive coding (LPC), using a filter order of 64. The excitation coefficients were then replaced by the F0 values obtained from the process described above. Each F0 value was assigned to the vowel of the syllable, at a point in time that was 25% into the vowel's duration. It was found that this gave a more natural-sounding

The output of the above processes is shown in Figure 3, where it is plotted with the F0 of the original utterance. output than if the F0 value were assigned at the very onset of the vowel, or halfway into the vowel. Once all values had been assigned, the F0 was interpolated between them.

Finally, F0 perturbations of 15 Hz were added at the boundaries between voiced and voiceless segments. This process reflected a physiologically-determined effect occurring in real speech at such boundaries. Although no attempt was made to allow for intrinsic vowel pitch and other perturbations, it was found that this one process greatly improved the naturalness of the synthesised output. The reference form of the original utterance is not the digitised version, but the version obtained by resynthesis from the LPC coefficients for the original. This was felt to be more comparable to the experimental resynthesised version, factoring out the effects of LPC resynthesis to display only the effects of alterations in the F0. In addition, the boundary between the second and third sentences was treated as a paragraph boundary, with complete resetting of the topline to its original value at the start of the third sentence. This was felt to be justified before the accented sentence adverb *however*, which was here functioning in an introductory, paragraph-initial fashion.

The match between the rule-synthesised F0 and the resynthesised original is good. To the ear, the match is even closer: a surprising discovery was that many discrepancies seen on the F0 plot in Figure 3 were not in fact perceptually salient. These discrepancies could be heard only on careful listening and in full knowledge of what to listen for, and seemed to be related to segmental micro-effects on F0 rather than to linguistically significant intonation. This suggests that attempts to match as precisely as possible to the original F0 may be unnecessary. A more useful metric is that of the **perceptual equality** of two F0 contours, as used by some Dutch workers on intonation synthesis (e.g. [7], [3]). Their notion of perceptual equality is based on linguistic and auditory indistinguishability, rather than on acoustic identity. Since no two utterances of the same sentence are ever completely identical acoustically, the notion of perceptual equality may well prove to be of value in the assessment of synthesised intonation.

4 Discussion

The investigations reported above may have implications for the way in which the synthesis of intonation is approached. An attempt has been made to use a theoretical model which expresses just those pitch movements that are linguistically significant in British English, and which has been used successfully for many years for the practical transcription and teaching of British English intonation patterns. The results so far support the view that the model chosen is able to account satisfactorily for the large-scale, linguistically-relevant features of pitch movement. If these movements are correctly specified, it is then possible to go on to consider segmental effects on F0, which affect the perceived naturalness of the output without contributing to the linguistic message.

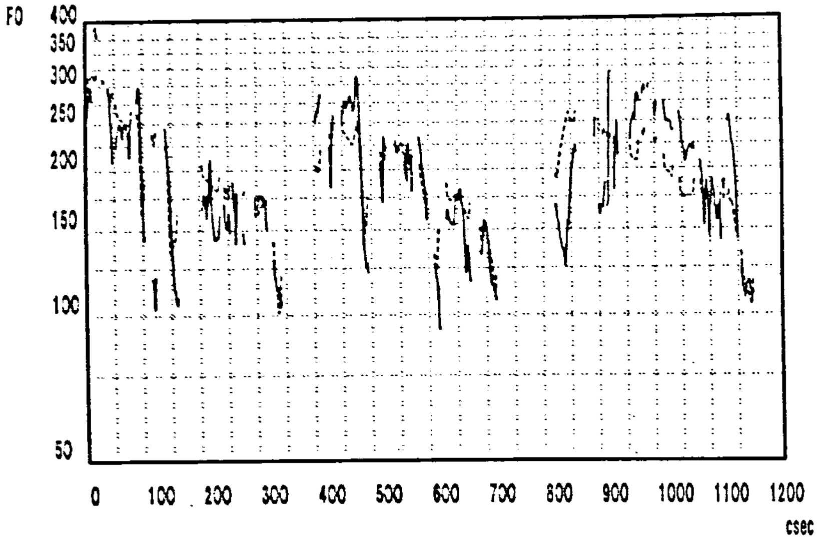


Figure 3. Original resynthesised F0 vs. F0 synthesised by rule

Solid line = F0 of resynthesised original utterance: hatched line = F0 of utterance synthesised by rule from prosodic transcription.

The assessment of intonation contours is peculiarly difficult, as it is rare for these to be definitively correct or incorrect: listeners will strive to fabricate a convincing scenario for an inappropriate intonation contour, rather than reject it out of hand. Thus it is difficult to find appropriate measures of the 'correctness' of synthesised intonation contours. As a first approximation to such a measure, we have used the F0 of the original utterance as a yardstick. However, the usefulness of this method is limited, as in no sense is the precise F0 of an original utterance to be taken as canonical. It is in this respect that the notion of perceptual equality is particularly useful. Two utterances that are perceptually equal in their intonation patterns can be said to be linguistically equivalent, carrying the same prosodic connotations. The synthesised utterances subjected to the process described in this paper seemed, on informal listening, to meet this criterion (in fact, in a few cases, the original and the rule-synthesised version were effectively indistinguishable). To establish the bounds of perceptual equality, however, more formal listening tests are required.

5 Beyond synthesis from annotated text

Having chosen a theoretical model for the representation of intonation, and having concluded that the units it provides are in fact of use in synthesising intonation, it is necessary to consider whether the model is capable of being related to other components of the grammar for the purposes of intonation synthesis from unannotated text. Bachenko et al. [1] outline a method of using the (surface) syntactic structure of an utterance to derive the prosodic representation, taking into account the syntactic constituent structure, grammatical function (head/modifier, etc.), and constituent length. In the context of a text-to-speech synthesis system, a syntactic parsing module will yield a syntactic representation giving the class of each word and the constituent structure (it is assumed that there will be no means of deriving semantic information, as the input will not be annotated in any way). The syntactic representation would be tagged with grammatical function to indicate the most likely points for intonational breaks (here interpreted as tone-unit boundaries). For Bachenko et al., there are four types of grammatical relations: these are shown below in order of strength, where the first is the most likely to cause an intonational break.

1. Sentence and adjunct: e.g. *Insert unit into correct shelf location - per detail instructions*
2. Subject and predicate: e.g. *The 48-channel module - has two di-groups*
3. Head and complement: e.g. *has - two di-groups; shows - you - how to fly your kite*
4. Head and modifier: e.g. *the echo cancelers - that are in that shelf; that are - in that shelf*

Some preliminary work has begun on specifying a prosodic representation according to criteria such as these, and the results indicate that it is indeed possible to use the type of model described above to derive intonation from syntactic structure. In this exercise, the criterion of success cannot be a match to the intonation of a particular token of that utterance, as there is no reason why the underlying prosodic representation should be the same in each case. What is required is that the intonation so derived should be at least a plausible pattern for that particular utterance, in that a listener should not need to stretch the bounds of possibility to make intonational sense of the resulting synthesised output.

At this stage, the most it is reasonable to aim for is a relatively neutral style of intonation without significant emotional colouring. Although it is debateable whether any intonation can truly be said to be 'neutral', it is a necessary idealisation in the present situation, where the relationship between syntax and prosodic structure is the least well understood aspect of intonation. In this respect the Spoken English Corpus described above is of great value, as it contains a large proportion of unemotional speech. It therefore provides data for the development of a basic intonational model which could then form the core of a more fully-specified theory of intonation that accounts also for emotional variation.

References

[1] Bachenko, J., Fitzpatrick, E. & Wright, C.E. (1986) The contribution of parsing to prosodic phrasing in an experimental text-to-speech system. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics.*

[2] Crystal, D. (1969). *Prosodic Systems and Intonation in English.* Cambridge: CUP.

[3] de Pijper, J.R. (1983). *Modelling British English Intonation.* Netherlands Phonetics Archives, Vol. III. Dordrecht: Foris Publications.

[4] O'Connor, J.D. & Arnold, G.F. (1961). *Intonation of colloquial English.* London: Longman.

[5] Pierrehumbert, J.B. (1980). The phonology and phonetics of English intonation. Unpublished Ph.D. dissertation, MIT.

[6] Svartvik, J. & Quirk, R. (1980). *A Corpus of English Conversation.* Lund Studies in English, no. 56. Lund: C.W.K. Gleerup.

[7] Willems, N. (1982). *English intonation from a Dutch point of view.* Netherlands Phonetic Archives, Vol. I. Dordrecht: Foris Publications.

[8] Williams, B.J. & Alderson, P.R. (1986). Synthesising British English Intonation using a Nuclear Tone Model. IBM UKSC Report no. 154.

THE PHONETIC BASIS OF ARTIFICIAL RUSSIAN SPEECH, ITS GENERATION BY COMPUTER AND ITS APPLICATION

Kálmán Bolla and Gábor Kiss

Department of Phonetics
Linguistics Institute of the Hungarian Academy of Sciences

Abstract

The authors describe their research experiences and results in their study (analysis and synthesis) of the phonetic structure of Russian speech in recent years. Based on their research findings, they developed a Russian language text-to-speech system, called RUSSON. Ine paper discusses some key phonetic questions related to RUSSON (letter-sound-phoneme-microelement, word stress, segmental and suprasegmental structure, palatalization-pharingalization) and describes the computer program of RUSSON.

Introduction

Production of artificial speech does not amount to a special scientific achievement. Lately, attention is focussed rather on the application of synthetic speech and on automatic speech recognition. In Hungary, the first sound and speech synthetizer systems were developed in the late seventies, early eighties as a result of research conducted at the Department of Phonetics of the Linguistics Institute of the Hungarian Academy of Sciences. Their primary aim was to aid scientific study of the sound structure of speech.

The present paper is an account of our research experiences and results accumulated in the past few years in the phonetic analysis and synthesis of Russian speech. Preliminary work and earlier results were reported in our book titled "A Conspectus of Russian Speech Sounds" published in 1981, as well as papers in the series "Hungarian Papers of Phonetics" No. 1-16. (1978--1986).

The instruments used for the analysis and synthesis of Russian speech were those available at the Departments of Phonetics of the Linguistics Institute of the Hungarian Academy of Sciences. The most important ones are as follows: a dynamic sound spectrograph, a pitch meter, a intensity meter, a four channel mingograph, a twelve channel oscillograph. The speech synthesis was done on a PDP11/34 computer and a OVE III/c formant speech synthesizer.

The authors first showed the RUSSON system to the public at an exhibition held in Moscow in 1985 to commemorate the 40th anniversary of Hungary's liberation.

RUSSON as a phonetic research aid

RUSSON was meant as a computer model of Russian phonetic processses. It provided a means to verify our analysis and to use the analysis-by-synthesis method. The synthesizing method enables us to alter any of the individual acoustic features of speech at will, to extract and analyse its physical and phonetic elements and structures, to filter out those constituents and features which have no linguistic function; to establish the language specific rules of sound linkage, the concomitance relations and compensatory ways obtaining between various constituents of sounds, the combination and variability of elements; to analyse the structural relevance of sound elements and the sound structures made up of these.

On some phonetic issues relating to RUSSON We can only touch upon some phonetic questions which relate directly to either the development of the application of RUSSON. (A more detailed version of the present paper will appear in No. 17 of Hungarian Papers in Phonetics.)

1. Writing, phonological system, sounding speech, acoustic structure, speech perception

The Russian writing system is a syllabic and morphophonemic system using the Cyrillic alphabet. One variant of our synthetic speech system produces sounding speech taking orthographic text in Cyrillic letters (including punctuation signs). This is the well-known text-to-speech system.

2. Segmental-suprasegmental sound structure

The two structures are relatively independent of each other, which means either can be extracted from the complex acoustic signal alone, or either can be produced separately.

3. Russian wordstress and temporal structure

Word stress in Russian is quantitative stress with special features of intensity and melody. The position of word stress is free varying in cases even depending on accidence.

The synthesized samples clearly suggested that lengthening always indicate stress, although in certain positions the duration of the stressed vowel (particularly in two syllable words) may be equal to, or even less than that of unstressed vowels. The reason for this is that stress is tied to the word form and is present in actual use even if unrealized by phonetic means.

4. The consonantal nature of the sound system palatalization and pharingalization

The Russian sound system is consonantal. In harmony with the consonantal character the articulatory and perceptual basis of Russian consonants is dominated by the consonants. The sound structure of Russian speech is basically determined by two factors: its duration is determined by its stress, its vocalic structure by the palatal-pharingal articulation.

5. Intonational structures, prosodemes

The text-to-speech system RUSSON uses the following matrix to produce the actual intonation forms. If our intonation experiments so require, the values of the matrix can be adjusted.

Operation of the Russian language text-to-speech computer system RUSSON

The program produces sentences of any content entered in correct Russian orthography in the following three main steps.

a) First, using a set of rules the program maps the letter sequence into a series of so-called microelements, which will ultimately form the segmental basis of artificial speech.

b) Next, on the basis of the sentence final punctuation mark the suprasegmental

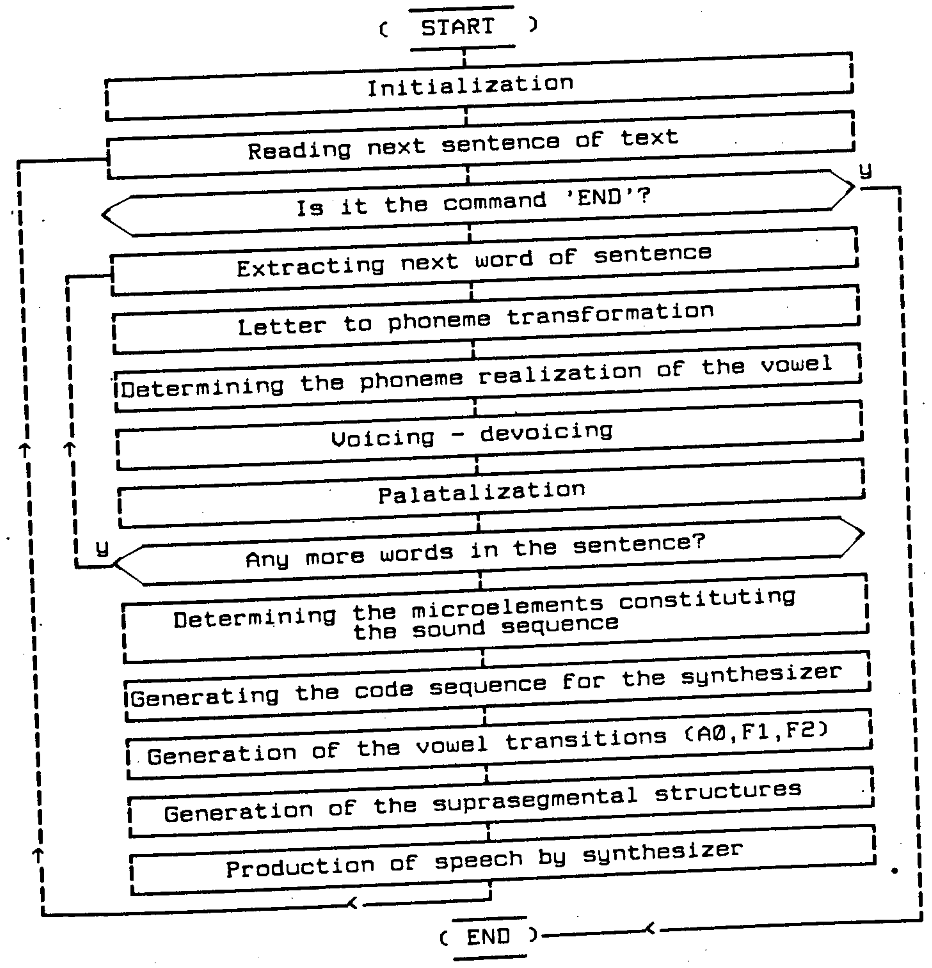

START
Initialization
Reading next sentence of text
Is it the command 'END'?
y
Extracting next word of sentence
Letter to phoneme transformation
Determining the phoneme realization of the vowel
Voicing - devoicing
Palatalization
y
Any more words in the sentence?
Determining the microelements constituting the sound sequence
Generating the code sequence for the synthesizer
Generation of the vowel transitions (A0,F1,F2)
Generation of the suprasegmental structures
Production of speech by synthesizer
END

Fig. 1. The main steps of the operation of the Russian language text-to-speech system RUSSON

structure is generated and then integrated with the segmental structure.
c) Finally, the code sequence resulting from the above two steps, which now maps the complex acoustic phenomena, is passed to the synthesizer, which will produce the sentence.

The operation of the program in more detailed steps is illustrated in the flowchart in Fig. 1.

The stock of micro elements

The control program produces the given sentence with the help of a system of rules and the inventory of microelements. The system of rules is implemented in tables and look up procedures. The stock of microelements contains the speech sounds and the pauses. Each sound is built up of 4 microelements. The RUSSON program produces the sound structure out of a possible set of 37 consonant and 35 vowel phoneme realizations. The pauses between words and sentences are generated out of 5 microelements of different length. Thus, the inventory of microelements must contain 34 * 4 + 35 * 4 = 292 elements.

The letter-to-phoneme transformation

The Russian text may consist of 31 letters as well as a soft and a hard mark. Going through the string of letters in the sentence the program selects out of the 21 consonants and 5 vowels those which correspond to the letters, simultenously carrying out any softening where requied. At this stage the program also registers word stress as well as possible sentence stress by storing the ordinal number of the stressed vowel.

Selection of vowel phoneme realizations

The program segment designed to establish the correct vowel phoneme realizations takes as input data the word to be processed and the vowel phonemes making up the word as yielded by the letter-to-phoneme transformation. They can be of the following five types: A, O, U, I, E. Taking these five vowels and their phonetic positions inside the given word the program selects one of the 35 possible vowel realizations.In defining the phonetic positions the program considers stress, pre- stress, word initial and word final positions as well as the quality of the preceding and following sound (whether it is soft or hard).

Selection of the consonant phoneme realization

The consonant phoneme number yielded by the letter-to-phoneme transformation is identical to the phoneme realization number. However, in the course of later processing the sequence of consonants may undergo change as a result of the program segments which check for voicing or palatalization.

Voicing and devoicing

The program extracts the two-member sound clusters from the words of the sentence one by one. If the cluster is made up of two consonants, both members will be checked to see if either of them belong to the exceptions. If the first member is listed as one undergoing no modification or the second member belongs to the set of consonants that do not change the preceding consonant, then the program passes on to the next cluster. When a modification is called for, it is carried out with the help of a table. Word-final consonant-consonant clusters require special treatment. First, the word-final sonorant is devoiced (if necessary) and then the preceding consonant is processed.

Тё'тя пьё'т ру"сский ча'й.

Тё'тя пьё'т ру"сский ча'й?

Сады' цвету'т весно"й.

Сады' цвету'т весно"й?

Ната'ша пое'хала нада"чу.

Ната'ша пое'хала нада"чу?

Execution of palatalization

Here again, the program first extracts two element sound clusters.If they are both consonants then the combinations not undergoing palatalization are filtered out. Where required, palatalization is executed by changing the number of the initial member of the cluster.

Defining the microelements

The suprasegmental structure corresponding to the sounds defined earlier is based on microelements.Four microelements are assigned to every phoneme realization. However, the program does not make use of all the four microelements in every instance. There are cases when only the second, third and fourth element is used. The function of the first microelement is to ensure a smooth, even onset of a sonorant sound.

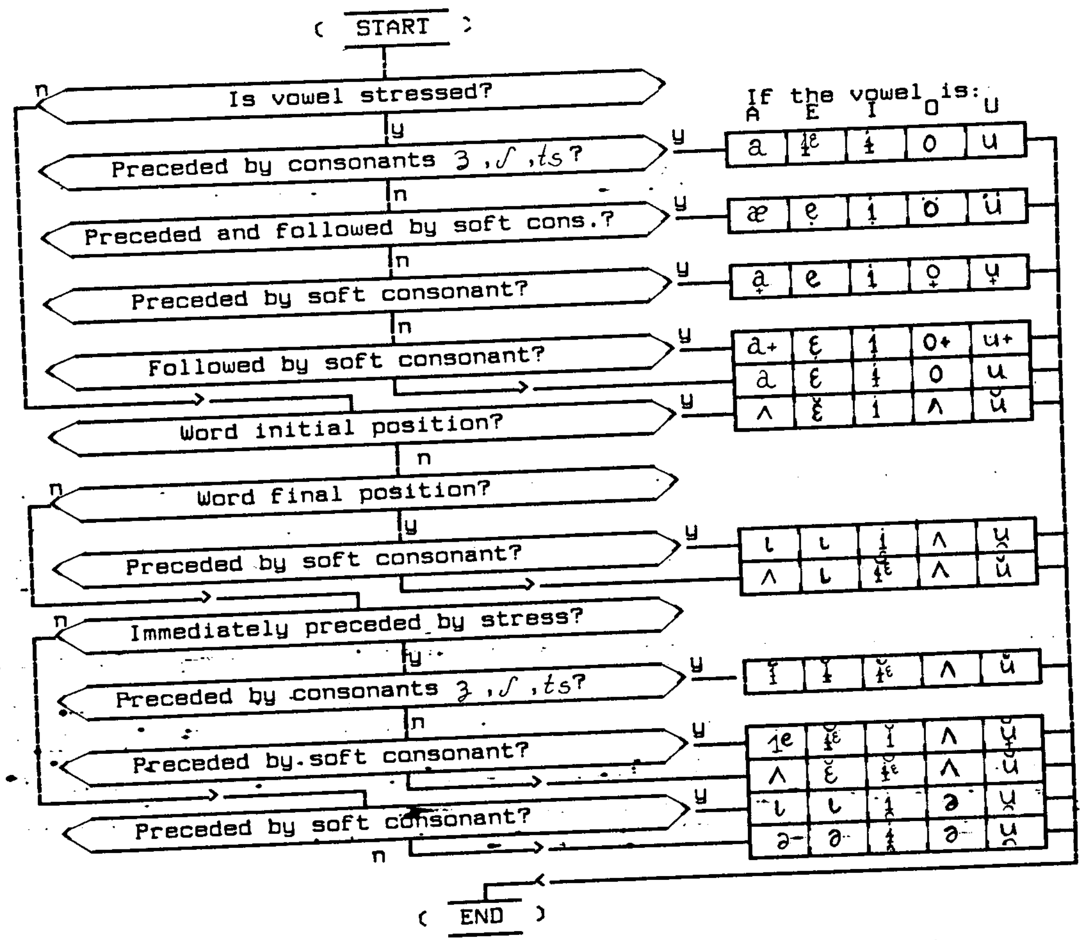

START
Is vowel stressed?
If the vowel is:
A E I O U
Preceded by consonants ʒ , ʃ , ts?
Preceded and followed by soft cons.?
Preceded by soft consonant?
Followed by soft consonant?
Word initial position?
Word final position?
Preceded by soft consonant?
Immediately preceded by stress?
Preceded by consonants ʒ , ʃ , ts?
Preceded by soft consonant?
Preceded by soft consonant?
y
n
END

Fig. 2. Determination of vowel phoneme realization on the basis of their phonetic position

Defining the transitions between vowel realizations

The vowel transitions are composed whenever a vowel occurs next to a consonant. In order to enhance faithful reproduction the vowel realizations have to be adjusted to the actual phonetic environment. This adjustment affects the first and the last microelement of the vowel realization. The modification concerns the adjustment of intensity (A0) and the first two formants (F1, F2) in such a way that they should conform to the corresponding values of the preceding or following consonant.

Generation of the suprasegmental structure

The suprasegmental structure is generated when the segmental structure of the utterance has been defined. The construction of the suprasegmental structure is aided by the sentence stress typed in the text as well as the sentence final punctuation mark. The temporal structure of the utterance is modified so that the duration of the vowel bearing sentence stress is doubled. The sentence final punctuation mark defines one of the eight possible intonation contours to be used. The RUSSON program recognizes the following sentenre final punctuation marks: . (full stop), :(colon), , (comma), ; (semi colon), ! (exclamation mark), ? (question mark), ?! (question mark - exclamation mark), ?? (double question mark).
With this operation completed, the complex sound structure is ready to be produced.

Control of the speech synthesizer

The sequence of code thus generated is passed on to the speech synthesizer to control its operation when it sets sound to the text.

PROSODIC ASPECTS OF POLISH WORD SYNTHESIS

JANUSZ IMIOŁCZYK RYSZARD CIARKOWSKI

Acoustic Phonetics Research Unit, Institute of Fundamental
Technological Research, Polish Academy of Sciences,
Noskowskiego 10, 61-704 Poznań, Poland

ABSTRACT

Four Polish words with varying number of syllables (dał, dobra, normalny and naturalnie) were synthesized using a COMPUTALKER CT-1 speech synthesizer. For each of the words 8 basic Polish intonation patterns were obtained by appropriately controlling the F0 parameter. Three variants of the intonation patterns were prepared (a quasi-natural variant and two types of approximation) of which the quasi-natural variant, elaborated on the basis of F0 values extracted from natural utterances of the four words, served as the model for the remaining two. The total of 70 synthetic intonation patterns were tested for recognizability and naturalness in a listening experiment. On this basis, the optimum approximation type was determined and the contours most typical for Polish questions and statements were selected for further research in the synthesis of intonation.

INTRODUCTION

The perceptual impression of accent is generally claimed to result from variations in fundamental frequency, duration and intensity within the vowel segment of the accented syllable ([2], [6], [10]). Of these three parameters, F0 has been found to play the most important role in signalling accent ([7], [9], [10], [14]), intensity having the least significant effect ([2], [9], [10]).

In analyses of linguistic functions of intonation two most general types of utterances are distinguished: (1) unfinished (general interrogative, continuative) and (2) finished (statements, demands, specific questions). Utterances of the first type are usually characterized by a rising intonation and utterances of the second type - by a falling intonation. In perceptual identification of each of these types the following three factors are of particular importance:

1) F0 level at the turning point, i. e. the point immediately preceding an F0 rise or fall
2) direction of F0 change and
3) its range.

For example, subjective impression of a general question is the stronger, the greater the F0 increase within the accented syllable ([8], [11]) and the higher the F0 value at the turning point ([8]). The effect of the F0 value at the turning point may even be more relevant than that of the final F0 rise, especially if the latter's range is relatively small. Moreover, the perceptual impression of accent is the stronger, the faster the F0 rise within the accented syllable ([13]).

TECHNICAL BASES OF WORD SYNTHESIS

For the purposes of the present experiment, four Polish words (dał, dobra, normalny and naturalnie) were synthesized using a COMPUTALKER CT-1 formant speech synthesizer controlled - via a minicomputer MERA 303 configuration (Fig 1; cf. also [4]) - by specially developed software. COMPUTALKER CT-1 simulates the transfer function of the vocal tract by means of formant filters connected in series. Apart from the synthesizing unit, composed of noise and glottal tone generators as well as nasal, noise and formant filters, it also includes a logistics unit the function of which consists in:

a) conversion of 8-bit digital parameters into the analog form controlling the elements of the synthesizing unit
b) short-term memory storing of the parameters.

A set of 9 parameters (amplitudes of the glottal tone, aspiration noise and nasal resonance, F0, F1, F2, F3, amplitude and frequency of the friction noise) has to be input to the synthesizer in digital form (1 byte per 1 parameter) for the synthesizer to be controlled. This is done by means of a parallel bus of data (each control parameter is given a 4-bit code) termed "frame". The rate of control data transmission in the system corresponds to an average speech rate (1 frame per 10 ms).

WORD SYNTHESIS. PRINCIPLES CONCERNING DURATION AND AMPLITUDE OF THE GLOTTAL TONE

Word synthesis was carried out using the existing library of synthetic diads of the CV and VC type ([3]). Appropriate transient segments were inserted between the diads to form natural-sounding words. At the initial stage each of the synthetic words received "intensity-durational" stress only, with constant F0 at 120 Hz over their whole length. Amplitude changes in the formant tract were of segmental character in almost all cases: within the steady state of any monosegmental phone amplitude values were not varied. Since the role of intensity in signalling accent is marginal (cf. above), amplitude values within stressed vowels were also held flat (the same approach was adopted by Abramson [1] and Mattingly [12]).

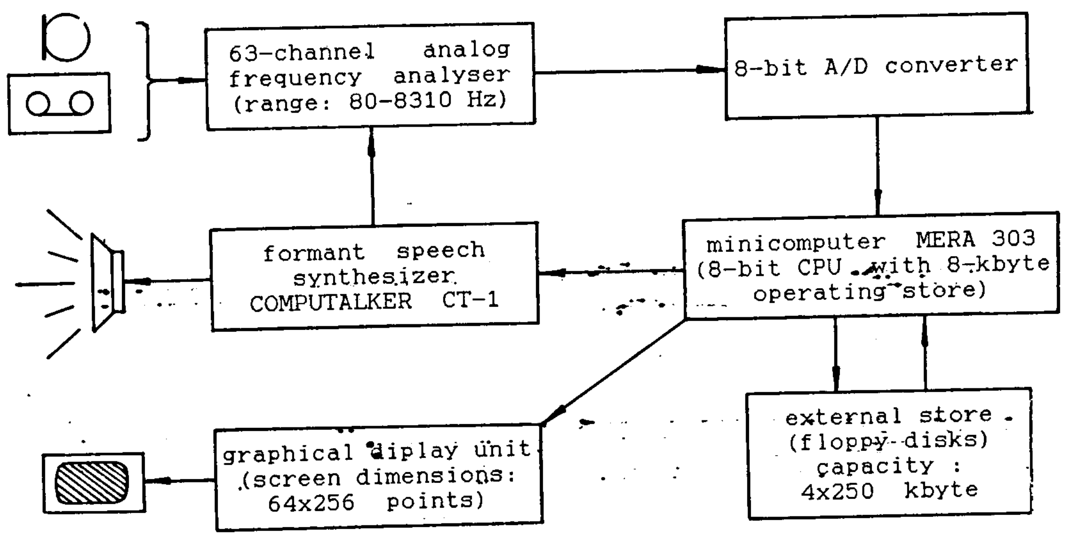

63-channel analog frequency analyser (range: 80-8310 Hz)
8-bit A/D converter
formant speech synthesizer COMPUTALKER CT-1
minicomputer MERA 303 (8-bit CPU with 8-kbyte operating store)
graphical diplay unit (screen dimensions: 64x256 points)
external store (floppy-disks) capacity : 4x250 kbyte

Fig. 1. Minicomputer configuration for speech synthesis and analysis

Durations of individual phones making up the four words were determined on the basis of the results given by Richter [15]. Preliminary listening tests led to the following duration of the steady-state part of the stress-bearing vowel within the total word duration:

dał - 190 ms (total - 480 ms)
dobra - 130 ms (total - 670 ms)
normalny - 120 ms (total - 910 ms)
naturalnie - 90 ms (total - 1060 ms)

SYNTHESIS OF INTONATION

For the purposes of intonation synthesis, 8 contours (corresponding to the most typical Polish word intonations) were selected from the inventory of Polish intonemes put forward by Steffen-Batogowa ([16]). These were:

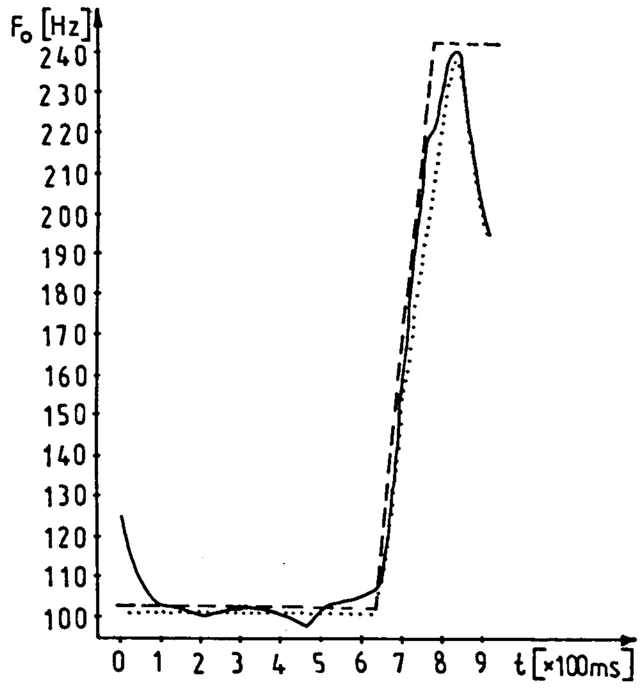
1) low rising (LR)
2) full rising (FR)
3) high rising (HR)
4) level (L)
5) full falling (FF)
6) low falling (LF)
7) full rising-falling (FRF)
8) low rising-falling (LRF)

Natural utterances of the four words, spoken with all these intonations by a skilled phonetician, were tape-recorded. As the utterances of the words dał, normalny and naturalnie containing the composite contours (LRF and HRF) sounded somewhat artificial, they were excluded from further examination. F0 patterns in the remaining utterances were analysed using a TM3 pitch meter. The resulting sequences of absolute F0 values, each corresponding to a 10 ms interval, were used to synthesize quasi-natural intonations which served as the model for the following two types of approximation of natural F0 contours:

1) Approximation by a broken line (A1) consisting of three or four segments. This is the most frequently applied approximation (cf. e.g. [5], [11]). Four-segment broken line was utilized in rising-falling intonations. Level intonations were approximated by a straight line.
2) Step-wise approximation (A2), in which a rapid change in F0 (10 Hz/10 ms), carried out within the accented syllable, occurred between two level segments.

Irrespective of F0 pattern type, neither of the two parameters responsible for signalling accent (i.e. duration and amplitude) was modified. It was assumed that variations in length of the accented vowel occurring among different types of F0 patterns do not affect accent perception and intonation type identifiability in any significant way.

Altogether, 70 synthetic intonation patterns were prepared. The total fundamental frequency range utilized in the synthesis covered frequencies from 77 Hz to 250 Hz. In the majority of cases, however, F0 values were not lower than 88 Hz and did not exceed 240 Hz.



quasi-natural contour ————
approximation by a broken line ··············
step-wise approximation - - - - - - - -

Fig. 2. NORMALNY - full rising intonation

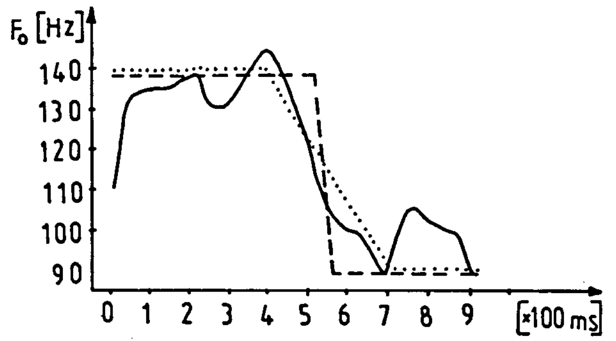
Examples of synthetic F0 contours are given in Figs. 2 and 3.

PERCEPTUAL EVALUATION OF SYNTHETIC INTONATIONS

Recognizability and naturalness of all the synthetic intonations were evaluated perceptually, in two listening tests, by 18 subjects divided into two panels (A and B). The task of panel A, consisting of 8 persons with professional experience with respect to speech melody, was to identify the type of intonation pattern presented (LR, FF. L etc.). Panel B, composed of 10 "naive" subjects, was to qualify the synthetic utterances as questions or statements.

Results

Panel A. For the total of 560 identification trials (70 patterns x 8 subjects) 401 correct responses were obtained (ab. 72 %). Identification rate of the intonations in the four words was the highest with the step-wise approximation (76 %) The quasi-natural intonations and the contours approximated by a broken line were recognized correctly in 69 % and 71 %, respectively.
Of the 8 types of synthetic intonations, low rises were recognized most efficiently (92 per cent of correct responses), whereas the composite patterns yielded the poorest identification results (only 21 per cent of correct responses were obtained for full rise-falls). The recognition scores for the remaining intonations were as follows: low fall - 78 %, level - 69 %, full fall - 81 %, high rise - 66 %, full rise - 73 %, and low rise-fall - 33 %.



quasi-natural contour ————
approximation by a broken line ··············
step-wise approximation - - - - - - - -

Fig. 3. NORMALNY - low falling intonation

Panel B. Two of the three synthetic rising intonations (full rise and high rise) proved to be nearly equally effective in signalling a question: they were perceived as interrogative 98 % of the time. With the exception of the low rise, which was misidentified most frequently, all the remaining intonations were commonly judged (from 88 to 100 per cent of responses) as typical for statements.
As the case of the level intonation indicated, an F0 fall is not an indispensable condition for an utterance to be perceived as statement-like.

DISCUSSION

In 87 % of the erroneous responses, the direction of F0 change was recognized correctly and the error pertained only to the range of variation . With the low intonations, however, the contrary tendency was observed to occur. Even though the LR and LF patterns were apparently fairly easy to identify (92 and 78 per cent of correct responses, respectively), they were also quite often confused with the level pattern. The factor responsible for this perceptual similarity was most probably the relatively small range of F0 variation in LR and LF.
As stated above, identification scores obtained for the complex patterns (LRF and FRF) were the lowest of all. Due to their less frequent occurrence in Polish one-word utterances, the two were often perceived as simple falling intonations.
The reason why the step-wise approximation provided the best identification results of intonation types was the characteristic, abrupt change in F0 preceded and followed by relatively long (hence easily perceptible) segments within which F0 value remained constant (level). The drawback of the intonation patterns thus produced was the peculiar "singing effect" of the utterances containing them.
Responses given by panel B pointed to the occurrence of a tendency for preferring the "statement" alternative. An intonation rise within the accented syllable was found to be the necessary condition of the "question" response. Moreover, the rise had to be characterized by a sufficiently wide range or a sufficiently high F0 level within the pre-accentual segment of the word. At least one of these conditions was met with FR and HR intonations, which were almost unanimously judged as typical for questions. On the other hand, a considerable divergence of responses occurred with LR intonations which, owing to the low F0 level within the pre-accentual segment and the small F0 increase, were perceived as indicating statements by the majority of subjects (cf. [11]).

CONCLUSIONS

The results obtained in the present experiment provide a number of cues which are essential for further research in Polish intonation synthesis. They suggest, among others, that due to both relatively high recognition rate and naturalness, the optimum approximation variant is the one utilizing approximation by a broken line. Of the two rising intonations judged as typical for Polish general interrogative utterances, the FR should be selected as the model one, as it is characterized by a wider frequency of usage and, thus, is more neutral. For similar reasons, the LF (and, perhaps, the FF) should be chosen as the model "declarative" intonation(s).

REFERENCES

[1] A.S. Abramson, Static and dynamic acoustic cues in distinctive tones, Lang. & Speech 21, 1978, 319-325.
[2] C. Adams, R.R. Munro, In search of the acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterances of some native and non-native speakers of English, Phonetica 35 1978, 125-156.
[3] R. Ciarkowski Minicomputer MERA 303-controlled synthesis of selected Polish diads and their perception (in Polish), IFTR Reports 7/1984, Warsaw.
[4] R. Ciarkowski, J. Imiołczyk, Analysis-aided formant speech synthesis, MELECON '85, vol.II, 171-173, North Holland, 1985.
[5] J.E. Clark, A low-level speech synthesis by rule system, J. of Phonetics 9, 1981, 451-476.
[6] P. Denes, J. Milton-Williams, Further studies in intonation, Lang. & Speech 5, 1962, 1-14.
[7] L. Dukiewicz, Intonation of Polish Utterances (in Polish), Ossolineum, Wrocław, 1978.
[8] K. Hadding-Koch, M. Studdert-Kennedy, An experimental study of some intonation contours, Phonetica 11, 1964, 175-185.
[9] W. Jassem, J. Morton, M. Steffen-Batog, The perception of stress in synthetic speech-like stimuli by Polish listeners, [in:] Speech Analysis and Synthesis (W. Jassem, ed.), vol.1, 289-308, Warsaw, 1968.
[10] I.Lehiste, Suprasegmentals, The M.I.T. Press, Cambridge, Massachusetts, 1970.
[11] W. Majewski, R. Blasdell, Influence of fundamental frequency cues on the perception of some synthetic intonation contours, J. of Acoust. Soc. of Amer. 45, 1969, 450-457.
[12] I.G. Mattingly, Synthesis by rule of prosodic features. Lang. & Speech 9, 1966, 1-13.
[13] S. Öhman, Word and sentence intonation: A quantative model, Speech Transmission Laboratory QPSR 2-3, 1967, 20-54.
[14] J.P. Olive, Fundamental frequency rules for the synthesis of simple declarative English sentences, J. of Acoust. Soc. of Amer. 57, 1975, 476-482.
[15] L. Richter, Statistical analysis of the rhythmical structure of utterances in the Polish speech (in Polish), IFTR Reports 7/1984, Warsaw.
[16] M. Steffen-Batogowa, Versuch einer strukturellen Analyse der polnischen Aussagemelodie, Zeitschr. f. Phon. u. allgem. Sprachwiss. 19, 1966, 398-440.

RESEARCHES ON THE FIELD OF SYNTHESIS OF THE ESTONIAN LANGUAGE

EUGEN KÜNNAP

Institute of Cybernetics
Tallinn, Estonia, USSR 200108

This paper reports the results of synthesis of the Estonian language. Constructions of synthesizers are described and the rules of synthesis are presented.

STATISTICS OF SPOKEN ESTONIAN

Every sound has his individual character and in speech process has some influence over neighbour sounds. Therefore the frequency of occurrence of phonemes, diphonemes and trigrams were investigated.
Arbitrarily choosed segments of speech were recorded, transformed into the phonetic symbols and analysed using digital computer. Analysis was made by syntagmas,i.e. by pauses in fluent speech. Selection contains 105942 phonetic symbols,which formed 19620 words and 4923 syntagmas.
In this work 31 phonemes were distinguished. In the Estonian alphabet there are 23 letters. In foreign names and loan words we can find some other letters,from which the letter f appears most frequently. It means that in written Estonian some phonemes were designated by the same letters. Consonants /l/,/t/,/n/,/s/ and /d/ can also be palatalized, which in fig.2 are marked with an apostrophe. Estonian /s/ is pronounced unvoiced, but sometimes, when it stays between vowels or after voiced consonants, vocal cords are also used. In this case/s/ is perceived as semivoiced and marked with /z/. /n/ can be palatalized and nasal.These phonemes have also the property of distinguishing between words. Usual /n/ can be in any phonetic constructions, but nasal only in /ng/ or /nk/ combinations. It is marked with two apostrophes. /b/,/d/ and /g/ are used as the indicators of short forms of /p/,/t/ and /k/. But in some cases they differ from /p/,/t/ and /k/ not only by intensity but also by spectrum and way of pronunciation. Therefore /b/,/d/ and /g/ are taken as different phonemes and conventionally named as semivoiced plosives. In this way we have phonemes:/a,b,d,d',e,f,g,h,i,j,k,l,l',m,n,n',n'',o,p,r,s,s',z,t,t',u,v,õ,ä,ö,ü/. The three most frequent phonemes are:/a/(11,61%),/e/(11,53%) and /i/(9,88%). In our material we have 99829 different diphonemes. The three most frequent are:/st/(1,77%),/le/(1,76%) and /te/(1,60%). All in all there are 94858 trigrams,11917 different types. The three most frequent are: /ele/(0,55%),/ist/(0,52%) and /sel/(0,49%). Average number of phonemes in a word was 5,4 and in a syntagma 22,5.
The Estonian language is a quantitative language. There are three distinctive degrees of length, while different degrees give the word different meanings.In written text not all of the degrees are distinguished.To obtain more natural sounding synthesized speech, in some cases 5 degrees of length are used.

SYNTHESIZER WITH ANALOG CIRCUITS

The first version of terminal synthesizer consists of four oscillators, connected in parallel, pitch impulse generator, four delay circuits, four amplifiers,summator and final amplifier. The frequencies of all oscillators, durations of delay, amplitudes of formant frequencies and the time of decay of formant frequencies are controlled by means of functional generators,described below. All oscillators are excited by pitch impulse generator. To synthesize fricatives the amplified noise of diodes was used. Four bandpass filters of noise have the range from 50 Hz up to 10 kHz. By means of this synthesizer short phrases were synthesized.

HARMONIC SYNTHESIZER

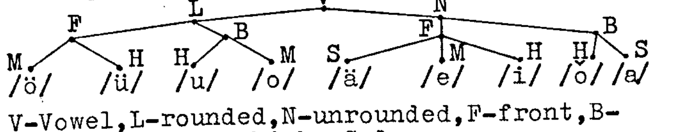
The voiced phoneme consist of formants. Each formant has his frequency,equal to the frequencies of fixed harmonics of pitch.As usual a formant is composed of most intensive neighbour including 2-3 harmonics,decaying in time. To have the oscillation of fundamental frequency and his harmonics, a generator of high frequency was constructed. Dividing the oscillation by means of trigger system the desirable harmonics were obtained. Received rectangular pulses were filtered and obtained sinusoidal oscillations were used as components of synthesized phonemes. The frequency of primary oscillation was obtained multiplying the fundamental frequency with the factors 11,9, 8,7 and 5. If the fundamental frequency has the value of 100 Hz, then the primary oscillation has the value of 5,54 MHz.Such choise of factors permits us to have all the harmonics, in practice needed to synthesize all phonemes.Primary generator consists of quartz generator of 30 MHz and generator, controlled by voltage in the range of 35 to 40 MHz. When they co-operate,the obtained beating has the range of 5 to 10 MHz and is used as primary generator. Synthesis of fricatives was made as described above.

FORMANT SYNTHESIZER WITH BANDPASS FILTERS

For the purpose to study several problems of synthesis of speech, besides synthesizer with oscillating circuits and harmonic synthesizer, a synthesizer with bandpass filters was constructed. Central frequencies of third-order analog filters were: F1 - 200+1000 Hz,F2 - 400+2000 Hz,F3 - 600+3000 Hz, F4 - 1,0+5,0 kHz,nasals F4 - 80+400 Hz. Fixed bandwidth of filters were - 80, 120, 150, 180 and 60 Hz respectively. To synthesize fricatives the filters with central frequencies of 800+4000 Hz and 1,2+ +6,0 kHz were made. To control the central frequencies of filters the method of pulse-width modulation was used.
The transfer function of vocal tract can be realised connecting resonators in parallel or in cascade, excited by pitch impulse or noise generators. In our synthesizer both methods can be used very easily, as well as the mixed connection of filters.
To control the parameters of synthesized sounds and to have the larynx-pulse generator, which can have the output voltage in any form, corresponding generators were worked out. The form of output voltage can be easily changed in wide varieties as well as during the experiments.
In our synthesizer 12 parameters were controlled. For this purpose 12 functional generators were worked out. Each generator has the matrix of wave-form oscillation,decipher with a system of diode keys and smoothing filter.For all of generators is one common pulse generator and comparator as circular counter of 8 triggers.The number of pulses of circular counter was chosen equal to 100. Matrix of wave-form oscillation has on his surface 32 stripes of foil. Each foil is under tension taken from voltage divider in limits of 0 to -7V. Across the foils are 100 metallic wires, each of them has a sliding silver contact. In the time of each pulse from pulse generator, pulses from circular counter were given to deciphers of all functional generators at the same time and in succession they switched on voltages from dividers of all function generators to input of smoothing filters. The outputs of these filters are used to control the parameters during the synthesis. When pulse generator has the frequency within the limits of 10 to 50 Hz, the duration of speech segments can be chosen from 10 to 2 sec.
The larynx-pulse generator has the same construction as previous generators. The frequency of pulse generator is electrically controlled in limits of 8 to 25 kHz,the matrix of wave-form oscillation has 130 stripes of foil,i.e. the voltage divider has 130 levels.The fundamental frequency can be changed from 80 to 250 Hz.

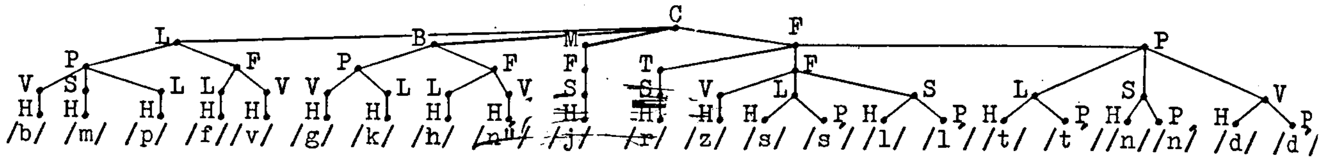
COMPUTER SYNTHESIZER

Further study of synthesis was made on the computer ES 1010. As usually, the model of vocal tract was formed by means of tunable second-order digital filters for the first three formants, fixed filter for nasals, the fourth and the fifth formants.The model consists of three branches,connected in parallel. One branch is the filter of nasals, the second consists of resonators of the third, second, first, fourth and fifth formants(fixed to 4500 Hz),connected in cascade, and the third branch consists of tunable bandpass filter for fricatives. Outputs of branches were summed up.As the source of tone the generator of triangle-form output voltage,and of noise, the generator of random numbers were used. For synthesis of nasals the branch of nasals and for synthesis of unvoiced fricatives the third branch were added to second branch. The controllable values were frequencies of pitch and first three formants, nasal and fricative formants,amplitudes of outputs of tone and noise generators, transitions,all in all 12 parameters.
To control the parameters of phonemes, they were divided by the articulary indication. In fig.1 the tree of the indication of vowels and in fig.2 the indication of consonants are shown. Every indication got his codemark and so they were stored into the memory of computer.For example phoneme /a/ has indication - VNBS, etc. The parameters of phonemes are given in table.
However, some parameters given in the table must be changed during the synthesizing process, depending on several circumstances. These changes and the rules of synthesis are as follows:1) If a vowel stands before plosives, then the first degree of length must be equal to 40 ms; 2) Duration of vowels in diphthongs must be equal to 120 ms; 3) If a vowel stands before /f/ of /v/, then the duration of transition must be equal to 60 ms;4) If before /r/ stands /ä/ or /õ/ and behind stands /a/, then the frequency of the first formant of /a/ must be equal to 750 Hz;5) If before /l/ stands /ä/ and behind stands /a/,then the frequency of the first formant of /a/ must be equal to 750 Hz; 6)/b/,/d/ and /g/ in the absolute first position in word must be synthesized as /p/,/t/ and /k/ respectively;7) /b/ in the last position in a word must be synthesized as /p/,but AN=40 units and F3= =F4=800 Hz;8) If /b/ is standing in the middle position between vowels,then the silent period before the noise burst must be equal to 60 ms;9) If /d/ is standing



V-Vowel,L-rounded,N-unrounded,F-front,B-back,M-middle,H-high, S-low.

Fig.1


C
L
B
M
F
P
F
P
F
F
T
F
S
L
S
V
V
S
L
L
V
V
L
L
V
S
S
V
L
P
H
H
H
H
H
H
H
H
H
H
H
H
H
P
H
P
H
P
H
P
H
P
/b/ /m/ /p/ /f//v/ /g/ /k/ /h/ ~~/n"/~~ ~~/j/~~ ~~/r/~~ /z/ /s/ /s´/ /l/ /l´/ /t/ /t´/ /n/ /n´/ /d/ /d´/

C- consonant
L- labial, B- back, M- medio-lingual, F- front
P- plosive, F- fricative, T- tremulant
V- voiced, S- sonorous, L- unvoiced; H- hard, P- palatalized

Fig.2

Table

Sym		TREE	DURATION	VOICE AMP.	NOISE AMP.	FRIC.AMP.	FORMANT FREQUENCY
				/AV/	/AN/	/AF/	
0		SPACE	5	0	0	0	2500/ 1500/ 500/ 3499
1	"	SPACE	10	0	0	0	2500/ 1500/ 500/ 3499
2	A	VNBS	8	50	0	0	2500/ 1100/ 750/ 3499
3	Ä	VNFS	8	15	0	0	2280/ 1600/ 870/ 3499
4	O	VLBM	8	50	0	0	2100/ 850/ 520/ 3499
5	Ö	VLFM	8	50	0	0	2280/ 1500/ 560/ 3499
6	U	VLBH	8	50	0	0	1500/ 600/ 350/ 3499
7	Ü	VLFH	8	50	0	0	2379/ 2080/ 300/ 3499
8	E	VNFM	8	30	0	0	2843/ 2000/ 480/ 3499
9	I	VNFH	8	30	0	0	3049/ 2450/ 280/ 3499
10	Õ (Q)	VNBH	8	50	0	0	2899/ 1200/ 400/ 3499
11	B	CLPVH	8/ 2	10/ 0	0/ 20	0/ 5	10/ 660/ 400/ 2000
12	F	CLFLH	8	0	100	0	2200/ 800/ 230/ 3499
13	G	CBPVH	8/ 2	0/ 0	0/ 20	0/ 0	2100/ 1200/ 600/ 3499
14	H	CBFLH	10	0	40	0	2599/ 999/ 400/ 3499
15	J	CMFSH	6	50	0	0	3049/ 2450/ 280/ 3499
16	K	CBPLH	10/ 3	0/ 0	0/ 40	0/ 0	2100/ 1200/ 600/ 3499
17	M	CLPSH	8	50	0	0	2000/ 900/ 200/ 3499
18	P	CLPLH	10/ 3	0/ 0	0/ 0	0/ 30	2000/ 660/ 400/ 2000
19	R	CFTSH	2/ 2/ 2	30/ 4/ 3C	0/ 0/ 0	0/ 0/ 0	2500/ 1500/ 500/ 3499
20	V	CLFVH	8	6	20	0	1650/ 625/ 170/ 3499
21	T	CFPLH	12/ 3	0/ 0	0/ 0	0/ 60	2599/ 1600/ 400/ 3999
22	Z,	CFFVH	8	20	20	0	2500/ 1500/ 500/ 4499
23	T´	CFPLP	16/ 1/ 1	0/ 0/ 0	0/ 0/ 0	0/ 60/(	2500/ 2000/ 600/ 3499
24	D,	CFPVH	8/ 2	0/ 0	0/ 0	0/ 10	2599/ 1700/ 400/ 3499
25	D´	CFPVP	8/ 2	0/ 0	0/ 0	0/ 20	2599/ 2000/ 600/ 3499
26	S,	CFFLH	8	0	0	3	2500/ 1500/ 500/ 4499
27	S´	CFFLP	8	0	0	7	2500/ 1500/ 500/ 4499
28	L,	CFFSH	8	30	0	0	2500/ 1400/ 400/ 3499
29	L´	CFFSP	8	30	0	0	2500/ 1700/ 400/ 3499
30	N,	CFPSH	8	50	0	0	2330/ 1800/ 200/ 3499
31	N´	CFPSP	8	50	0	0	2699/ 2100/ 310/ 3499
32	N"	CBFVH	8	50	0	0	2000/ 1250/ 290/ 3499

In the column DUR (duration) the units are in 10 ms. VA-voice amplitude, NA-noise amplitude, FA-fricative amplitude - are given in the relative units, where the unit 0 corres ponds to 0 dB and 100 to 40 dB. Frequencies are expressed in Hz.

in the middle position between vowels,then the silent period before the noise burst must be equal to 80 ms and duration of vowels must be lenghtened to 120 ms; 10)/d/ in the last position in a word must be synthesized as /t/;11) If /b/ stands before /u/, then it is necessary to decrease the duration of the noise burst to 10 ms and AN to 50 units;12) /g/ in the absolute last position in a word must be synthesized as /k/, while the amplitude of noise burst must be equal to 30 units and the duration of silent period before the noise burst equal to 40 ms;13) Duration of /f/ in the first position in a word must be equal to 120 ms, and AN=50 units;14) The first degree of length of /f/ must be equal to 120 ms,the second - 180 ms and third - 240 ms; 15)/b/,/d/ and /g/ must be synthesized as /p/,/t/ and /k/ respectively,if behind them in the middle of the word stand o' [illegible] plosives or /s/;16)If /h/ stands befor /v/, then the duration must be equal to 120 ms;17)/j/ in the first position in a word must be synthesized as /i/,only the duration must be equal to 40 ms;18) If /k/ stands before /u/,then F3 of /u/ must be equal to 1500 Hz;19)If /k/ stands in the middle position in a word,then AN=10 units 20)If /m/,/n/ or /n´/ stands in the middle position in a word,then AN=40 units;21)If /n/ stands before /a/,/o/,/u/ or /o´/,then F2=1600 Hz;22)If /n/ stands behind /e/,/i/ or /ä/,then F1=250,F2=2100,F3=2500 Hz;23) If /m/ or /n/ stand between vowels in an unstressed word,then the duration must be

equal to 60 ms;24)If /m/,/n/ or /n´/ are in unstressed word,then AN=40 units;25) If /p/ is in the first position in a word, then AN=60 units;26) If/p/ is in the last position in a word, then F3=F4=800 Hz;27) If behind /l/,/m/,/n/,/r/ or /v/ in the middle of word stand /s/ or /h/, then they must be synthesized only by means of noise source;28)If plosives are in the middle position in a word, then the duration of noise burst must be equal to 5 ms;29)Duration of /v/ in the first position in a word must be equal to 120 ms,in an unstressed syllable between vowels - 40 ms; 30)If /v/ is in the first position in a word,then AN=60 units and AV=10 units;31) In compuond words between simple words must be a pause 10 ms;32)Duration of vowels of the first degree of length must be 120 ms,in an unstressed word - 80 ms,second - 180 ms,third - 240 ms, in the last position of word, when word is stressed-300ms; 33)Duration of voiced consonants of the first degree of length must be equal to 80 ms, second - 180 and third - 240 ms;34) Duration of nasals of the first degree of length must be equal to 80 ms,but if /n/ stands before vowels,then 40 ms,second for /n/ - 120 ms,for /m/ - 140 ms,third - 180 ms, in a stressed syllable - 240 ms;35)Duration of the silent period before noise burst for plosives of third degree of length - 240 ms;36)Duration of unvoiced fricatives of the first degree of length must be equal to 80 ms, second - 150 ms for /h/ and 120 ms for /s/, third - 240ms for /h/ and 280 ms for /s/.If they stand between vowels in an unstressed word,then the first degree of length must be equal to 60 ms;37)The synthesis of /n´/ and /l´/ must begin with synthesizing /i/ with the duration of 40 ms;38)The synthesis of /d´/, /t´/ and /s´/ must begin with synthesizing /i/ with the duration of 60 ms.
The durations are given for the middle rate of speech,i.e.8-10 phonemes in sec.

SYNTHESIZERS,CONTROLLED BY MICROCOMPUTERS

In the first version of terminal synthesizer,controlled by means of a microcomputer, the functional generators of the synthesizer,described above,were replaced with a microcomputer.The parameters of phonemes were stored into the memory of constants by digital keyboard.Every cell of memory has his address. To synthesize the speech signals,the contents of constant memory were fed into the memory of control parameters by means of alphabet keyboard.Both memories were connected together with a control-block of logical circuits.The task of the control block - to feed the contents of memory of constants by addresses to registers of synthesizer by commutator,using an alphabetic keyboard.The contents of all 12 registers of memory (12 controlled parameters) are fed at the same time by D/A converter to synthesizer.By means of indicator of 7 light diodes it was possible to check the contents of memories by the addresses.The table of indicator was formed of 8 diod complexes.The last version of synthesizer,controlled by means of microcomputer is more flexible and perfect.The model of vocal tract is,as described above, composed of third-order analog filters.To control the central frequencies of filters, the pulse-width modulation is also used. Central frequencies can be controlled in the range of 8 bits,but less bits are sufficient in practice.The central frequencies of the first and the third formant filters are controlled in the range of 5 bits,the second - 6 bits,filters of unvoiced fricatives and plosives 2 bits.The range of filters:F1-150+750 Hz,F2-500+2100 Hz,F3-1,5+3,2 kHz,FF-1,5+4,7 kHz,F4-3,5 kHz,F5-4,5 kHz,FN-200 Hz.The rate of transitions is controlled by 2 bits (20,40,60 and 100 ms),frequency of pitch - 3 bits (from 100 to 154 Hz),amplitudes of tone and noise generators and output amplifier - 2 bits each.ASCII - coded Estonian text is transformed into the form of discrete control signals.The microprocessor system consists of a processor unit (KP580IK80), ROM with the capacity of 6 kbyte,RAM with capacity of 2 kbyte and input-output interface.Digital control signals from the microprocessor are converted into continuous-time analog signals to control the parameters every 10 ms.

CONCLUSION

The synthesizers,described above, allowed us to research several aspects of synthesis of the Estonian language to achieve speech sounding close to natural.

VOCAL JITTER AS AN INDICATOR OF CHANGES IN PSYCHOPHYSIOLOGICAL AROUSAL

Erkki Vilkman
Phoniatric Dept.,
Tampere University
Central Hospital,
SF-33520 TAMPERE
FINLAND

Olavi Manninen
Dept. of Public Health,
University of Tampere

Eija-Riitta Lauri
Phoniatric Dept.,
Helsinki University
Central Hospital

Tarmo Pukkila
Dept.of Mathematical
Sciences,
University of Tampere

ABSTRACT

The possibilities of using cycle-to-cycle changes in fundamental frequency (jitter) for estimating changes in subjects' psychophysiological arousal were studied. The subjects (n=20) were exposed to four different combinations of dry bulb temperature (20°C or 35°C), noise (90 dBA) and whole-body vibration (sinusoidal 5 Hz) in a special exposure chamber. The exposure lasted the whole day. The jitter was measured manually of an excerpt from a text which the subjects read during rest and exposure separately from morning and afternoon samples. Only in the afternoon samples did the changes in jitter caused by exposure to 20°C (increase) and 35°C (decrease) temperatures differ significantly (p<0.05, df=8).

INTRODUCTION

The tremor of one's voice is a well-known feature of anxiety, prompted, for example, by performing situations. Thus it is natural that there have been attempts to use voice changes, for instance, in detecting deceptive behavior. Commercially available apparati which have been claimed to reflect a person's emotional state have been developed in this field. The most popular of these is the so-called Psychological Stress Evaluator (PSE), which in addition to forensic science /1,2/ has also been used in the field of psychology /3/.

In brief, PSE consists of an electrical integration circuit which filters the acoustic signal so that a fluctuation at a rate of about 10 Hz of the baseline can be seen. In anxiety states this tremor is claimed to diminish /1,2/. The origin of the tremor is somewhat obscure at present. The reliability and validity of PSE has been seriously questioned /1,2,4/.

Another fluctuation phenomenon of the human voice is the cycle-to-cycle variation in the fundamental frequency, i.e. the jitter. Jitter has been studied rather extensively as a sign of vocal pathology /5/, but it has also been found to be associated with the emotional contents of speech /6/. It has been claimed that in so-called "stress situations" the jitter scores tend to lower with increasing threat /7/.

Temperature, noise and vibration are important exposure factors for research, because in modern society people are exposed to them almost daily. However, there is relatively little knowledge concerning their individual effects and hardly any concerning their combined effects on the psychophysiological arousal of human subjects. In an earlier study on changes in prosodic features of speech due to environmental factors /8/, we noticed that different combinations of temperature, noise and whole-body vibration caused changes in the average fundamental frequency, intensity, spectral characteristics and durational variables of speech.

The changes in the prosodic features could be interpreted in terms of existing knowledge of psychophysiological changes related to similar exposure conditions /8/. The aim of this preliminary study was to determine whether jitter could be used as a sensitive indicator of psychophysiological arousal. For this purpose we analysed excerpts of speech samples from four different exposure combination cells of the earlier study.

SUBJECTS AND METHODS

Excerpts of reading samples of twenty healthy male subjects were analysed. The subjects (n=20; 5 in each) were exposed to the following exposure combinations: 1) 20°C temperature (T) , no noise (N), no vibration (V) (T1N0V0); 2) 35°C temperature only (T2N0V0); 3) 35°C and 90 dB(A) noise (T2N1V0) and 4) 35°C temperature, 90 dB(A) noise and 5 Hz sinusoidal whole-body vibration along Z-axis (T2N1V1). The experiment was carried out in a special exposure chamber. During the test, subjects sat in a vibration chair. See /9/ for a detailed description of the exposure arrangements.

The samples of speech were recorded during a pause in the test. Altogether four speech samples were recorded. Rest session samples were recorded from 9:13 to 9:15 in the morning before the exposure and in the afternoon from 12:13 to 12:15. The first exposure sample was recorded in the morning between 11:08 and 11:10 after an exposure session lasting 80 minutes. The second was recorded in the afternoon at 2:08-2:10 after total exposure (lasting altogether176 minutes). See /8,10/ for the recording arrangements and further details.

The excerpt on which the cycle-to-cycle analysis was performed consisted of vowels and voiced consonants (/on lauan.../). This excerpt starts a new chapter and is very emphatic. The excerpt ends at a voiceless consonant(/t/). The sampling rate of the signal was 10 kHz. In the analysis of recorded sinusoidal sounds the jitter of the apparati fell below the accuracy of the measurements. The measurements were carried out blindly. On an average (X±SD) 45.7±9,8 successive periods were measured. Measurements were made manually from oscillographic displays by means of a cursor. The results of the measurements were stored on disk and drawn on plotting paper using microcomputer-based (Motorola Exorset) programs.

The jitter value was formed as the difference between the observations and the five point moving average to avoid the influence of the general trend of the fundamental frequency. If the original observations are denoted by X_t, the five-point moving average value is obtained from the formula

$$T_t = (X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2})/5.$$

The jitter J_t is therefore obtained as the difference

$$\begin{aligned} J_t &= X_t - T_t \\ &= (- X_{t-2} - X_{t-1} + 4X_t - X_{t+1} - X_{t+2})/5, \end{aligned}$$

i.e. J_t is obtained as a five-point moving average such that the sum of the weights is zero.

The standard deviation (SD) of the jitter J_t was used as the jitter index [J(*tot*), J(*asc*)]. The mean of the fundamental frequency (Fj) of the measured excerpt was calculated. The analyses were run separately for the ascending part (ΔJ(*asc*) (pitch rise) and the total excerpt(ΔJ(*tot*) (c.f. Fig.1).

Before statistical treatment, both the morning and afternoon exposure sample results of jitter were standardized by subtracting the results of the preceding rest samples. The correlation of jitter [ΔJ(*tot*)] with the changes in fundamental frequency (ΔF0), intensity (ΔI) total reading time (ΔTt), articulation time (Ta), and number and errors in /tapaka/ and /pataka/ word repetitions obtained from the earlier study /8/ was calculated.

RESULTS

The mean and the standard deviation of Fj of the excerpt was 112.5±19.3 Hz The mean J(*tot*) value of the total excerpt was 2.9 Hz and the mean J(*asc*) of the ascending part 2.0 Hz.

Fig. 1. shows an example of FO curves in rest and exposure.

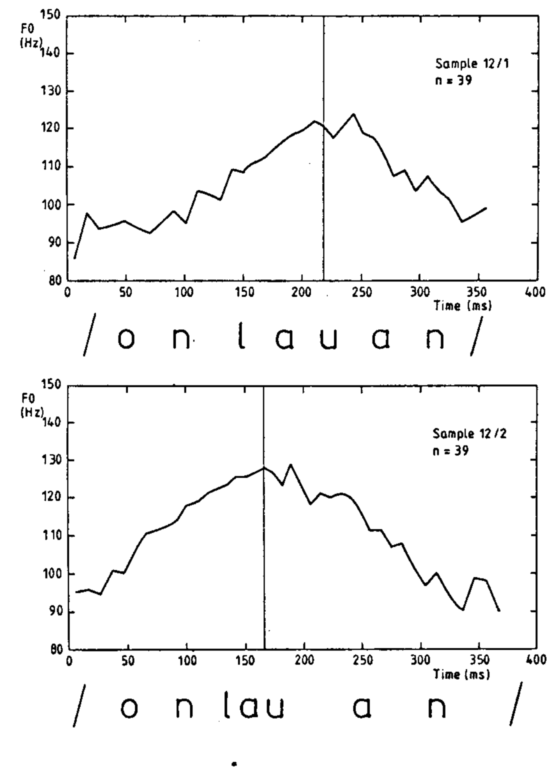


Fig. 1. The F0 curves measured cycle-to-cycle on the excerpt from the text (/on lauan.../) for one subject (exposure combination T2N1V0). The jitter is bigger for rest (top) [J(*tot*)=2.29 Hz, J(*asc*)=2.39 Hz] than exposure (bottom) [J(*tot*)=1.54 Hz, J(*asc*)=1.21 Hz]. F0 is higher in the exposure sample. The vertical line shows the estimated end of the ascending part (start of /u/) of the curve which was analysed separately.

The mean changes in the measurements of the excerpt are shown in Table I by exposure combinations. It can be noted that the changes in jitter values [ΔJ(*asc*) and ΔJ(*tot*)] are small in general and that interindividual variation is big. Differences between the effects of the exposure in the morning and afternoon samples are also clear. In the morning samples the clearest drop in jitter [ΔJ(*tot*)] has occurred due to exposure to combination T2N1V0. This combination raised F0, I, Tt and Ta values of the total sample, and the rate of word repetitions (/tapaka/ and /pataka/) increased /8/. In the afternoon samples the effects of T1N0V0 and T2N0V0 on ΔJ(*tot*) differed statistically significantly (p<0.05, t=2.32, df=8).

Table I. Average changes (X±SD) in jitter and fundamental frequency measurements of the text excerpt. See text for symbols.

	ΔJ(*asc*)	ΔJ(*tot*)	ΔF(j)
a.m.	(Hz)	(Hz)	(Hz)
T1N0V0[1]	-0.5±1.4	0.5±2.0	-1.1.±2.9
T2N0V0	-0.2±0.9	-0.1±1.4	6.1±7.1
T2N1V0	-0.7±1.2	-2.0±1.9	5.3±6.5
T2N1V1	0.2±0.7	-0.6±1.5	7.8±5.9
p.m.			
T1N0V0	0.0±1.4	1.5±2.5	4.4±4.4
T2N0V0	-0.4±1.3	-1.3±1.2	5.1±3.0
T2N1V0	0.3±0.8	0.0±1.0	7.7±6.8
T2N1V1	0.2±0.7	-0.4±2.5	1.2±5.6

[1] Exposures were T1=20°C, T2=35°C temperature, N0=no noise, N1=90 dBA, V0=no vibration,V1= 5 Hz sinusoidal vibration.

Table II shows the correlation coefficients of jitter measurements with prosodic features of speech. It can be seen that in general the correlations are weak. The ΔJ(*tot*) values, however, show a tendency to negative correlation with ΔF0 and ΔI. Very interesting are the correlations between ΔJ(*tot*) and the /pataka/ word repetitions and especially the statistically significant correlation with the number of errors in the afternoon samples and consequently also in a.m.+p.m. calculations. ΔJ(*tot*) also correlated weakly with the ΔFj value: in the morning samples the correlation was negative (r=-0.26) and in the afternoon positive (r=0.32).

Table II. Product moment correlation coefficients (*r*) between changes in jitter and prosodic variables.See text for symbols.

					Δ/pataka/		Δ/pataka/	
	ΔF0	ΔI	ΔT t	ΔTa	n:o	errors	n:o	errors
a.m.								
ΔJ (*tot*)	-0.28	-0.14	0.03	-0.25	-0.15	-0.11	-0.31	0.17
ΔJ (*asc*)	-0.15	0.07	-0.12	-0.22	-0.01	-0.36°	-0.19	0.18
p.m.								
ΔJ (*tot*)	-0.24	-0.28	0.05	0.08	0.17	0.05	-0.17	0.58**
ΔJ (*asc*)	-0.04	0.09	-0.14	0.16	0.02	0.33	-0.07	0.25
a.m.+p.m.								
ΔJ (*tot*)	-0.25	-0.29°	0.05	-0.03	0.02	-0.03	-0.24	0.43**
ΔJ (*asc*)	0.05	0.07	-0.10	0.02	-0.01	0.13	-0.16	0.23

° $p<0.10$, * $p<0.05$, ** $p<0.01$, a.m.&p.m. df=20, a.m.+p.m. df=40

DISCUSSION

The changes in fundamental frequency (ΔFj) of the excerpt did not show significant exposure-specific changes. This may be due to the fact that such an average value reflects more the changes in reading style, which is more complicatedly related to changes in psychophysiological arousal, than the supposed tension changes underlying the long-time average F0 (c.f. /11/).

A five-points moving average was used to calculate the jitter. According to Kitajima et al. /12/ the use of four points already produced a satisfactory smoothening effect. The absolute values of the jitter of the present study cannot be compared to other studies because of the differing types of jitter indices /13/. The jitter values of the present study may be biased due to a relatively low sampling rate, which has been found to increase the magnitude of jitter /5/ and the random effect of the jitters of the recording and the analysis system (c.f. /7/).

The magnitude of the physiological jitter has been reported to be related to the asynchronous firing of the motor units of the cricothyroid muscle; thus the decrease in jitter is caused by an increase in neural input and arousal /7,14/. A drop in jitter values /6, 7/ and a rise in F0 and intensity /4,15,16/ have been noticed to be connected with psychological stress. In terms of these reports it can be supposed that the exposure to the combination of 35°C temperature and 90 dB(A) noise caused a rise in arousal in the morning samples. The effect of T2N0V0 on the ΔJ(*tot*) in the afternoon samples also implies psychological stress.

The positive correlation between the jitter and errors in /pataka/ word repetitions might imply a common basis for these changes. In an earlier study we found that /pataka/ word repetitions and errors might be useful in assessing changes in a person's arousal /8/. It can be hypothesized that the increase in magnitude of jitter is associated with lowered arousal, and thus the tendency to increased errors can be interpreted in terms of motivation (the correlation was highest in the afternoon samples). And vice versa, the performance becomes better with higher arousal and an increase in motivation.

In conclusion, the results of this preliminary study suggest that the jitter is not a more sensitive indicator of a person's psychophysiological arousal than the prosodic features of speech. However, it has to be kept in mind that different vocal and speech variables may reflect some independent specific changes in arousal due to the exposure. This would be in line with recent psychophysiological studies on the demand-specificity of changes in various physiological measures /17/.

REFERENCES

/1/ VanDercar DH, Greaner J, Hibler NS, Spielberger CD, Bloch S. A description and analysis of the operation and validity of the psychological stress evaluator. J Forens Sci 25:174-188, 1980.

/2/ Horvath F. Detecting deception: the promise and reality of voice stress analysis. J Forens Sci 27:340-351, 1982.

/3/ Smith GA. Voice analysis for the measurement of anxiety. Br J Med Psychol 50:367-373, 1977.

/4/ Hollien H. Vocal indicators of psychological stress. pp. 47-72. In: Wright, Bahn, Rieber (eds.) Forensic psychology and psychiatry. Ann N Y Acad Sci vol 347, 1980.

/5/ Heiberger VI, Horii Y. Jitter and shimmer in sustained phonation. Speech Lang: Adv Basic Res 7:299-332, 1982.

/6/ Lieberman P. Perturbations in vocal pitch. J Acoust Soc Am 33:597-603, 1961.

/7/ Brenner M, Shipp T, Doherty T, Morrissey P. Voice measures of psychological stress. pp. 240-248. In: Titze, Scherer (eds.) Vocal fold physiology. Center for Perf Arts, Denver 1983.

/8/ Vilkman E, Manninen O. Changes in prosodic features of speech due to environmental factors. Speech Comm 5: 331-345,1986.

/9/ Manninen O. Hearing threshold and heart rate in men after repeated exposure to dynamic muscle work, sinusoidal vs stochastic whole body vibration and stable broadband noise. Int Arch Environ Health 54:19-32, 1984.

/10/ Vilkman E, Manninen O. Changes in prosodic features of speech under complex exposure conditions. pp. 145-167. In: Manninen (ed.) Combined effects of environmental factors, Proc Ist Int Conf Combined Effects of Environmental Factors, Tampere 1984. Keskuspaino, Tampere, Finland 1984.

/11/ Scherer K, Ladd DR, Silverman KEA. Vocal cues to speaker affect: testing two models. J Acoust Soc Am 76:1346-1356, 1984.

/12/ Kitajima K, Tanabe M, Isshiki N. Pitch perturbation in normal and pathological voices. Studia Phonol 9:25-32, 1975.

/13/ Zyski BJ, Bull GL, McDonald WE, Johns ME. Perturbation analysis of normal and pathological larynges. Folia Phoniat 36:190-198, 1984.

/14/ Baer T. Vocal jitter: A neuromuscular explanation. pp.19-22. In: Lawrence, Weinberg (eds.) Transact 6th Symp Care Prof Voice. The Voice Foundation, New York, 1980.

/15/ Scherer K. Vocal indicators of stress. pp. 189-220. In: Darby (ed.) Speech evaluation in psychiatry. Grune & Stratton, New York, 1981

/16/ Williams CE, Stevens KN. Vocal correlates of emotional states. pp. 221-240. In: Darby (ed.) Speech evaluation in psychiatry. Grune & Stratton, New York, 1981.

/17/ Lyytinen H. The psychophysiology of anticipation and arousal. PhD Diss. University of Jyväskylä, Jyväskylä, Finland, 1984.

ACKNOWLEDGEMENTS

This study was financed by the Academy of Finland. The authors wish to express their gratitude to Professor Jaakko Lehtonen for technical and practical support.

SPEECH PRODUCED UNDER ADVERSE CIRCUMSTANCES

Z. S. Bond
Department of Linguistics
Ohio University
Athens, OH USA

Thomas J. Moore
Armstrong Aerospace Medical
Research Laboratory
Wright-Patterson AFB, OH USA

Speakers sometimes are required to function under adverse or demanding speaking circumstances. We have been examining the effects of two physically adverse conditions, acceleration and high noise levels, on the acoustic-phonetic structure of speech.

INTRODUCTION

Over the past few decades, the acoustic-phonetic structure of speech has been investigated in considerable detail. Almost all of this work has described speech produced carefully, with minimal distraction or disturbance of the speaker; in short speech in benign circumstances. Yet speakers sometimes have to function under circumstances which either impose or require changes in speech. First, speech may be produced in different styles, such as very clear, slow, loud, and so forth, as seems appropriate for a specific audience. Speakers may also be influenced by psychological states such as excitement, fatigue, discomfort or distraction. Finally, speakers may be influenced by the physical circumstances under which they are required to function, such as high ambient noise levels or various forms of physical motion.

The acoustic-phonetic consequences of style differences have received some attention (for example, Schulman, 1985; Picheny, et al., 1986); states of psychological arousal have been investigated primarily from the point of view of assessing the condition of a speaker, the question of interest being whether it is possible to detect stress by examining characteristics of speech. The effects of physically adverse circumstances on the acoustic-phonetic characteristics of speech have received the least investigation.

PHYSICALLY ADVERSE CIRCUMSTANCES

Over the past two years, we have been studying the speech of speakers in physically adverse circumstances: while hearing high noise levels, experiencing high sustained acceleration or whole-body vibration. We have found that speech produced in adverse circumstances differs systematically from speech produced in benign environments. In this report, we wish to characterize briefly the physical environments which we have investigated and summarize our findings to date.

Acceleration.

Acceleration vectors are classified according to the direction in which they act on the human body. Headwards acceleration, which tends to displace body tissue footward, is termed positive G or +Gz. High sustained acceleration is defined as exposure to acceleration forces of 6G or greater for periods in excess of 15 seconds (Burton, et al., 1974). High sustained accelerations would be encountered in some aircraft.

Method. We have examined the acoustic-phonetic structure of isolated words as produced by two male speakers in two conditions: 1) while sitting in the gondola of the centrifuge at the Armstrong Aerospace Medical Research Laboratory without acceleration (at 1G) and at an acceleration level of +6Gz. The speech of the subjects was recorded through the M-101 noise-cancelling military microphone which was located within a standard Air Force oxygen mask. Speech analysis was performed using the program SPIRE (Zue and Cyphers, 1985) on the Symbolics 3670 computer. Measurements from SPIRE displays were made of the first three formants of vowels, word durations, vowel durations, intervocalic obstruent durations, and fundamental frequency in stressed and unstressed syllables.

Results. Speech produced under high acceleration sounds quite normal, even natural. Whether a word has been produced under acceleration is not obvious to a listener. However, some acoustic-phonetic characteristics appear to differ between speech produced under acceleration and speech produced in a benign environment. Differences were detectable both in the timing and spectral composition of segments.

The formant structure of vowels shifted under acceleration. Only the third formant did not exhibit systematic changes. The first formant of most vowels was somewhat higher; the second formant was lower for front vowels and higher for back vowels. The F1/F2 vowel space of one speaker is given in Fig. 1. The vowel space shrinks, suggesting lessened mobility of the articulators.

Mean fundamental frequency in stressed syllables increased for both speakers, by 10 Hz for Speaker 1, somewhat less for Speaker 2.

All but one of the words measured increased slightly in mean duration for Speaker 1; for Speaker 2, however, the mean duration of some word increased, of others decreased. Word duration shifts resulted almost entirely from shifts in the duration of vowels, so that under acceleration Speaker 1 produced longer vowels while Speaker 2 used variable vowel durations. The duration of intervocalic obstruents decreased slightly for both speakers under acceleration. Because of variability in response, it is difficult to determine whether changes in segment and word durations are a function of speaker characteristics or of acceleration levels. Further details of this study are available in Bond, Moore, and Anderson, 1986.

High Ambient Noise.

When in the presence of high ambient noise, speakers tend to increase the level of their speech, presumably to maintain what they judge to be an appropriate level of sidetone. This increase in loudness is typically accompanied by an increase in pitch, reflected in fundamental frequency. While these relationships have been noted repeatedly and described in the extensive literature on the Lombard effect, only recently have other changes of speech produced under noise received attention. Pisoni, et al., (1985) have reported that vowels as defined by the first two formants become less distinct and that the distribution of energy within the speech spectrum shows an increase in high frequency components.

Method. We have examined the speech of one male speaker, a 20 year old student at a Midwestern university, in a number speaking conditions in conjunction with noise exposure. The speaker was recorded on four separate days in five speaking conditions and two recording environments.

The recording environments can be characterized as 'operational', the speaker wore a standard Air Force helmet equiped with an oxygen mask and an M-101 microphone, and 'laboratory', the speaker was wearing a boom microphone.

The speaker recorded two repetitions of ten spondee words in three noise exposure conditions: white noise at 85dB, 95dB and 100dB SPL; he also recorded the same materials in quiet and when instructed to be 'loud'. We will limit our report to a description of speech produced under the two highest noise levels, in comparison with speech produced in quiet and speech intended to be 'loud'. In all cases, the noise was presented over earphones.

The durations of words and segments, the fundamental frequency and energy at the mid-point of both syllables, and the formant structure of vowels and diphthongs were measured from SPIRE displays.

Results. The speaker reported some discomfort while speaking in the operational condition, particularly when he was also exposed to noise. In all conditions, however, his speech was intelligible and produced in a relatively casual conversational style. The first syllable of each word, receiving stress, was produced with a higher fundamental frequency than the second. Since the speaker was producing words in isolation, the second syllable was longer than the first, a result of pre-pausal lengthening.

Speech in noise. Average word duration varied by approximately 100 msec. from speaking in quiet to speaking in noise. The majority of the variability was a function of vowel durations. In quiet, the average duration of the first syllable was 156 msec. and of the second syllable, 234 msec. The first syllable was longest when speaking in 100 dB noise, increasing to 178 msec. The second syllable increased to 265 msec. There was considerable variability, however; vowel segments did not invariably lengthen in noise.

The second syllable was produced with less energy than the stressed first syllable in all noise conditions. In quiet, the second syllable was produced 9dB lower than the first. In the two noise conditions, the differences between the two syllables decreased to 2 dB and 4 dB.

As would be expected, the fundamental frequency of both syllables increased when speaking in noise, though there was some variability and the increases were not directly proportional to noise levels. In quiet, the first syllable was produced at an average F0 of 138 Hz, the second syllable at 109 Hz. At 100 dB noise, the two syllables were produced at an average F0 of 147 and 119 Hz. The absolute levels differ, but the F0 difference between the two syllables is roughly proportional.

Speaking in noise had a detectable effect on vowel formants. Noise was associated with a higher F1, and a lower F2 for front vowels. The effects were most marked for high vowels. The formant shifts associated with noise are given in Fig. 2.

Oxygen mask. Wearing an oxygen mask had a detectable effect on speech in and of itself and the mask also tended to modify some of the changes associated with noise.

Word and segment durations tended to be longer when the speaker was wearing the mask. With the mask but in quiet, mean word duration was 775 msec; at 100 dB noise, it was 768 msec., effectively the same. The stressed first syllable was produced at a mean duration of 172 msec. in quiet, 173 msec. at 100 dB noise; the lengthened second syllable varied from 248 msec. in quiet to 259 msec. in noise. On the average, therefore, wearing the mask tended to cause the speaker to lengthen segments but noise exposure had no additional effect.

The same can be said of energy differences between the two syllables. The second syllable was produced 5 dB lower than the first in quiet, 4dB lower at 100 dB noise.

The average fundamental frequency for the stressed syllable in quiet was 129 Hz, almost 10 Hz lower than without the mask. Noise at 100 dB increased average F0 to 150 Hz, a value comparable to the increase without the mask. The unstressed second syllable was produced at a mean 109 Hz in quiet, 121 Hz at 100 dB noise, values comperable to speaking without the mask.

The vowel space associated with the oxygen mask is given in Fig. 3. The oxygen mask appears to have an effect similar to noise in that F2 tends to lower, particularly for front vowels relative to speech produced without the mask. Noise, however, does not seem to have any additional effects on vowel formants over those associated with the oxygen mask.

Loud speech. When asked to be deliberately loud, the speaker produced words with average vowel durations and fundamental frequency and amplitude values comparable to those characteristic of speech in noise, speech which might be characterized as unconsciously loud. Without the oxygen mask, 'loud' vowels in the two syllables were 181 msec. and 273 msec. in duration. The corresponding fundamental frequency values were 158 Hz. and 123 Hz. The second syllable was 6 dB lower than the first.

When wearing the oxygen mask while attempting to be loud, the speaker produced similar values: mean vowel durations were 170 and 240 msec.; mean F0 was 150 Hz and 119 Hz; the difference in engergy between the syllables, however, was only 3 dB.

The vowel space plot for loud speech is given in Fig. 4. The vowels of loud speech were very similar whether the speaker was wearing an oxygen mask or not. The vowels were shifted, however, from the values of quiet speech: F2 for front vowels lowered and F1 raised, particularly for high vowels.

DISCUSSION

Our primary observation is that the acoustic-phonetic structure of speech can be systematically affected by the physical environment under which it is produced. The observed changes can be correlated with the specific circumstances of speech production.

In order to maintain vision and consciousness at higher accelerations, so called anti-G maneuvers are necessary. These involve pulling the head down, tensing the skeletal and abdominal muscles as much as possible, and increasing intrathoracic pressure by forcibly exhaling against a partially or completely closed glottis. These straining maneuvers undoubtedly affect laryngeal tension and vocal tract configuration, and may be responsible for the changes observed in speech under acceleration. Increased laryngeal tension would be responsible for the observed increase in fundamental frequency. Tension in the pharyngeal region would tend to reduce tongue mobility, resulting in a decreased vowel space.

When speaking under high levels of noise, the speaker increased loudness (energy) and pitch (fundamental frequency). These same changes were associated with deliberately loud speech. The inference is that loud speech is the same, whether due to external physical circumstances or to speaker intent.

According to our subject, the oxygen mask restricts the mandible so that there is some resistance to jaw lowering. However, in previous work (Shulman, 1985), an increase in loudness was associated with a larger mouth opening and a raised F1. We would hypothesize that a speaker who is increasing the loudness of his speech and using a larger mouth opening would tend to shift the point of maximum constriction towards the back, raising F1 and lowering F2 for front vowels. When jaw movement is restricted by the mask, tongue mobility would decrease with approximately the same acoustic effects.

REFERENCES

Bond, Z. S., Moore, T. J., and Anerson, T. R., The effects of high sustained acceleration on the acoustic-phonetic structure of speech: a preliminary investigation. Armstron Aerospace Medical Research Laboratory Technical Report, AAMRL-TR-86-011, 1986.

Burton, R. R., Leverett, S. D., and Michaelson, E. D., Man at high sustained +Gz acceleration: a review, Aerospace Medicine 45 (1974) 1115-1136.

Picheny, M.A., Durlach, N.I., and Braida, L. D., Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech, Journal of Speech and Hearing Research 29 (1986) 434-445.

Pisoni, D.B., Bernack, R. H., Nusbaum, H. C., and Yuchtman, M., Some acoustic-phonetic correlates of speech produced in noise, IEEE International Conference on Acoustics, Speech, and Signal Processing, 1985.

Schulman, R., Dynamics and perceptual constraints of loud speech, paper presented at 110th meeting of the Acoustical Society of America, Nashville, Tennessee, 1985.

Zue, V. W., and Cyphers, D. S., The MIT SPIRE system, Proceedings of Speech Tech '85, New York: Media Dimensions, 277-279.

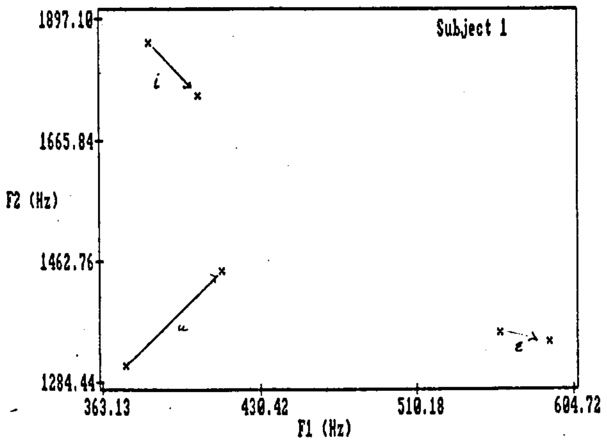


Fig. 1. F1 and F2 at 1G and 6Gz; the arrow points toward the 6Gz values.

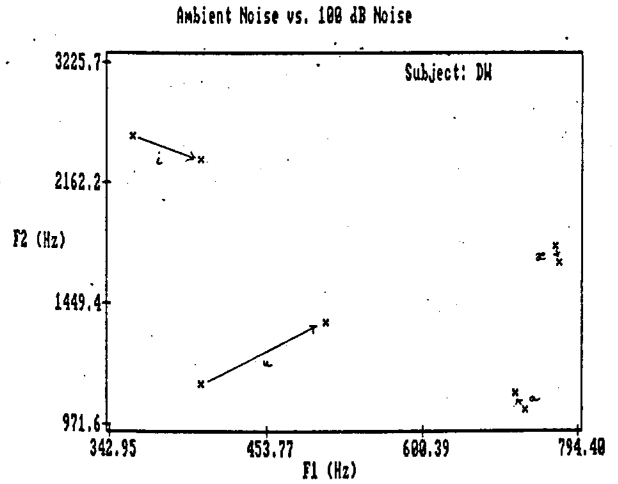


Fig. 2. Formant shifts associated with speaking in noise.

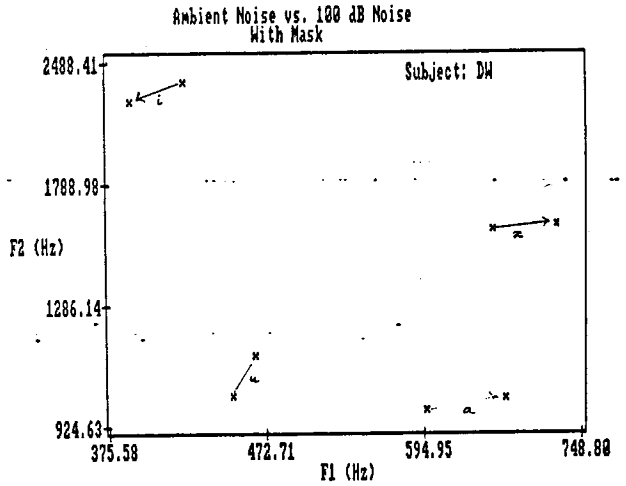


Fig. 3. Formant shifts associated with noise. The speaker is wearing an O_2 mask.

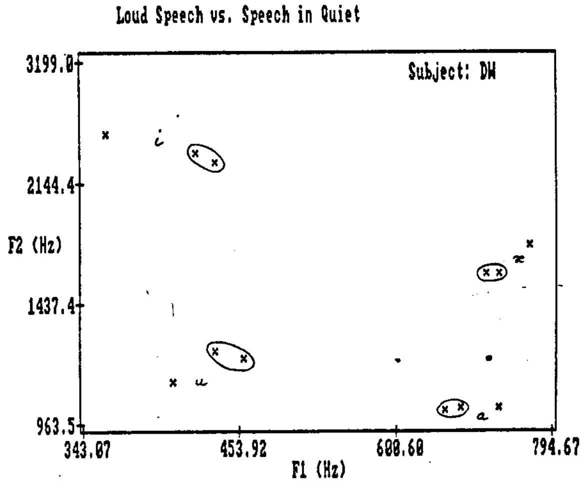


Fig. 4. Speech intended to be loud is enclosed in ellipses.

THE EVALUATION OF MISINTERPRETATIONS OF SPEECH SEGMENTS UNDER NOISE-TEST CONDITIONS

PREMYSL JANOTA
Department of Linguistics and Phonetics
Philosophical Faculty
Charles University
116 38 Prague 1, Czechoslovakia

ZDENA PALKOVÁ
Department of Linguistics and Phonetics
Philosophical Faculty
Charles University
116 38 Prague 1, Czechoslovakia

ABSTRACT

Two parallel tests containing the same language material and differing solely in the noise levels used were run using a fairly large number of listeners. The results obtained from speech segments listened to under noise-test conditions reveal on analysis a snift in values in the same general direction, differing mainly in terms of quantity.

The results appear to reveal trends in the behaviour of the individual phonic qualities the perception of which was constrained by noise. Analysis of erroneous identifications of speech segments obtained by the noise-test furnishes comparative material for further research into speech perception with special reference to automatic speech recognition.

INTRODUCTION

The present paper follows on in part from our experiences with testing speech signals masked by noise (cf. /1/; further literature ibid.), and in part from an unpublished research report concerning the discrimination of a limited set of comments, also masked by noise, given by a group of different speakers.

The investigation was conceived as a probe to contribute to the problem of automated speech recognition.

The material consisted of 20 one-word commands performed by 10 speakers, and it was masked by gradually increasing levels of white noise. The results obtained were assessed in terms of the words confused and in terms of the influence of the different speakers. The test was run in two variants, the testees previously knowing or not knowing the speech material to be used.

The results of the probe revealed significant differences in the degree of difficulty of the different items, while they also pointed to differences in the intelligibility of the speech of the different speakers.

In this paper what we describe are the results of a more extensive experiment which picks up the first range of results from that earlier probe, i.e. we are not looking for differences brought about by the pronounciational idiosyncracies of different speakers, but are using the interpretations of a single speaker whose pronounciation can be treated as standard in terms of orthoepy and speaking technique.

We concentrate on the analysis of errors by a largish number of listeners in understanding noise-masked speech material. In our analysis we not only note the degree of deformation, but also attempt to evaluate it qualitatively. The results are presented chiefly in the form of tables giving absolute figures and percentage relations; some of these are distributed as handouts.

THE EXPERIMENT

The material for the test consisted of a group of 100 Czech words selected according to preliminary criteria in such a way as to facilitate the composition of 5 relatively homogeneous subsets of 20 words each. The criteria used were as follows:

a) All the words were nouns in the nominative singular.

b) Their frequency lay in the 1000-10000 zone (see /2/).

c) The following were excluded: words of visibly foreign origin, emotionally laden words, specialist terms, proper names.

d) Each subset contained the same ratio of words classified by length in syllables, the words being of from one to five syllables in the ratio 5:7:6:1:1.

e) each subset contained repetitions, in individual words, of the same composition of VC elements. The basis for the selection of syllable structure types were the statistical data given in /3/. We sought to include all high-incidence types, but with some reduction in the use of the most frequent types CV and CVC.

f) Each subset contained a word in which the syllable peak was r or l.

g) Approximately half of the words used were more concrete in meaning, the other half being abstracts.

h) It was not possible to standardise phoneme frequency; however, comparison of the overall data with the relative frequencies for Czech /4/ revealed statistically significant agreement, as did comparison of the frequency of phoneme pairs across the subsets (their rank correlation).

The five 20-word subsets as realised by one speaker were ordered into a continuous test in which the speech signal was masked by noise which was stepped up between the separate parts of the test. Two variants of the test were run, using different steps in the noise level. (Variant A: - 40, - 15, - 9, - 3, + 3dB and variant B: - 9, - 6, - 3, 0, + 3dB).

Both variants of the test were given to listeners whose native language was Czech (100 testees, all students registered for modern language courses in their first and second years at the Philosophical Faculty of Charles University).

The performance of each testee was assessed by data on the total number of wrong answers. Comparison of the results within each group showed the normal distribution (χ^2 - test, 5%). The average result in test A : $\bar{x}$ = 17.8, s = 3.49; in test B: $\bar{x}$ = 28.3, s = 4.04.

Comparison of the frequency of mistakes with individual words reveals that the result is influenced by two factors above all: the level of noise, and the individual characteristics of the different words.

RESULTS OF THE EXPERIMENT

Results acquired on the basis of the overall data of the number of errors at different noise levels can be summarized as follows:

a) The degree of difficulty of variants A and B was different. The influence of different steps in the noise level confirmed our assumptions. In test A the first two sections were error-free, while in test B errors were distributed throughout.

b) The underlying tendency for errors to increase within the classification used in the table remains essentially the same, irrespective of differences in the difficulty of the test.

c) The stability of words proved dependent on the number of syllables: the longer the word, the lower the number of wrong answers. The highest percentage of errors is within monosvllables.

d) The number of syllables proved to be a relatively stable attribute of a word. Errors of syllable number ammount to only 1% of responses in test A and 2% in test B. The dependence of errors in syllable numbers on words-length does not share the tendency noted in c). Results for individual groups of words according to the number of syllables are fairly evenly balanced, the least stable words being disyllables (see in particular test B).

e) Failure of testees to respond at all ("0-judgments") is also not directly dependent on word-length; at higher noise levels the testees resorted to this solution more frequently with di- and trisyllables than with monosyllables.

f) The link between a word's stability and its length comes out most strongly in the section, giving the number of syllables remaining the same.

The set of errors where at least the number of syllables was preserved in both test A and test B was submitted to further analysis in terms of their phoneme composition.

Results obtained from analysis of vowel switches

a) Under test conditions vowels remain fairly stable. Of the mis-heard words (with the right number of syllables) less than half have the error in a vowel. In test A-2 the figure is 36%. The higher percentage in test A-1 is due to the single figure of the higher number of errors in the third syllable of trisyllabic words; in test A-2 this tendency does not reappear. The causes would appear to do with something other than sound; it concerns just one word in each column: pracovna - pracovník, horlivost - horlivec.

b) Errors in quantity are less frequent than changes in the quality of a vowel.

c) The vowel in monosyllables is conspicuously stable.

d) It may be similarly assumed that the first syllable of polysyllables will have its vowel better preserved than those in the other syllables. This tendency is indeed strong in trisyllables. In disyllables in test A-2 the ratio of errors in the two syllables is fairly evenly balanced. The reason is the high number of errors in the first syllable of the word in column 13 of the test. Once more the result is based on confusion in two words only, but this time there can be no doubting the influence of sound factors. The cases are confusions of důkaz - výtah (56 out of 83 errors)) and příval - úval (29 out of 61 errors). Insofar as there is a tendency for greater stability in the first syllable of disyllables, it is not so strong as to outweigh other phonic properties of the word.

Mutual substitutions of vowels separately

a) The direction of substitution seems not to be arbitrary since there are some discernible tendencies.

However, in interpreting the results consideration has to be given to those cases where there is a high incidence of substitution in one word and where the motivation may be other than phonic (most often it is conditioned morphologically). These are the cases of the above-mentioned substitution if the ending pracovna - pracovník (46 instances a - í in test A-1). Similar cases znalec - znalost (42 instances of e - o) and horlivost - horlivec (45 instances of e - e) may be explained as changes of grammatical morphemes as well, but a strikingly similar tendency of this vowel substitution may be pointed out in test A-2, in which a possible influence of a morpheme change is not probable.

b) The vowel a appears to be relatively stable, especially long á. By contrast most errors affected the vowels í and ú. These two vowels showed a tendency to mutual substitution in the material. Interchange between a and o is also relatively frequent. Syllabic l tends to survive better than syllabic r.

Results obtained by analysis of mis-heard consonants.

The analysis of mistakes affecting consonants and consonantal clusters was also carried out on the basis of the set of mis-heard words where the number of syllables was preserved. Consonants have not yet been looked at individually, the overall picture of substitutions having been worked out with respect to certain pre-stated types of errors.

6 basic types of change were distinguished.

x_1 - simplification of consonantal clusters by the loss of one or more consonants (e.g. for vzdech - vdech or dech);

x_2 - loss of a consonant or consonants, the consequence of which is the loss of the consonantal element in the given position altogether (e.g. for dozor - ozón, for vzdech - zde);

x_3 - addition of a consonant or consonants where there was already, i.e. creation or expansion of a consonantal cluster (e.g. for jih - mnich, for vzdech - vzhled);

x_4 - addition of a consonant where no consonant existed before (e.g. for ořech - konec, for mluva - průvan);

x_5 - simple substitution of a single consonant (e.g. for jih - nit', for střed - střep);

x_6 - substitution of an entire consonantal cluster, or one of its elements with retention of the right number of elements in the cluster (e.g. for zřetel - dveře, for blesk - vlek);

x_7 - syllables with changed open/closed character.

Thus in processing the results we also distinguished positions of consonantal elements before and after a vowel, for various reasons including the information which this offered on the change in the character of a syllable in terms of its being open or closed.

To obtain more telling values for comparison of the obtained frequencies, the following characteristics were added:

y_1 - number of correctly heard consonants in erroneously received words;

y_2 - number of correctly heard consonants clusters in erronesouly received words;

y_3 - number of syllables with retained open/closed character in erroneously received words;

y_4 - sums of erroneous words with retained numbers of syllables;

y_5 - sums of erroneous words with retained numbers of syllables.

On the basis of interpretations made to date the following may be stated:

a) As expected, the number of mis-heard consonants is conspicuously higher than the figure for vowels. Among the mis-heard words with the right number of syllables retained erroneous identification of consonants ammounts to 65% in test A-1 and 75% in test A-2. Relating these erroneous identifictions to the simple total of mis-heard words this represents 175.6% in test A-1 and 195.2% in test A-2, i.e. approximatively two errors per word on average affect consonants.

b) By contrast with the foregoing, the character of the syllable as closed or open proves a highly stable property. Error frequency of this type is lower than the frequency of wrong vowels. The results of the two tests are very evenly balanced, whether the ratio of wrong and right identifications (A-1: 0.14, A-2: 0.13) or the relation to the number of wrong words (A-1: 20.6%, A-2: 18.6%) is used as a characteristic.

The analyses show further the need to distinguish the position of the syllable in the word. Of particular stability are monosyllables and the first syllables of polysyllables. On the contrary, endings preserve the character of a syllable to a considerably lesser degree. Again the reasons may be other than phonic. This may have something to do with the fact that the destruction of the character of a syllable at the different noise levels does not rise in proportion to difficulty in listening, but peaks in both tests at the penultimate level.

c) Comparison of the right and wrong interpretation of the test items reveals a clear tendency for a consonant to be more stable before a vowel than after one. This tendency is observed at all noise levels and applies to words of different length. However, it is not tied more to the first syllable. In the material given, the most stable consonant or consonantal element is at the head of the second syllable of di- and trisyllables.

With more detailed processing of the results of the test, there are a number of further tendencies discernible, some also to do with the actual nature of specific substitutions.

For example, with deformation of word-initial consonant clusters that consonant which immediately precedes the vowel is often preserved, e.g. in test A-2 the ratio of wrong and right solutions for consonants in the first syllable immediately preceding the vowel is 1.5 (for the whole consonantal unit in this position it is 3.76); a similar tendency is found in monosyllables, in both tests moreover.

In the material used the nasal consonants proved relatively stable, and it is precisely their nasality which survives. The commonest error with nasals is their substitution by a different nasal consonant. For example, the ratio of wrong and right solutions in test A-1 at the most difficult noise level is 0.63 (i.e. correct responses are in the majority), and if we take as correct also those cases where the substitute was also nasal, the value drops to 0.17; similarly for test A-2 a ratio of 1.06 drops to 0.31.

It is expected that additional modifications will be made to the parameters used in order to ascertain more exactly which of the phenomena discovered contribute effectively to the identification of speech.

REFERENCES

/1/ P.Janota, Z.Palková: Testing Perceptive and Productive Skills in Language Learning, AUC, Phonetica Pragensia V. 1976, 15-28

/2/ J.Jelínek, V.Bečka, M. Těšitelová: Frekvence slov, slovních druhů a tvarů v českém jazyce, Praha 1961

/3/ H. Kučera, G.H.Monroe: A Comparative Quantitative Phonology of Russian, Czech and German, New York 1968

/4/ M.Ludvíková, J. Kraus: Kvantitativní vlastnosti soustavy českých fonémů, Slovo a slovesnost 27,1966, 334-344

Categorical Perception and Difference Limens in Helium-Oxygen Speech

Kolbjørn Slethei

Department of Linguistics and Phonetics
University of Bergen
N-5007 Bergen

ABSTRACTS

Speech distorted by helium in the breathing gas, as is the case in saturation diving at depths below 50 msw, is rendered unintelligible by an upward frequency shift of spectral components.

When responents label spectras of linear interpolations between /i:/ and /e:/ spoken in air, we get a categorical transition in the vowel continuum. Interpolation spectras from 54, 120 and 300 msw are being categorised with decreasing accuracy, while the respondents' ability to label appropriately increases from 300 to 500 msw.

The difference limens (DL) for F1 and F2 for the vowel /i:/ have been investigated for the same depths. DL for F1 remains relatively stable, with a rise from 300 to 500 msw. DL for F2 is raised from 0 to 300 msw, and lowered from 300 to 500 mws.

These findings will be discussed.

INTRODUCTION

In saturation diving the nitrogen and most of the oxygen in the breathing gas is replaced by helium. Table 1 lists typical compositions of breathing gases at various depths, in this case from an experimental dive in pressure chambers at The Norwegian Underwater Technology Center (NUTEC) in Bergen, Norway. (The small quantity of N_2 is a consequence of an unintentional contamination.)

	Depth (msw): 0:	54:	120:	300:	500:
Pressure in atm.:	1	6.4	13.0	31.0	51.0
O_2 in %:	air	8.6	4.1	1.6	0.9
N_2 in %:	air	0.1	3.4	1.6	0.0
He in %:	air	90.4	92.6	96.9	99.1

Table 1. Contents of breathing gases at various depths /1/.

Intelligibility can be described by the output from a Modified Rhyme Test (MRT), where intelligibility is the percentage of correct identifications in a multiple forced choice test, adjusted to correct for the potential guesswork involved.

Table 2 list typical MRT-scores for speech at depths between 0 and 500 msw.

	Depth (msw): 0:	100:	200:	300:	400:	500:
MRT-score:	97	56	50	46	42	47

Table 2. Intelligibility as a function of depth. Data modified from Slethei /2/.

The decrease in MRT-score is mainly caused by the helium, but the increase in ambient pressure contributes to the effect /3/. (Rank-order correlation between MRT-score and the proportion of helium is -0.99, between MRT-score and ambient pressure -0.77, based on the data in Tables 1 and 2.)

The loss in intelligibility is small from 200 to 400 msw, and from 400 to 500 there is even an increase. This flattening and rise in MRT-score cannot be accounted for by changes in depth, ambient pressure or composition of the breathing gas. A somewhat more detailed analysis seems to be needed.

It should be borne in mind that the MRT-score may disguise differences that are pertinent to the understanding of real speech, because the vowel phoneme is the same for all words that are candidates for the respondents' best forced choice. This might suggest that studies of the perception of vowels could shed some more light into the auditory darkness at depths below 200 meters.

In order to approach some of the problems related to the perception of helium speech, we have made two studies; one deals with categorical perception of vowels in helium-oxygen speech (Part I), the other deals with how difference limens (DLs) behave in this breathing gas (Part II). DLs will be studied for F1 and F2 separately.

At the time of finishing this paper, both parts comprise data from 15 respondents with no prior experience with helium-oxygen speech. Both studies will be extended to 20 respondents.

METHOD.

Part I: Formant parameters (F1-F4, BW1-BW4) for the Norwegian vowels /i:/ and /e:/ spoken in air /4/ by one diver were used as end point values, and 18 linear interpolations were calculated. All the 20 vowels were synthesized by a LPC-based formant cascade synthesizer. The same procedure was repeated for the same vowels spoken in atmospheres for 54, 120, 300 and 500 msw.

Each of the 5 sets was headed by the /i:/-/e:/-pair 3 times to serve as anchoring points for the identification tasks. Each set was randomized individually. These 5 sets, together with pauses and some sinusoid control signals, were DA-converted directly onto analog audio tape.

The stimuli were presented to the respondents via earphones, and the respondents were asked to tick off their best identification as either /i:/ or /e:/ for each stimulus. The empirical material for Part I thus consists so far of 1500 individual and independent data points.

Part II. F1 for the vowel /i:/ spoken in air was varied with respect to frequency and bandwidth and pairs of vowel-like stimuli were produced. Stimulus pairs were organised to fit into an AX-paradigm, where F1 frequency for the X spectrum was varied from 2% to 12% above that of the A spectrum in a cumulative manner. X_1 has a first formant frequency 2% above A, X_2 has a first formant frequency 2% above that of X_1 and so on. AX_1 and AX_2 would then constitute two different pairs of vowel stimuli.

This procedure was carried out for all depths for F1 and F2.

The test material consisted of 800 vowel pairs. They were presented to the respondents via earphones, and the respondents were asked to determine whether A and X were identical or different in quality by ticking off appropriate boxes on a response sheet. The empirical material for Part II consists of 12000 individual and independent data points.

RESULTS.

Part I. Table 3 presents the results for Part I. Stimulus No 1 is the end point /i:/ and No 20 is the end point /e:/.

Stim. No:	Number of respondents identifying S as /i:/ for each depth: 0:	54:	120:	300:	500:
1	15*	15*	7	5	11
2	15*	12*	4	11	11
3	15*	13*	4	8	13
4	15*	11	3	5	10
5	15*	13*	3	7	9
6	15*	11	3	7	8
7	10	8	3	6	3
8	14*	13*	7	7	8
9	6	10	4	5	7
10	2*	10	1*	9	4
11	1*	2*	3*	7	5
12	1*	3*	4	5	3*
13	2*	3*	3*	3*	1*
14	0*	1*	2*	11	5
15	0*	2*	4	10	4
16	0*	1*	2*	7	5
17	0*	1*	5	3*	3*
18	0*	0*	1*	4	1*
19	0*	0*	1*	7	1*
20	0*	0*	3*	7	4

Table 3. Number of respondents identifying stimuli 1-20 (/i:/-/e:/) as /i:/ for the depths 0 ,54, 120, 300 and 500 msw. Significant identifications as N* .

In Table 3 we have indicated which of the identifications were statistically significant according to the binomial distribution (alpha=0.05). A non-directional hypothesis was considered for depth = 0, and a directional hypothesis for depth > 0. In the stimulus continuum, the turnover point (between S 7 and S 8) was used as dividing point for directing the null-hypothesis towards /i:/ or /e:/.

From Table 3 we can calculate the number of statistically significant identifications for each depth. The results from this calculus are presented in Figure 1.

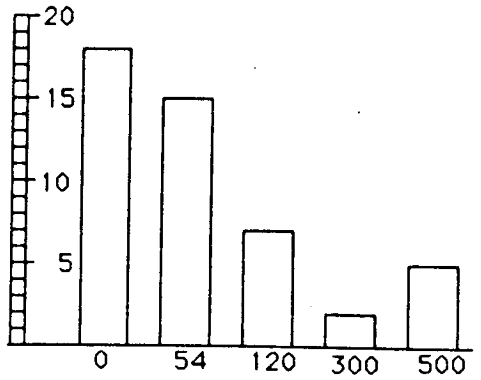


Figure 1. Number of statistically significant identifications in vowel stimulus continuum per depth.

If we disregard 0 msw and consider the cases where a stimulus is correctly identified as either /i:/ or /e:/, i.e. when stimulus number is either < 8 or stimulus number is > 7, we find that only 4 out of 28 stimuli have been identified correctly as /i:/, while 25 out of 52 have been correctly identified as /e:/. Testing these proportions against an expected equal proportion, we find that the difference is statistically significant. (Chi-square goodness of fit, with expected equal proportions.)

Part II. In an AX-paradigm, the DL is defined as the minimally detectable difference between A and X. For 15 respondents, we can reject a hypothesis that a threshold has been detected erraneously if 11 out of 15 responents agree that A and X are different. (Binomial distribution, nondirectional, alpha = 0.05.)

Table 4 lists the DLs as the mean percentage of difference between A and X, and the typical formant frequency when the difference has been detected by 11 respondents. The typical formant frequency is the mean of the A and X values.

Formant:	Depth:	DL:	SD:	Typical frequency:
F1:	0	10.7	*	366
	54	11.2	1.2	786
	120	10.4	*	1225
	300	10.4	3.1	1597
	500	15.0	4.6	1917
F2:				
	0	9.9	2.7	2451
	54	12.7	3.9	4823
	120	14.9	3.2	5237
	300	22.0	6.8	5974
	500	18.8	5.4	6185

Table 4. DL and standard deviations as a function of depth. (* indicate insufficient data.) Typical formant frequencies when DL is detected.

The decrease in DL for F2 from 300 to 500 msw becomes more apparent when presented in graphical form in Figure 2.

DISCUSSION

Part I. The labelling tasks performed in Part I clearly demonstrate that vowels simulating helium speech spoken in isolation differ in difficulty as objects for labelling. The Modifyed Rhyme Test disguises this difference. This calls for developing descriptive techniques which combine the reliability of the MRT with an ability to exploit variation within the linguistic material. MRT is only able to differentiate between those combinations of VC- and CV-structures that are included in the finite set of response words.

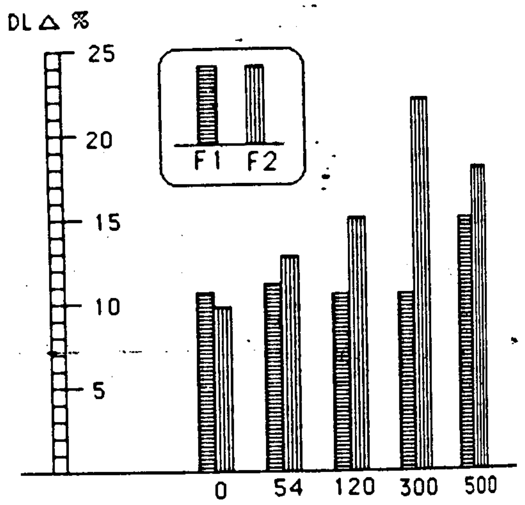


Figure 2. Difference limens for F1 and F2 per depth.

Algorithms aiming at reconstructing the speech signal as it would have been in air at 1 atmosphere, take the physical properties of the breathing gas into consideration. This is of course necessary, but auditory and perceptual aspects are being neglected. The upward shift in frequencies causes the formants to be moved out of the auditory region where pitch resolution is optimal for speech perception. For the vowels /i:/ and /e:/ this means that the second formant gradually loses its importance as cue for identification, until the first formant has been enabled to take its place. This is the most probable explanation for the effect shown in Table 4 and in Figure 1.

Part II: Flanagan's findings /5/ of DL for vowel formant frequencies in the region of 3-5% have recently been questioned by Ghitza and Goldstein /6/, who report DLs in the region above 12%. Our findings are largely in accordance with those of Ghitza and Goldstein, with an increase for F1 in the region above 1.5 kHz.

The considerable increase in DL for F2 from 0 to 300 mws (Table 5 and Figure 2) is only to be expected. The decrease from 300 to 500 mws is unexpected. Although the standard deviations are uncomfortably large, it is worth while asking why there is such a decrease.

The same explanation may suffice here. The frequency of F1 increases in relative importance as a cue to detecting the DL for F2 as the F2 frequency is moved out of the region of optimal pitch resolution.

CONCLUSIONS

Developing instruments for improving the efficiency of communication system is a demanding task, where the knowledge and skills from various professions may contribute.

There is a considerable room for improving the methods and techniques which describe the efficiency of such systems.

Knowledge from the fields of auditory and perceptual phonetics may contribute to the development of instrumentation for communication systems.

ACKNOWLEDGEMENTS

The Norwegian Underwater Technology Center (NUTEC) has given valuable assistance in the research reported here.

Instituut voor Perceptie Onderzoek (IPO) in Eindhoven generously allowed my to use their computer facilities to produce the speech synthesis.

REFERENCES

1. Belcher, E.O. and Hatlestad,S. (1983): Formant frequencies, bandwidths and Qs in helium speech. Journal of the Acoustical Society of America, 72(2): 428-432.
2. Slethei,K. (1984): Perception of Speech in a Hyperbaric Helium-Oxygen Atmosphere. Van den Broecke, M.P.R. and Cohen,A. (eds.): Proceedings of the Tenth International Congress of Phonetic Sciences, 438-442. Foris Publications.
3. Fant,G. (1980):Tal i vätgas- och heliumblandingar vid höga tryck - en teoretisk studie. Internal Report, Dept. of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.
4. Belcher,E.O. and Hatlestad,S.(1982): Analysis of isolated vowels in helium speech. NUTEC Report No 26/82, Norwegian Underwater Technology Center, Bergen.
5. Flanagan,J.L. (1955): A difference limen for vowel formant frequencies. Journal of the Acoustical Society of America, 27:1223-1225.
6. Ghitza, O. and Goldstein,J.L. (1986): Scalar LPC Quantization Based on Formant JND's. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34:697-708.

ANALYSE DE LA COMMUNICATION VERBALE DANS DIFFÉRENTS MILIEUX DE TRAVAIL - PROBLÈMES DE LA PERCEPTION

MARIE DOHALSKÁ-ZICHOVÁ

Institut de Phenétique
Université Charles
116 38 Prague

RÉSUMÉ

Problème de la perception de la parole dans de mauvaises conditions acoustiques - communication verbale dans l´organisation moderne du travail dans les différentes branches de l´industrie, du transport, etc. Facteurs influençant la qualité de compréhension.

L´emploi de nouvelles techniques modernes dans l´organisation et la gestion de l´industrie, des mines, du transport et d´autres branches de l´économie nationale nous a permis d´observer que la parole y joue un rôle de plus en plus important, car le message verbal représente aujourd´hui non seulement un instrument de communication, mais il est devenu un instrument de travail. Tenant compte de cette situation il est important d´étudier les conditions et les circonstances qui favorisent l´optimalisation de la qualité de compréhension dans tous les domaines de l´économie nationale.

Pour qu´un message verbal important (ayant sa fonction dans l´organisation du travail) accomplisse sérieusement sa fonction informative, il est nécessaire que tous ses composants soient parfaitement compréhensibles pour chaque membre du collectif sans laisser la moindre ambiguïté, aussi négligeable qu´elle puisse être. Une information peu explicite ou déformée risque de troubler non seulement les conditions de travail, mais aussi d´engendrer des accidents de travail.

De ce point de vue, le sujet parlant doit tenir compte à la fois de la capacité de compréhension de l´auditeur, de sa rapidité à fixer l information et enfin à réagir, en fonction du signal émis, pour accomplir un acte de travail adéquat. Il est toujours nécessaire de savoir comment et en combien de temps une information peut être reçue.

Pour pouvoir classer les différents types de communication nous avons rassemblé des échantillons représentatifs de la communication verbale des chemins de fer, du transport urbain et aérien, des mines, de la haute métallurgie et de l´agriculture. Le matériel obtenu est assez riche et en même temps hétérogène - ce qui répond à l´hypothèse préliminaire d´un degré différent des déformations phonétiques de ce genre de communication.

Nous nous sommes posés la question jusqu´à quelle mésure il serait possible de saisir certains types de déformations du signal acoustique pour en faire une classification. En tout cas, les déformations du signal acoustique ne peuvent pas être étudiées d´une façon isolée, il est nécessaire de les suivre non seulement d´après leur position dans les groupes rythmiques, mais aussi d´après le caractère complexe de l´énoncé (le débit, la durée, le genre de l´information, etc.). Chaque type de P.C. (poste de commande - c.-à-d. "dispatching") a sa façon spécifique pour transmettre les messages verbaux et par conséquent il faut se rendre compte qu´il y a même une différente quantité de l´information dans les messages "simples" (c´est-à-dire instructions ne contenant pratiquement que des mots-clé) par rapport aux phrases "complexes" dans lesquelles déjà le contexte et la structure grammaticale jouent un rôle important.

Les premiers sondages ont montré qu´il existe différents types de déformations des éléments consonantiques, vocaliques, syllabiques ainsi que les déformations des structures rythmiques de la phrase. Dès le début de nos recherches, nous avons réalisé plusieurs tests différents et nous avons comparé la façon de perception des professionnels - employés dans le milieu de travail donné - à celle des prophanes qui ne connaissent pas les "mécanismes de compensation" typiques pour chaque profession. Pour illustrer certains problèmes importants, nous présentons brièvement quelques résultats du test suivant.

D´après les recherches préliminaires nous avons préparé un test pour les allocutaires prophanes (25 étudiants, philologues, 18-20 ans). Dans ce test, notre attention a porté non seulement sur l´intelligibilité générale, mais tout particulièrement sur l´intelligibilité des indications numériques. Nous avons donc procédé à l´analyse de nos matériaux de manière à obtenir un aperçu spécial de la déformation des indications numériques qui sont en général dans tous les types de P.C. assez nombreuses et leur perception exacte se montre très importante.

Notre test a été composé de 178 segments dont:

8	dans l´intervalle	0 - 0,9 s.
47	dans l´intervalle	1 - 1,9 s.
62	dans l´intervalle	2 - 2,9 s.
40	dans l´intervalle	3 - 3,9 s.
18	dans l´intervalle	4 - 4,9 s.
3	dans l´intervalle	5 - 5,9 s.

L´intervalle le plus favorable pour la perception s´est montré entre 1s.- 4s. (ce qui représente 83,7% de tous les segments testés). Les intervalles au dessous d´une seconde, nous les avons choisis seulement pour savoir de quelle maniere l´allocutaire serait capable de percevoir des segments tres brefs, et au contraire, ceux qui dépassaient 4s.- - pour nous rendre compte quelle sarait le nombre limitatif de données perçues par les allocutaires.

Les différences dans les résultats se sont montrées d´abord dans les réactions des allocutaires

a) d´après le type de P.C.;
b) d´après la composition syntaxique du segment;
c) d´après la perception des mots-clé.

Les 178 segments contenaient au total 471 indications numériques, tels qu´on les trouve dans les phrases spontannées des P.C. . De ce nombre de 178 segments 14 segments seulement ont été perçus avec une intelligibilité de 100% (dont 6 ne contenant que des indications numériques, 3 segments "combinés" et 5 segments contenant des indications non-numériques).

Nous pouvons résumer les résultats de ce test spécifique dans les points suivants:

1) Le degré de l´intelligibilité dépend d´abord de la précision de l´articulation, du rythme et du débit de locuteur et en même temps de la composition syntaxique des segments. Ces facteurs se montrent encore plus importants au moment où le bruit général du milieu de travail s´accroît ou si la qualité de transmission du signal acoustique diminue.
Si le locuteur prononce son message distinctement en respectant la répartition rythmique de l´énoncé, l´intelligibilité est suffisante même dans un bruit assez important.

2) Pour la communication dans de mauvaises conditions acoustiques il s´avère que ce sont les voyelles qui représentent les éléments portant l´information. Nous nous sommes rendus compte qu´il existe souvent une certaine "matrice des voyelles" qui reste assez stable.
P.ex.: l´indication numérique "522" a été perçue comme: "222", "2,5,22", "9,22";

"522" /pjetsetdvacetdva/ e-e-a-e-a
"222" /dvjestédvacetdva/ e-e-a-e-a
"2,5,22" /dvjepjetdvacetdva/e-e-a-e-a
"9,22" /devjetdvacetdva/ e-e-a-e-a

Tous les changements indiqués sont dans ce type de P.C. théoriquement possibles, car ils y donnent un sens réel.
Si un changement des voyelles se manifestait nous avons examiné, quel type de changement se montrait comme typique, et au contraire, quel type nous pouvions considérer plutôt en tant que périphérique.

3) La durée des messages bien intelligibles est de 3s. environ (même s´il y a des différences dans la structure des messages). Les mots initiaux et finales sont aussi bien intelligibles que tous les autres (excepté les mots monosyllabiques commençant par les sifflantes au début d´un message ou des mots très longs à la fin d´un message dépassant 6s.).

4) L´intelligibilité des "notions-codes" au début ou à la fin d´un message représente un problème spécial.

Nous présumons qu´il est important d´enregistrer dans les analyses préliminaires tous les types de changements, même ceux qui nous paraissent marginaux. Dans la communication courante, l´allocutaire "complète" automatiquement les différentes informations grâce à la connaissance du milieu du travail, de l´ensemble des mécanismes de compensation. Mais dans une communication professionnelle qui dirige des systèmes du transport ou de la production importants, il est nécessaire de demander de telles conditions, dans lesquelles l´intelligibilité soit réellement de 100%.

ACOUSTICS AND PERCEPTION OF SPEECH IN VARIOUS MODES OF ARTICULATION

B.M.Kolesnikov L.M.Zakharov

Dept. of Philology Moscow State University
Moscow, USSR, 119899

ABSTRACT

This paper presents preliminary data and observations from research on acoustic modifications of speech in various modes of articulation. We consider acoustic variables of speech within a general model of speech production in which the mode of articulation (MA) is an independent source of acoustic change. Intelligibility scores of the word-lists heard under conditions of noise differ due to the degree of physical manifestation of phonetic features which is the highest in forced speech and the lowest in sloppy speech.

INTRODUCTION

Quite a number of papers have recently been published on acoustic properties of speech and its perception. The bulk of this research deals with speech in a comparatively narrow range of acoustic change. However, speech often may differ acoustically due to external conditions of communication or to the internal state of the speaker. One example is forced speech (FS), speech which is produced in the mode of forced articulation. As a rule FS is louder than normal speech (NS) and therefore less subject to distortions and more intelligible due to its more effective use of the hearer's attention. FS occurs in many situations of everyday life and is a universal means of overcoming distance, ambiant noise, an interlocutor's dullness or a child's disobedience (in the latter two cases FS is not logically motivated and has negative emotional connotations).

Another modification of speech has not yet been sufficiently analysed. It is the speech produced in a state of extreme weariness or intoxication. This variety of speech with slurred articulation is commonly called sloppy speech (SS). It's acoustics are markedly different from those of normal speech (NS) and forced speech. We distinguish three modes of articulation (MA) and three corresponding modifications of speech: normal, forced and sloppy respectively.

METHOD

In our experiments acoustic peculiarities and perception of FS and SS as compared with NS have been the main object of interest. Word-lists of I4 words have been read by 6 speakers possessing certain dramatic skills. Speakers were asked to imagine a situation where it was necessary to "out-voice" ambient noise or imitate the speech of an operator exhausted by 48 sleepless hours.

Recordings thus obtained were presented to a group of subjects who were asked to describe the speaker's condition and any peculiarities of his speech. Here follow some examples of such descriptions: "neutral speech" (about normal articulation), "speech most likely produced at a meeting (about forced articulation), "indifference, or rather weariness near to drowsiness" (about slurred articulation). The recordings were used for further acoustic analysis.

The sonagrams of the recorded stimuli were made by means of the "Kay Elemetrics" Sona-graph. The measurements of fundamental frequency, duration and spectral characteristics of the stimuli were made on the sonagrams. The main data are presented in the following tables.

In table I FS and SS are compared with NS. A plus-sign stands for increased measured duration of a segment in various MAs as compared with NS, a minus-sign stands for decreased measured duration and ∅ stands for equality of measured duration.

The table shows that while a word in general becomes longer in FS, stressed vowels and to a lesser extent preceding consonants are consistently lengthened. The remaining segments (unstressed vowels and other consonants are not necessarily lengthened. Other consonants are lengthened 55% of the time for speaker I and 82% of the time for speaker II. Unstressed vowels are lengthened 64% of the time for speaker I and 86% of the time for speaker II.

In SS the character of duration change is different for both speakers. Thus speaker II lengthens half of all the words, and in only 29% of these cases lengthens vowels stressed or unstressed. Speaker I, who lengthens 40% of the words, constantly makes this by lengthening stressed vowels. He lengthens unstressed vowels only 62% of the time. He does not lengthen preceding consonants. However, consonants other than preceding ones are lengthened more consistently - I00% of the time for speaker I and 9I% of the time for speaker II.

The tendency to lengthen consonants was confirmed in experiments on a larger scale employing

5 and 6 speakers and a greater number of words. The consonants in FS do not always increase in duration in unstressed syllables. Lengthening occurs as follows: sonants - 42%, fricatives - 13%, plosives and affricates - 59% as compared to their duration in NS. In FS for consonants preceding stressed vowels lengthening occurs thus: sonants - 61%, fricatives - 43%, plosives and affricates - 52%.

Table I. Duration change in FS and SS as compared with NS

Stimulus word	speaker	forced speech: general	stressed vowel	unstressed vowel	preceding consonant	other consonants	sloppy speech: general	stressed vowel	unstressed vowel	preceding consonant	other consonants
ia-ma	II	+	+	+	+	−	+	−	ø	ø	+
	I	+	+	+	+	+	+	+	−	−	+
azu-	II	+	+	+	+		+	+	+	−	
	I	+	+	+	+		+	+	−	−	
a-lpha	II	+	+	+		−	+	−	+	+	+
	I	+	+	+		+	+	+	+	+	+
uzda-	II	+	+	+	+		+	+		−	
	I	+	+	−	+		+	+		−	
i-m'a	II	+	+	+		+	−	−	−		+
	I	+	+	+		ø	+	+	−		+
uzhe-	II	+	+	+	+		−	+	+	−	
	I	+	+	+	+		+	+	+	+	
e-ta	II	+	+	+		+	−	−	−		+
	I	+	+	+		ø	+	+	+		+
ke-pka	II	+	+	+	(ø)	+	−	−	−	(+)	+
	I	+	+	−	(+)	+	+	+	+	(ø)	+
i-kry	II	+	+	+		+	+	−	+		+
	I	+	+	+		+	+	+	+		+
t'o-sh'a	II	+	+	+	(+)	+	+	−	−	(+)	+
	I	ø	+	−	(+)	+	+	+	+	(+)	+
kho-lodno	II	+	+	−		+	−	−	−		−
	I	+	+	−		+	+	+	−		+
za-ponki	II	+	+	+	+	+	−	+	−	−	+
	I	+	+	+	+	+	+	+	−	+	+
pała-tka	II	+	+	+	+	+	−	ø	−	ø	−
	I	+	+	−	+	−	+	+	+	ø	+
analgi-n	II	+	+	+	ø	+	+	+	−	+	+
	I	+	+	+	ø	−	+	+	+	ø	+

In SS consonants other than preceding ones tend to lengthen thus: sonants - 69%, fricatives - 67%, plosives and affricates - 76%. The preceding sonants in SS tend to maintain their duration or shorten in the following manner: sonants - 57%, fricatives - 69%, plosives and affricates - 48%.

Table 2 shows that FS is characterized by lengthening of vowels. However, such lengthening depends on a number of circumstances. Thus for example only 4 out of 6 speakers lengthen stressed vowels. One speaker might shorten all the vowels, whereas another might lengthen the stressed /u/ and /i/ and shorten all the rest. One circumstance might be the tempo chosen by a speaker - vowel lengthening is characteristic of slower tempo. Another might be the degree to which articulation is forced. Curiously enough speakers I and IV are the same person. The second set of his recordings were made after an interval of 6 months. The data are totally different: the first recordings display shortening of almost all vowels, but the second recordings display lengthening of the stressed vowels. It is worth mentioning that the tempo in the first case was slower than in the second. Five out of six speakers lengthen the stressed /i/ and /u/ to a greater degree than /a/, /o/, /e/ in FS.

As far as the unstressed vowels are concerned, there are not consistent differences between various MAs. Only one speaker out of six constantly lengthened unstressed vowels and only one of six constantly shortened the unstressed vowels (the same speaker also shortened the stressed vowels).

The duration of unstressed vowels was not found to depend on their identity, position relative to stress or consonant environment. The duration change of unstressed vowels did not reveal any evident regularity.

The SS is characterized by an irregular and individualized manner of vowel duration change. 2 speakers (I and III) tend to lengthen almost all vowels, three speakers (II, V, VI) tend to shorten almost all vowels. One speaker lengthens almost all stressed vowels and shortens almost all unstressed vowels. Moreover, the character of vowel duration change in SS does not depend on the speaker's chosen tempo. The duration of all vowels was not found to depend on their identity, position relative to stress or consonant environment.

One can point out two contrasting tendencies in intensity characteristic of FS as compared with NS.

I. Greater prominence of the stressed vowel (the difference between maximum intensity (Imax) for stressed and unstressed vowels being less in NS). When forcing is very strong a speaker just "shouts out" the stressed syllable while the rest of the word is almost inaudible.

2. The levelling of intensities of stressed and unstressed vowels (the difference between Imax of the stressed vowel and Imax of the unstressed ones in FS is less than in NS). In the extreme case of FS a speaker begins verbally to scan all the syllables.

Some speakers displayed the first tendency, others displayed the second, and some displayed both tendencies. However, speakers more often displayed only one tendency with polysyllabic word. The first tendency is predominant in forced speech if in the normal speech the stressed vowel is not distinguished by its intensity. Such is the case if the stressed vowel in a word is /i/ or /u/ and the preceding vowel is /a/.

In SS also both tendencies are displayed. However, the second one (levelling of intensities) is more frequent. In such cases the intensity contrast may disappear, that is, vowel intensity may coincide with consonant intensity, the consonants being sonants as well as voiced and voiceless fricatives. It is not infrequent

that all sounds in a word are of equal intensities.

Table 2. Duration change of vowels in FS and SS as compared with NS (msec)

vowel / speaker MA		stressed /a/	stressed /o/	stressed /e/	stressed /u/	stressed /i/	pretonic /a/	pretonic /u/	posttonic /a/	posttonic /'a/	posttonic /ɨ/	posttonic /i/	posttonic /o/
I	NS	200	185	210	260	180	110	120	140	160	150	140	65
	FS	+25	+15	+40	+30	+40	+10	-10	+10	+15	+30	-25	+10
	SS	+65	+50	+90	+95	+65	+10	+40	+15	∅	+55	-10	-10
II	NS	145	130	130	105	105	95	90	75	95	80	75	50
	FS	+25	+25	+25	+55	+30	+25	+25	+30	+30	+15	∅	+25
	SS	-10	-25	-15	+15	-25	-15	+25	-10	-25	+55	-70	-10
III	NS	140	120	145	160	140	75	105	80	80	95	65	50
	FS	+55	+55	+75	+55	+55	+15	∅	+10	+15	-15	+10	+10
IV	NS	230	210	235	265	230	160	130	160	175	130		
	FS	-30	-15	-25	-65	-80	+15	-25	-40	-40	-10		
	SS	+65	+40	+55	-15	+65	+25	+55	-10	-25	-30		
V	NS	195	195	175	200	200	120	130	130	170	120		
	FS	+25	+40	+40	-15	+15	+30	-15	-25	-40	+10		
	SS	-30	-75	-40	-40	-105	+25	-50	∅	-75	-15		
VI	NS	160	155	195	185	140	170	115	95	80	90		
	FS	-10	-30	-10	+10	+25	-30	+40	-15	-10	+55		
	SS	-10	-15	-40	+25	-10	-15	-30	-25	∅	+25		

Table 3. Intensity of vowels in FS ɑnd SS as compared with NS

stimulus word / speaker	forced speech II	IV	I	III	V	VI	sloppy speech II	IV	I	III	V
pala-tka	+−	++				++	+−	+−			
za-ponki	ø−	+−				++	+−	−−			
a-lpha	−	−	−	+	+	−	−	−	−	+	−
uzda-	−	ø	+	−	+	+	−	+	+	+	−
ia-ma	−	−	+	+	−	+	+	+	−	+	−
kho-lodno	−−	++				++	−+	++			
t'o-sh'a	ø	+	+	+	−	+	ø	−	+	ø	−
uzhe-	+	+	−	+	−	++	−	+	−	ø	−
e-ta	−	−	−	+	+	+	−	−	+	−	−
ke-pka	−	+				+	+	−			
azu-	+	+	+	−	+	+	−	+	+	+	+
analgi-n	ø+	++				++	ø−	++			
i-kry	−−	++	+	−	+	++	++	ø+	+	+	−
i-m'a	−	−	−	−	+	+	−	+	−	−	+

+ (stressed) vowel more intensive
- (stressed) vowel less intensive
ø (stressed) vowel equally intensive

SPECTRUM

Only visual estimates of the spectrum have been made. Due to these estimates there are no regular substantial changes in the spectrum in FS as compared with that of NS. F_1 and F_2 of the vowels remain at the same frequencies and retain their normal intensity values. It is important that F_1-F_2 difference does not change perceptibly in FS even for /i/, though this contradicts some observations reported in the literature. Despite increasing intensities of higher formants, errors in perception can be more satisfactorily explained by errors in the horizontal rather than the vertical position of the tongue in articulation, that is, information about F_1 is more easily perceived than that about F_2 (cf. frequent substitutions /u/↔/i/↔/ɨ / and /o/—/e/). However, in FS there are occasional peculiarities of spectrum that serve to increase its intelligibility as compared with NS. The peculiarities are as follow:

I. Vowel formants occupy the most characteristic frequencies in the spectrum (e.g. F_1 for /a/ and F_2 for /u/ are higher than in normal speech).

2. Consonant noise is amplified at more

characteristic frequencies than in NS.

3. The formants of sonants are physically more distinct (e.g. better physical manifestation of nasal formant, etc.).

In general these peculiarities may, together with the lengthening of sounds frequent in FS, explain the increase in the intelligibility of FS as compared with NS.

FREQUENCY

Table 4. Fo_{max} and Δ Fo of the stressed vowels in various MAs

Vowel	MA speaker	NS		FS		SS	
		$f_o max$	ΔF_o	$F_o max$	ΔF_o	$F_o max.$	ΔF_o
/a/	I	I32	I5	200	52	II7	3
	II	I30	I3	I90	38	II3	20
	IV	I3I	I6	I48	44	I06	9
	V	I29	6	272	I02	II7	II
	VI	I4I	I9	I73	34	-	-
/o/	I	I35	I3	208	55	I37	8
	II	I40	I0	202	35	I37	25
	IV	I34	I7	I56	40	II9	I2
	V	I36	9	280	80	II6	II
	VI	I40	2I	I74	34	-	-
/u/	I	I22	5	2I0	56	II7	2
	II	I42	I2	I98	42	I27	I5
	IV	I44	I9	I73	6I	I08	8
	V	I38	9	276	I06	II3	8
	VI	I43	20	I93	40	-	-
/t /	I	I50	20	240	85	I35	I5
	II	I50	I0	2I5	30	I40	25
	IV	I40	I7	I55	28	II3	8
	V	I40	I3	283	75	I20	I0
	VI	I4I	23	I87	48	-	-
/i/	I	I33	I3	235	55	II8	I0
	II	I43	I5	I95	50	I23	23
	IV	I30	I4	I70	43	III	3
	V	I45	9	285	I06	II8	9
	VI	I4I	I5	I85	40	-	-
/e/	I	I28	5	2I8	50	I23	7
	II	I40	I3	202	38	I32	20
	IV	I38	20	I55	47	I03	7
	V	I32	II	28I	88	II7	8
	VI	I38	I5	I80	37	-	-

As could be seen in table 4, an increase of Fo of 50% in the average is characteristic of FS (the range of Fo-increase is from II% for speaker IV to II3% for speaker V). In SS Fo_{max} increases by about I3% (the range of Fo-increase is from I% for speaker I to 25% for speaker IV). In addition, there is difference between Fo_{max} and Fo_{min} over the same vowel four times greater in FS than in NS (the range is from I,6 times for speaker VI to I7 times for speaker V).

These speech events are consistent and reproducable and as such can serve to distinguish between various modes of articulation.

It is probable that the degree of forcing determines Fo values. The intensity increase and the rise of Fo (maximum and change) are absolute indicators of FS, while the intensity decrease and the fall of Fo (maximum over stressed vowels) are absolute indicators of SS.

PERCEPTION

To investigate perception of speech in various MAs four word-lists containing 3I stimuli each were read by 3 speakers. The record level was adjusted so that all stimuli were equal in intensity. Thus the factor of intensity was excluded since it effects perception greatly, FS being about 3 times more intensive and SS two times less intensive than NS. Three groups of subjects listened to the recordings under the conditions of noise. The intelligibility score for each word-list has been calculated. To neutralize memorizing the order of presentation was as follows: NS, FS, SS.

The average intelligibility scores were 58% for NS, 66% for FS and 48% for SS. The factor of intensity being neutralized three main variables: fundamental frequency, duration and spectrum-determined speech intelligibility. Energy distribution among vowels effects perception as well. Duration, on the other hand, does not effect intelligibility as such (in SS duration is often greater than in NS without any evident effect). The accuracy of articulation is an important factor in the increase of intelligibility in FS and reduction of intelligibility in SS.

It is not infrequent in FS that the better physical manifestation of formants and formant transitions accounts for better vowel identification and better identification of place for adjacent consonants. Consonants in FS are charaterized by amplified parts of the spectrum relevant for their identification.

There are certain errors in FS, such as substitution of fricatives for plosives, which may result from release lengthening in plosives (substitution KH for K, etc.) and inserted vowels (/cɨ-kl/ is perceived as /tɨ -pel/, /ku-pol/, etc.). These errors, however, are compensated for by better recognition of other sounds, final vowels and consonants in particular.

Thus the increased intelligibility of FS as compared with NS and SS is explained not only by its being louder but by the change in other parameters like duration, fundamental frequency and spectrum. On the other hand, the increased loudness and more "imperative" sound of FS mobilize the hearer's attention to a greater degree. This may be one reason for the inappropriate use of FS to overcome an interlocutor's dullness or a child's disobedience.

CONCLUSION

FS and SS may be considered as special varieties of speech characteristic of everyday speech communication and, as such, may be of theoretical and practical interest. The peculiarities of FS and SS may prove useful for automatic speech recognition and high quality synthesis of speech.

INTELLIGIBILITY OF ENGLISH, FRENCH, GERMAN, AND SPANISH CONSONANTS GENERATED BY RULE OVER SIMULATED TELEPHONE BANDWIDTHS

BATHSHEBA J. MALSHEEN MARISCELA AMADOR-HERNANDEZ MELANIE YUE JAMES T. WRIGHT

Speech Plus, Inc.
640 Clyde Ct.
Mountain View, CA USA 94043

ABSTRACT

The intelligibility of synthetic English, French, German and Spanish initial consonants was tested under normal and telephone bandwidth conditions. Segments were synthesized by language-specific rules using the CallText 5000 text-to-speech converter, based on a cascade-parallel formant synthesizer derived from MITALK-79. The data were compared with those of a human speaker of each language. Nonsense syllables were presented in an open-response format. The results show that (1) on the average synthetic segments for all four languages are 35% less intelligible than human ones, and (2) telephone bandlimiting only slightly degrades synthetic consonants. The findings from nonsense syllables differ from those previously reported for real English words, which were substantially degraded by bandlimiting.

INTRODUCTION

Over the past decade several researchers have tested the segmental intelligibility of a number of English synthesis-by-rule systems [1,2]. The results of these tests have indicated that the intelligibility of high-quality text-to-speech systems is approximately 90% for consonants in real English words. Nevertheless, the level of intelligibility for synthetic phonemes is still significantly lower than that of natural speech. Pisoni, et al suggest that synthetic speech lacks the full compliment of perceptual cues necessary to decode speech. Put in a different way, synthetic speech lacks the acoustic-phonetic redundancy of natural speech, and is acoustically "impoverished".

In order to further investigate the properties of synthetic speech, we measured the intelligibility of synthetic initial consonants for English, French, German, and Spanish, and compared the results of each language to those of a human speaker. In addition, we set out to determine the effects of telephone bandlimiting on the intelligibility of synthetic consonants, in order to learn more about how listeners decode and process these segments. In a previous paper, Wright, et al.[3] found that when consonants in monosyllabic real English words were tested under a telephone bandwidth condition, both human and synthetic consonants suffered significant losses in intelligibility. In this study, we were interested in ascertaining differences in segmental intelligibility between human and synthetic speech when stimuli were nonsense syllables which contain no semantic information to aid in phoneme identification. Finally, we wished to determine whether any consistent phonemic error patterns could be identified for a number of language-specific synthesis-by-rule systems.

Segments were synthesized by rule using the CallText 5000 text-to-speech converter, which is based on a cascade-parallel formant synthesizer derived from MITALK-79 [4]. Whereas the CallText English synthesis-by-rule system is a commercial product which has undergone considerable development and linguistic improvement over the past few years, the other language systems are in earlier stages of development.

METHOD OF TESTING

The intelligibility of initial consonants for all four languages was tested in an open-response format. Test stimuli for each language were CV nonsense syllables which included all of the phonemes occurring in initial position in each language. Three tokens of each phoneme followed by /i,a,u/ were presented to six native speakers in each listening condition. All testing was conducted in a sound-treated room at the Phonology Laboratory of the University of California at Berkeley. None of the subjects had ever before heard synthetic speech.

For English, human and synthetic stimuli were tested in normal and telephone bandwidth conditions. Synthetic stimuli for French, German, and Spanish were tested in normal and telephone bandwidth conditions. In addition, the same stimuli were recorded. by French, German, and Spanish native speakers. All recorded human speakers were male.

An average U.S. long-distance telephone line was simulated for the telephone bandwidth condition. Our motivation for simulating a long-distance telephone connection rather than using an actual long-distance line was the variability in transmission performance reported in the "1982/83 End Office Connection Study" [5]. The variability reported suggests that it would be difficult to replicate the actual telephone channel characteristics at different times or to ensure that an actual given connection was typical. The telephone channel introduces many distortions which we potentially could have modelled. We chose to simulate the telephone connection's frequency response and some aspects of its noise characteristics. Our simulator included an octave filter bank to create a transfer function closely matching the average long-distance frequency response reported in the End-Office Connection Study. Because the data in the study represent only the measurements from one central office to the other, the filter bank was adjusted to incorporate the additional losses of the end loops. These losses are estimated to be 1.75 dB per 1000 Hz at each end. The simulator used a codec to maintain a constant signal-to-noise ratio over a wide range of signal levels.

RESULTS

Figure 1 compares mean percentages of error for natural and synthetic consonants in the normal bandwidth listening condition. Note the considerable increase in errors for the synthetic stimuli when compared with the human ones. The differences in mean error percentages between the four synthetic language systems reflect the varying degrees of system development.

Figure 2 compares mean error percentages for English human and synthetic consonants in normal and telephone bandwidth listening conditions. As shown in this figure, error rates for the human stimuli are more affected by bandlimiting than those for the synthetic ones. The mean percentage of error for human consonants increased in the telephone bandwidth condition from 5% to 12%, but the synthetic one increased only from 33% to 35%.

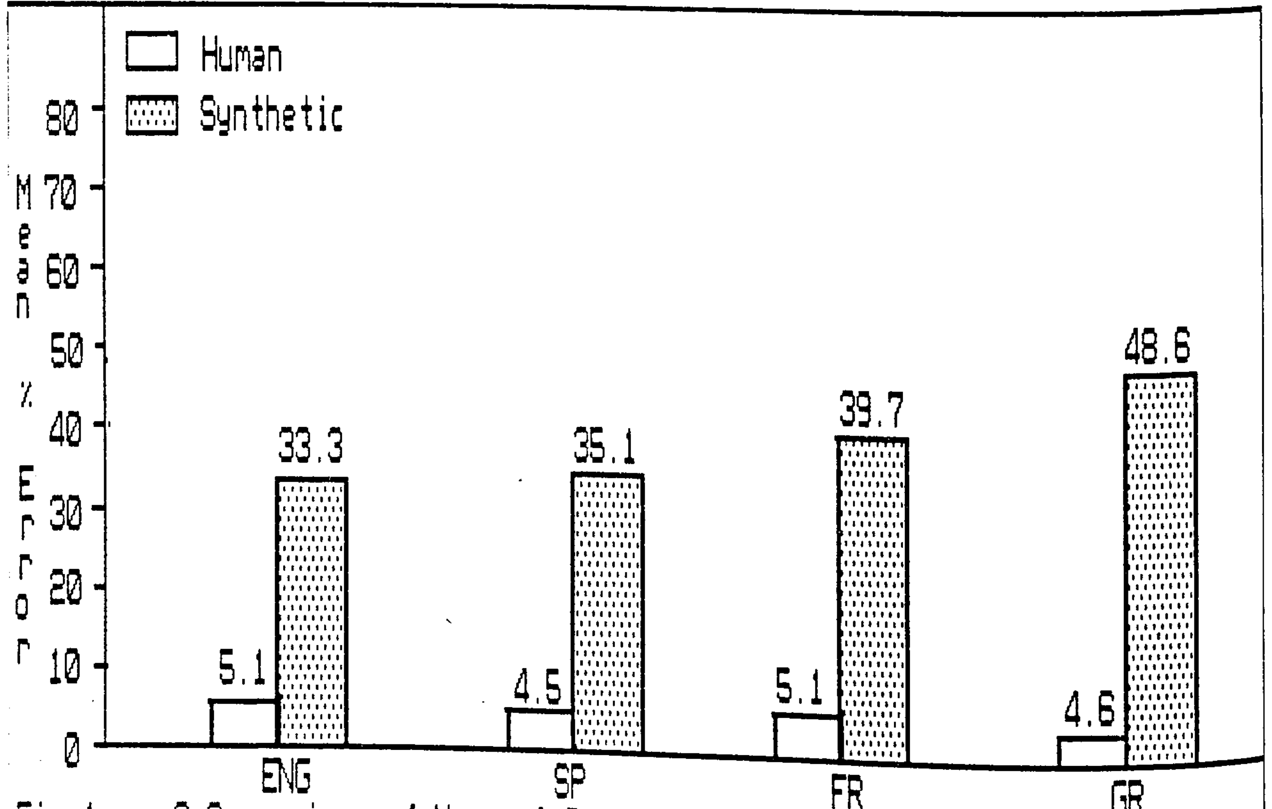


Fig 1 A Comparison of Human & Synthetic Consonants for Four Languages

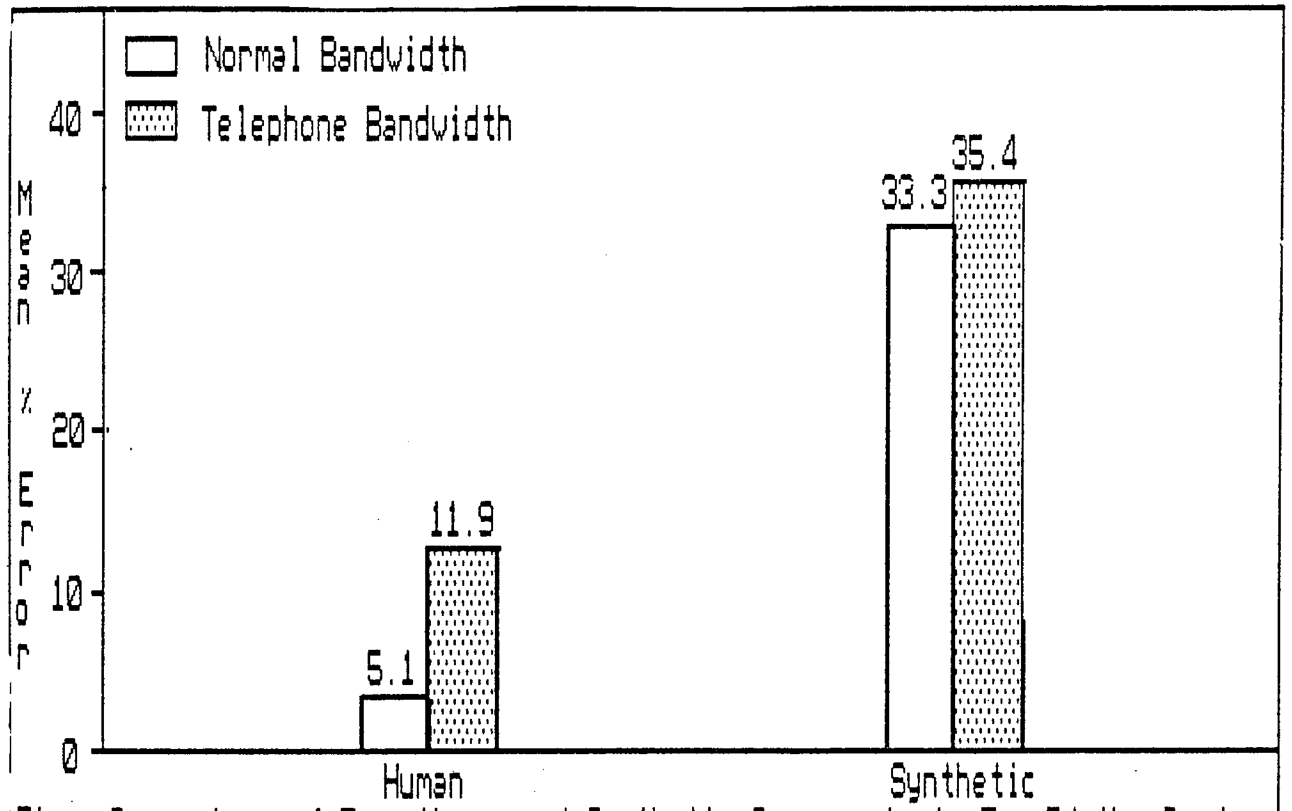


Fig2 Comparison of Eng Human and Synthetic Consonants in Two Bdwth Conds

Figure 3 compares intelligibility results for synthetic segments in all four languages for normal and telephone bandwidth conditions. The synthetic segments, which have high error rates in the normal bandwidth condition, degraded only slightly in the bandlimited condition. The average increase in errors for the telephone condition was only 4% from the normal bandwidth condition.

CONCLUSION AND DISCUSSION

Although the data in this study represent only six subjects per language, the results clearly indicate that synthetic nonsense syllables are much less intelligible than human nonsense syllables. Additionally, the data show that the synthetic segments, unlike human ones, suffer only slight degradation due to the attenuation of higher frequencies. These results for synthetic consonants in nonsense syllables are different from those found for real words by Wright, et al. Consonants in real words which were generated by the same synthesizer, degraded substantially when bandlimited. These consonants, however, had intellibility levels of approximately 90% under the normal bandwidth condition. It appears that once intelligibility drops below a certain level the effects of bandlimiting are minimal.

Our findings support Pisoni's contention that synthetic speech is acoustically impoverished. Presumably, a great deal of the acoustic-phonetic information necessary to signal phonemic distinctions is missing or incorrectly specified in the synthetic stimuli. Nevertheless, the fundamental phonetic information which is correctly specified in the synthetic stimuli--the information responsible for the 65% intelligibility of the segments--appears to be sufficiently robust to withstand bandlimiting.

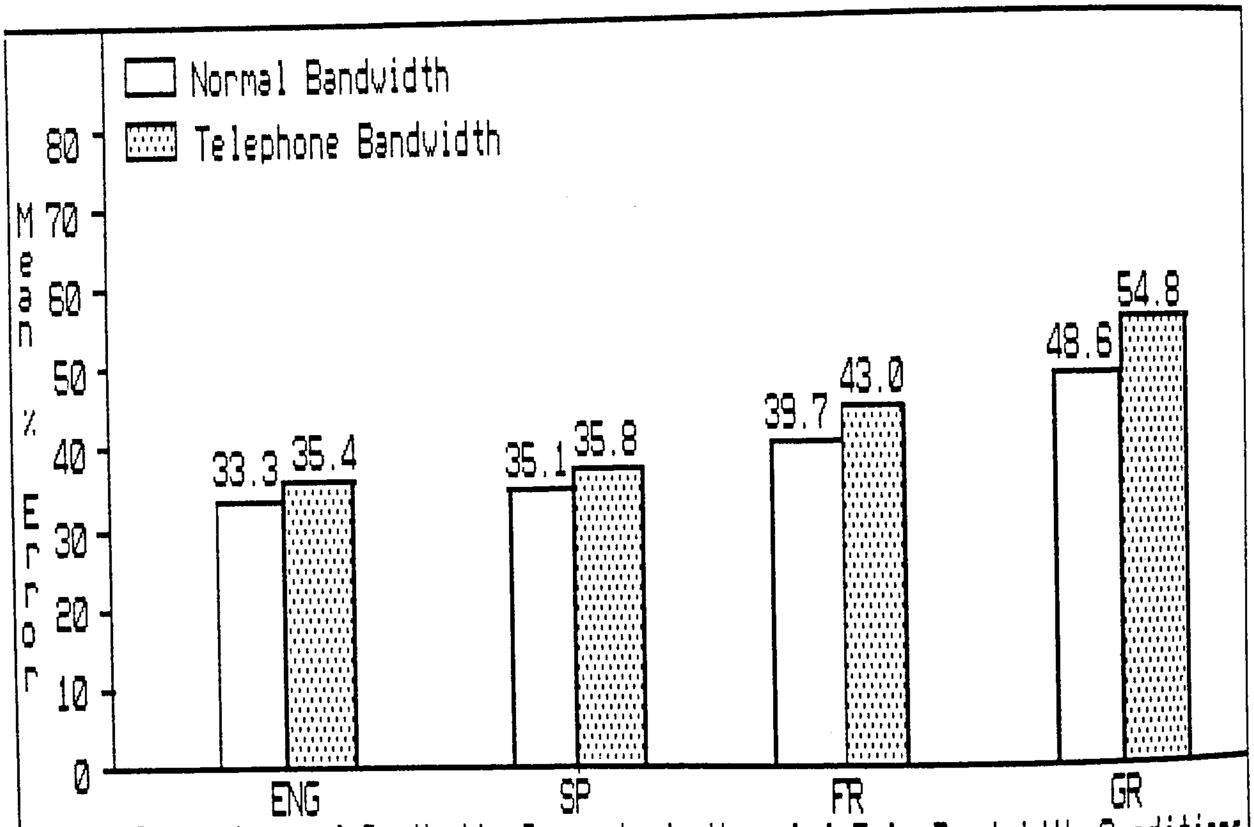


Fig 3 Comparison of Synthetic Segments in Normal & Tel. Bandwidth Conditions

REFERENCES

[1] Pisoni, D.B., Nusbaum, H.C. & Greene, B.G. (1985). Perception of Synthetic Speech by Rule. Proceedings of the IEEE, 73, 1665-1676.

[2] Pisoni, D.B., (1986). Some Measures of Intelligibility and Comprehension. In J. Allen (Ed.), From Text to Speech: The MITALK System, Cambridge U.K.: Cambridge University Press.

[3] Wright, James T., Malsheen, B.J., & Peet, Margot (1986). Comparison of Segmental Intelligibility and Pronunciation Accuracy for Two Commercial Text-to-Speech Systems, Proceedings of American Voice Input/Output Society, 235-261.

[4] Allen, J. (Ed.) (1986). From Text to Speech: The MITALK System. Cambridge UK: Cambridge University Press.

[5] Carey, M.B., Chen, H.J., Descloux, A., Ingle, J.F., & Park, K.I. (1984). 1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network. Bell Systems Technical Journal 63, 2059-2119.

TIME AND FREQUENCY RESOLUTION CONSTRAINTS ON SYNTHETIC SPEECH INTELLIGIBILITY

J.E. CLARK

Speech Hearing & Language
Research Centre
Macquarie University
North Ryde, N.S.W. 2113
Australia

R.H. MANNELL

Speech Hearing & Language
Research Centre
Macquarie University
North Ryde, N.S.W. 2113
Australia

D. OSTRY

Radiophysics Division
C.S.I.R.O.
North Ryde
N.S.W.
Australia

ABSTRACT

The effects of time and frequency resolution properties of resynthesised natural speech on its intelligibility were investigated at the phonological level. An automatic analysis-resynthesis channel vocoder was used to manipulate the time and frequency properties of the synthetic speech. The original natural speech and a high quality formant vocoder provided the comparative performance benchmarks. The test materials were noise-masked monosyllables. Results showed that vowels made the greatest demands on frequency resolution, with both consonants and vowels showing similar overall demands on time resolution. The higher information rate channel vocoders were markedly superior in consonant intelligibility to the formant vocoder benchmark.

INTRODUCTION

This investigation was motivated by a general interest in the performance of speech synthesis systems, and in the parametric coding required to represent the phonologically related information content with perceptual adequacy.

Limitations in the intelligibility and perceptual robustness of synthesised speech have been observed since the time of Stewart [1]. There has been accumulating quantitative evidence of this limitation in more recent times [2], [3], [4], [5], prompting Pisoni et al [4] to comment that "..it seems more advisable to use a low-cost synthesizer to provide spoken confirmation of database entries than as a voice response system in the cockpit of a jet fighter or a helicopter." (p.1675).

OBJECTIVES

The broad objectives of this study were:

1. To try and determine some of the ways in which the intelligibility of synthesised speech is constrained by resolution of the information (in its time and frequency domains) of the information contained in its resynthesis parameters.

2. To relate the findings on synthesis parameter manipulation to the intelligibility of the original natural speech and a known high quality formant vocoder as benchmark comparisons.

METHODOLOGY

Speech Processing Systems

A classical channel vocoder was chosen as the means for manipulating the parametric information content of the resynthesised speech signal. This class of vocoder has time and frequency resolution properties which are explicit in their structure. Moreover, they make few apriori assumptions in their parametric encoding about the nature of the phonologically related information bearing properties of the time-varying spectrum of speech signals. They do, of course, make some necessary assumptions in relation vocal tract excitation sources, about the nature of its periodicity and aperiodicity.

The channel vocoder is the earliest electronic speech analysis-resynthesis device. It was first developed some 50 years ago, motivated by an interest in reducing telephone transmission bandwidths. This is achieved (without great coding efficiency) by only transmitting the relatively slow time-varying changes in the energy envelopes of the speech signal spectrum as sampled by a filter-bank analyser spanning the range of frequencies of interest in the signal to be processed. The output of each analysis filter is detected and processed to produce the necessary slow time-varying envelope signal, and this information is then transmitted for resynthesis at the other end of the transmission path. The resynthesis is achieved using a corresponding filter-bank excited by periodic and/or aperiodic functions of uniform spectral energy, or a mix of both, as appropriate The actual excitation level appropriate to each filter is set using a multiplier controlled by the energy envelope signal derived from the corresponding analysis filter channel. Excitation function information defining whether it is periodic or not, and in the former case the period itself, is derived direcly from the input signal and transmitted to the resynthesis component of the system separately. Fig. 1 shows the structure of the vocoder in schematic form.

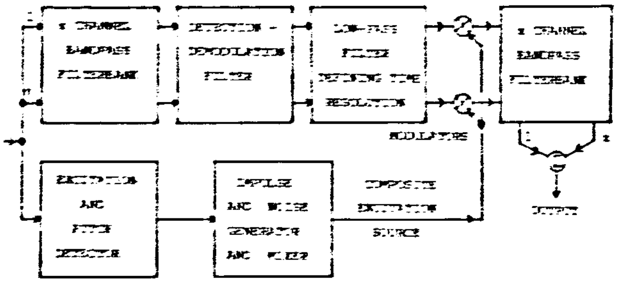

N CHANNEL BANDPASS FILTERBANK
DETECTION - DEMODULATION FILTER
LOW-PASS FILTER DEFINING TIME RESOLUTION
N CHANNEL BANDPASS FILTERBANK
MODULATORS
INPUT
EXCITATION AND PITCH DETECTOR
IMPULSE AND NOISE GENERATOR AND MIXER
COMPOSITE EXCITATION SOURCE
OUTPUT

Fig.1 S.H.L.R.C. Research Vocoder

The vocoder was realized as software on a VAX 11/750, and makes no attempt to meet any particular criteria of encoding or computational efficiency, given that it is only intended as a signal manipulation device. Identical filterbanks were used for analysis and resynthesis.

Despite the venerable age of this speech processing device, there are several quite basic questions about its design parameters which are not clearly resolved in the literature. It is not the purpose of this paper to discuss vocoder design, but it is worth noting that in develcping the vocoder used in this present investigation, several different analysis-resynthesis filter types were tried together with several forms of analysis filter energy detection before settling on the configuration used in this investigation.

Opinions in the literature [7], [8], on requirements for analysis/synthesis filter properties vary. Despite some claims that it is desirable to use filters with relatively shallow skirt slopes and having well damped impulse responses, and that filter skirt response overlap is relatively unimportant because of the large amount of correlated energy occurring in adjacent bands, it was found in this study that such filters produced speech of unacceptable quality and intelligibility. By contrast, each of several filters tried with steep skirt slopes and much more restricted response overlap produced far better speech quality.

The effective frequency resolution of the system is set by the number of filters, and may be selected from 6, 12, 24, and 48, to give uniform bandwidths of approximately 800, 400, 200 and 100 Hz respectively. The sampling rate of the vocoder is 10KHz, and hence the frequency range of the filterbank is 0 - 5KHz in all cases.

The question of optimal criteria for filterbank energy detection systems also seems unresolved in the literature. For the present investigation, the need for independent manipulation of the vocoder data rate made a Hilbert filter a suitable choice to meet the output ripple and response speed criteria. This is rather specific to the uses of the vocoder, being designed to provide an output energy envelcpe with a maximally rapid impulse response.

A separate set of low pass filters were used to limit the bandwidth of the energy envelope signals, so simulating changes in the vocoder data transmission rate. This controls the effective time resolution of the information transmitted for resynthesis. The cut-off of this filter may be set to give effective parameter update rates of 10, 20, 40, and 60mS. It may also be bypassed to give a limiting vocoder time resolution set by the combined effects of the analysis filterbank and energy detection systems.

Algorithms for deriving pitch and voicing status excitation data abound; the scheme used here is not claimed to have any special merit, but was a time-domain type specifically tailored to the needs of this vocoder. The excitation signals for resynthesis in the vocoder are derived by direct extraction of smoothed pitch data, and a voicing detection system which determines whether the signal is periodic, aperiodic, or a mixture of both. The detection system contains hysteresis to minimise voicing decision jitter.

The formant vocoder used was a standard high quality system at the Joint Speech Research Unit, using a copy of the master recording of the benchmark natural speech materials as input. The resynthesis uses a four formant systems based on the well known J.S.R.U. synthesiser, chosen because of its reputation for very high quality speech output.

Listening Tests

The perceptual properties of the acoustic speech signal were tested in conditions of near minimal linguistic context to minimise the confounding effects of top-down processing by listeners. A set of 11 /h-d/ words and 19 CV nonsense syllables representing a selection of the common vowels and consonants of English respectively, were employed.

Natural speech versions of the test materials that were used as input to the vocoder were tested to provide a benchmark for the vocoder speech intelligibility data. The original natural speech was recorded to professional standards in an echo free sound treated room. The full range of time resolutions were tested using a 24 channel vocoder, and the full range of frequency resolutions were tested using a 10mSec time resolution (data update rate). The formant vocoded speech was processed with a 10mSec time resolution.

All the speech types were tested unmasked, and masked, the latter at signal to noise ratios of +6, 0 & -6 dB. The masking noise had a sloping spectrum approximating the long term spectrum of male speakers of English, and all the test stimuli were level normalised using the standard Leq method. The test stimuli presentations were all randomised and recorded with a 500Hz tone preceding each stimulus, and an interstimulus interval of 5 Sec. The stimulus and test tape generation was done digitally on the VAX 11/750. Listeners were drawn from amongst students and staff at Macquarie University. No listeners experienced with the task or with speech synthesis were employed, and listeners were not used for more than a single test session. Prior to the test sessions listeners were given a simple speech discrimination test to ensure that they could accurately identify common monosyllabic words down to a presentation level of 45dB s.p.l. before being included in the test crew. All the test materials were presented at +70dB s.p.l. using TDH49 headphones with standard cushions and circumaural seals in a sound treated room.

Analysis Procedures

The response data was entered into a computer program which produced intelligibility scores by individual test condition, and pooled intelligibility scores.

RESULTS

Fig. 2 shows intelligibility by frequency resolution by masking. The vowels are overall more resistant to masking than the consonants, with the formant vocoder vowels the most resistant of all. Both the 48 and 24 channel vocoders produce highly intelligible vowels at all but the deepest masking condition, whilst the poor performance of the 12 and 6 channel vocoders demonstrates the importance of frequency resolution. Note the rising intelligibility with noise in the 6 channel case.

The synthesised consonants show lower overall intelligibility than the vowels, although the 48 and 24 channel vocoders show a resistance to masking which is comparable to or better than natural speech in conditions of moderate masking. The formant vocoder is a little poorer than the 12 channel vocoder, except in moderate to heavy masking.

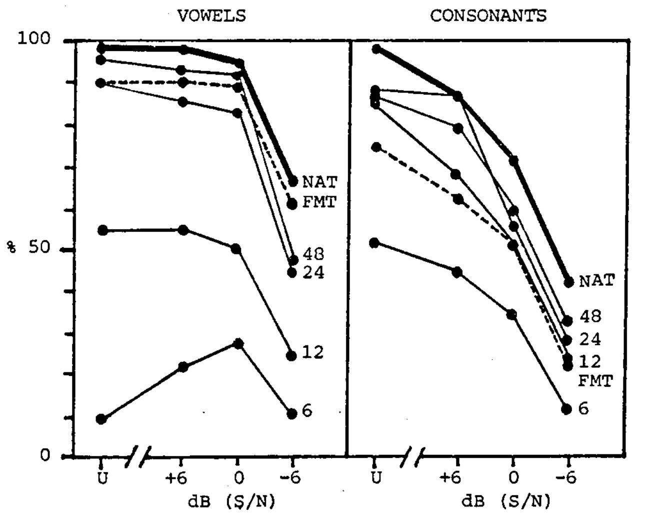


Fig.2 Intelligibilty X Frequency X Masking

Fig. 3 shows intelligibility by time resolution by masking. The vowels are relatively tolerant of reduced time resolution with no degradation until the 40mS condition, and a slight rise in intelligibility with moderate masking. The consonants show a similar pattern but with more rapid degradation at 40 and 60mS. The 10mS condition is least resistant to moderate masking. The formant vocoder has a performance which is comparable to or slightly better than a 40mS 24 channel vocoder.

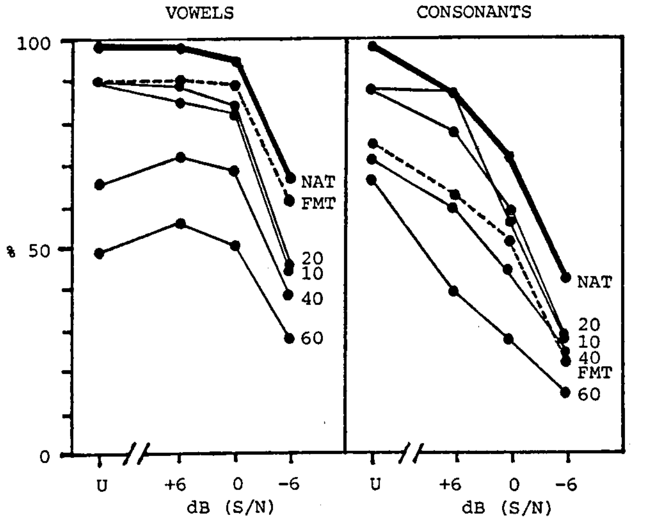


Fig.3 Intelligibility X Time X Masking

Figs. 4 & 5 show intelligibility by frequency and time resolution respectively, with pooled masking data. Vocoder vowel intelligibility decreases rapidly below 24 channel frequency resolution, and requires very high frequency resolution to approach formant vocoder performance. Consonant performance is far more tolerant of reduced frequency resolution, and suggests the formant vocoder to have a performance similar to that of a 12 channel vocoder.

Time resolution effects on performance are more consistent for both vowels and consonants, with appreciable reductions in intelligibility occurring at 40mS and above.

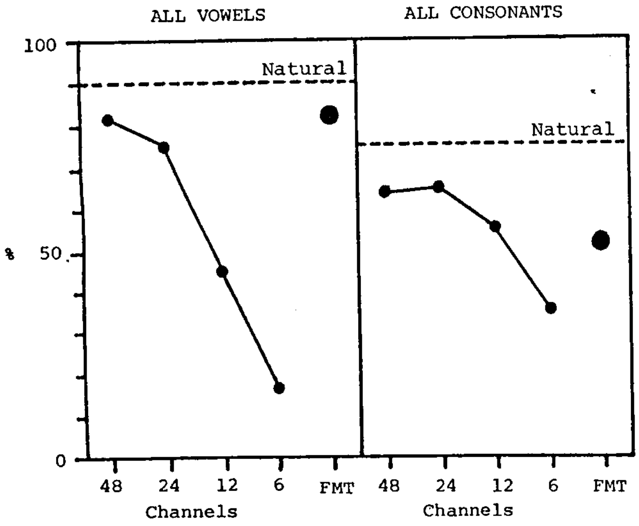


Fig.4 Pooled Intelligibility X Frequency

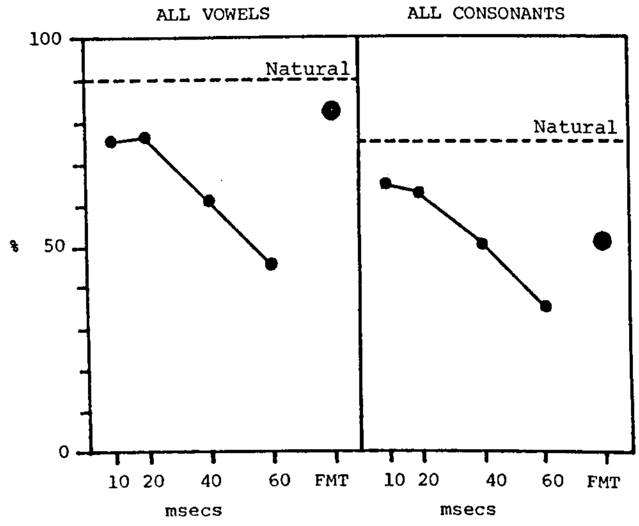


Fig.5 Pooled Intelligibility X Time

CONCLUSIONS

1. Overall intelligibility is more degraded by reduction in frequency resolution than by reduction in time resolution under the conditions tested (insofar as the two domains can be compared).

2. The comparative intelligibilities of vowels and consonants are reversed by progressive reduction in frequency resolution, but not time resolution. This illustrates the more stringent demand on frequency resolution in vowel parameter coding.

3. Time resolution reduction has a more consistent effect overall than frequency resolution reduction.

4. The formant vocoder shows a much greater performance differential between vowels and consonants than the channel vocoders. This generally poorer performance of consonant intelligibility with the formant coded speech suggests that it has appreciably less adequate parametric coding of consonantal information content than the 48 and 24 channel vocoders.

ACKNOWLEDGEMENTS

This work has been supported under the Macquarie-C.S.I.R.O. research grant scheme, and the A.R.G.S. The contribution of the technical staff of the S.H.L.R.C. is gratefully acknowledged, and we would like to thank Alison Brierley for running many of the listening test sessions, and tabulating the listener response data.

REFERENCES

[1] J.Q. Stewart, "An electrical analogue of the vocal organs", Nature, 110, 311-312, 1922.
[2] R. Carlson et al, "Evaluation of a text-to-speech system as a reading machine for the blind", STL-QPSR 2-3, 9-13 1976.
[3] J.E. Clark, "Intelligibility comparisons for two synthetic and one natural speech source", J. Phonetics, 11, 37-49 1983.
[4] D.B. Pisoni et al, "Perception of synthetic speech generated by rule", Proc. IEEE, 73, 1665-1675, 1985.
[5] J.E. Clark et al, "Cue enhancement by stimulus repetition: Natural and synthetic speech comparisons", JASA, 78, 458-462, 1985.
[6] H. Dudley, "Synthetic Speech", Bell Labs Record, 15, 98-102, 1936.
[7] M.R. Schroeder, "Vocoers: Analysis synthesis of speech", Proc. IEEE, 54, 720-734, 1966.
[8] I.H. Witten, Computer Speech, Academic Press, 1982.

PERCEPTUAL SPACES AND THE IDENTIFICATION OF NATURAL AND SYNTHETIC SENTENCES

N. BACRI [*]
Laboratoire de Psychologie Expérimentale

C.N.R.S. - E.H.E.S.S.
54, bd Raspail 75006 PARIS

Is synthetic speech just degraded speech or is it processed as a specific perceptual space? The identification responses to 8 phonetically balanced lists of ten sentences each, using several syntactic structures, were studied for four sets of stimuli (natural speech, LPC speech, synthesis by diphones using two text-to-speech systems). All the stimuli were intensity equalized, then degraded by a masking pink noise. Phonetic and prosodic cues effects were strong, while the effect of syntax was weak. The choice of sentence identification strategy depends on the natural vs synthetic nature of the speech used and on SNR: a step-by-step decoding for impoverished synthetic speech and a SNR below 8 dB, backward lexical interpretation for natural speech or a low noise. Acoustic cues redundancy and masking noise level impose the choice of specific cognitive processing modalities.

In the case of spoken language, sentence perception and comprehension imply the interaction of both acoustic and linguistic sources of knowledge to identify word boundaries, select word candidates and construct a meaningful sequence. According to identification tasks using a gating paradigm in which signal duration is varied /3, 4, 9/, data support the assumption of a parallel and interactive processing of acoustic-phonetic information and of syntactic-semantic information provided by the sentence context. It is the redundancy of lower-order and higher-order sources of information which can explain the listener's ability to understand speech even under degraded conditions. But the redundancy of acoustic-phonetic cues by themselves is also of importance. It is possible to evaluate its weight by comparing sentence recognition performance for natural speech and synthetic speech of different qualities.

Previous research has demonstrated that synthetic speech is more difficult to recognize than natural speech /8/. This is perhaps due to what Nusbaum and Pisoni /7/ call the "noisy speech" hypothesis i.e. the fact that acoustic structure of synthetic speech is somehow degraded, as is the case of natural speech in noise. But according to the "impoverished speech" hypothesis, the rather bad performance for synthetic speech corresponds to a specific cognitive processing. Listeners must adapt their perceptual and identification strategies to a signal which is in its nature different from natural speech: they have to build a new perceptual space.

The present experiment aims at studying how naive listeners, without a previous knowledge of synthetic speech, can manage to understand sentences with different degrees of syntactic complexity, either natural or digitized, or generated by good vs. low-cost text-to-speech systems. Moreover stimuli were degraded by adding varying amounts of pink noise. The main hypothesis is that the level of performance and kind of errors will be linked to the quality of the sets of stimuli i.e. to the characteristics of the potential perceptual space. In any cases they will be significantly different for natural and synthetic speech. Another assumption bears on the effect of syntactic and semantic complexity. As speech becomes less intelligible, according either to its quality or to speech-to-noise ratio (SNR), listeners will rely more heavily on linguistic structure, so that easy-to-parse sentences would be better understood than less predictable ones, specifically as the quality of synthetic speech becomes worse /7/. Finally, following the researches on synthetic speech training /2/, it can be hypothesized that the results will improve from the first to the second session.

Speech materials and systems

Eight phonetically balanced lists of ten sentences each, covering a range of syntactic structures and semantic degrees of plausibility /1/, were read by a trained female speaker, with a neutral intonation and a 4.27 syllables/second speech rate. The first set

[*] This paper was supported by the CNET. Experiment was carried out in col. with P. Graillot, when the author was on leave at CNET, Lannion-A. She would like to thank him, C. Sorin and other members of RCP Department for their help.

of stimuli, A_1, consisted of these naturally spoken sentences. Audio tapes of the original sentences were then sampled at 16 kHz (16 coefficients), digitized by a linear prediction coder, and stored on disk by a PDP-11/34 computer. This second set of materials will be referred to as A_2. The two other sets were generated using synthesis by diphones according to two text-to-speech systems. The high-quality one, A_3, was processed with all frames set to 13 ms using a PDP-11/34 computer, and generated from a diphone dictionary recorded by a male speaker at a 3.42 syl./sec. speech rate. Prosody was a good approximation of natural speech. The last set of stimuli, A_4, was processed by a low-cost system using a diphone dictionary recorded by a female speaker at a 3.18 syl./sec. speech rate. This dictionary was implemented on a micro-processor (26 ms period). Some rough prosodic markers were added.

Procedure and subjects

Mean intensity of all the stimuli was equalized at 71/72 dB lin. The stimuli were masked by pink noise the intensity of which decreased from trial to trial. In the first trial, SNR were of - 2 dB for natural speech, + 4 dB for LPC speech and high-quality system, + 8 dB for low-cost system. These values were chosen so that no correct response could be given at the first presentation. At each of the 6 successive presentations, the level of noise was diminished by 2 dB steps for natural speech, 2 dB then 3 dB steps for synthetic speech. Four groups of 5 subjects each participated in the experiment during 2 sessions, at an interval of 5 days. All groups were given the same recognition task. Subjects had to say what they had understood after each presentation of each sentence. Order of presentation was counterbalanced, and the systems were crossed with the lists according to a latin square design. For each group the factorial design was as follows:

$$S_5 * L_8 <A_4 * D_2> * Se_{10}$$

(S: subjects, L: lists, A: systems, D: test session, Se: sentences).

Speech-to-noise ratio at the identification threshold for all the responses, correct response percentages for each list, sentence or system, perceptual confusions and SNR at the identification threshold (IT) for correct responses were analysed.

Results and discussion

An ANOVA was performed on the SNR at the IT, after the IT reached by a subject in erroneous responses in the last trial was increased by 4 dB. Overall analysis showed that all the factors had a significant effect, especially the factor Systems (F(3, 48) = 682, p<.0001). Interactions were also significant. Three major findings were obtained for post hoc comparisons. First, the main discrepancy is between natural speech and low-cost text-to-speech system, as was expected, while the weakest is between LPC and high quality text-to-speech systems (F(1, 16) = 29.88, p<.01). This last result confirms the good quality of this synthesis, as well as the basic difference between natural and coded or synthetic speech (Table I).

Systems	Mean SNR	sd
A_1	2.59	.805
A_2	10.45	1.269
A_3	12.01	1.801
A_4	15.80	2.735

Table I - Mean SNR at the identification threshold (dB) as a function of the systems. A_1: natural speech; A_2: LPC speech; A_3: high-quality text-to-speech system; A_4: low-cost text-to-speech system.

Second, the effect of syntactic-semantic differences between sentences is significant (F(9, 144) = 9.59, p<.001), but it is higher for natural speech than for synthetic speech, and is not related to the intrinsic quality of synthetic speech. For synthetic speech systems, the presence of easy-to-parse sentences does not facilitate identification, compared with less expected structures. Third, differences between sessions are significant (F(1, 16) = 13.41, p<.01), but this effect is only due to the contrast between the reality of a kind of training for the "poor" system A_4 and the lack of learning in all the other conditions (Table II).

	Sessions		
Systems	1	2	Change
A_1	2.48	2.70	+ 0.22
A_2	10.72	10.18	- 0.54
A_3	12.22	11.80	- 0.42
A_4	16.99	14.62	- 2.37

Table II - Mean SNR at the identification threshold (dB) as a function of systems and sessions.

A complementary study of only the correct responses confirmed these findings. Mean SNR for A_2 and A_3 are nearly the same (Table III). But it would be misleading to conclude that these two systems present the same degree of intelligibility, for the distribution of erroneous responses indicates that text-to-speech systems are less intelligible than coded speech (Table IV).

Systems	Mean SNR	sd
A_1	2.24	.894
A_2	9.04	2.446
A_3	9.00	2.960
A_4	12.72	2.837

Table III - Mean SNR at the IT (dB) as a function of systems, for correct responses.

One can also see that a training effect appears only for the "bad" synthetic speech. Moreover, the extent of the improvement from the first to the second session varies depending on the sentence. But it is worth noting that some easy-to-parse sentences are less well understood than more difficult ones. These two results suggest that acoustic-phonetic cues play a role as well as syntactic or semantic information.

	Sessions		
Systems	1	2	Change
A_1	3	4	- 1
A_2	14	9.5	+ 4.5
A_3	26	23	+ 3
A_4	28	18.5	+ 9.5

Table IV - Percent erroneous responses, for each system and each session.

Analysis of perceptual confusions revealed systematic errors only for the text-to-speech systems. For example, initial /m/ and the nasal opposition, initial /v/ and the opposition /v-f/ led to numerous identification errors. On the contrary, no systematic error appeared for LPC and natural speech. From a morpho-syntactic and syntactic point of view, monosyllabic pronouns and prepositions were well identified, whereas mono or polysyllabic nouns in a subject noun phrase brought about errors. For all the positions, adverbs and adjectives were often omitted or modified. Generally speaking, errors located at the beginning of a sentence were usually not corrected, irrespective of syntactic structure. The only syntactic structure that was misunderstood was of the injonctive type (3 sentences). On the other hand, semantic plausibility played a role only when it was very low, irrespective of speech quality.

Though the verb is generally considered as the main component of a sentence /6/, the large number of misleading identifications of the first lexical items suggests that listeners processed information from left to right, according to a step-by-step decoding strategy. Sequential processing did not prevent backward error rectifying in some cases /4/. However backward corrections occurred just when first responses had exhibited a good degree of approximation to the signal. Fruitful corrections were always supported by a correct identification of sentence "scaffolding" provided by pronouns and prepositions. These results correspond to what can be called a comprehension strategy: locating syntactic marks allows the listener to restore missing phonemic or syllabic information. But a striking result of error analysis is that this kind of restoration appears only either for natural speech and coded speech or during the last trials for the other systems i.e. when the noise was very weak. The choice of a comprehension strategy is constrained by listening conditions as well as by the quality of signal. SNR analysis agrees with this assumption.

Intelligibility per se of the signal was evaluated for correct responses as a function of SNR at the identification threshold. In spite of some restrictions related to the range of syntactic structures /5/, cumulative frequencies for each SNR are a reliable indicator of intelligibility /10/. Identification data for each system are shown in Figure 1 as percentages of correct responses according to SNR carried out in each condition.

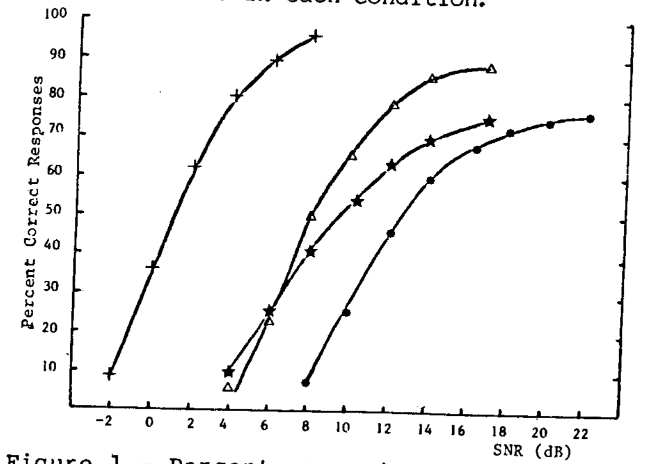


Figure 1 - Percent correct responses as a function of SNR, for each system. A_1: + ; A_2:△ A_3:★ ; A_4: ✱.

- Intelligibility gain (IG) for natural speech reaches about 13% for a SNR of -2dB to +2dB, improves by 9%/dB between SNR of +2dB and +4dB, then stabilizes around 3%-4%/dB.
- As for coded speech, IG is 8%/dB between SNR of +4dB and +6dB, to 13%/dB between SNR of +6dB and +8dB, then decreases to 6%-8%/dB for a +8dB to +12dB SNR range, and to 3.6%/dB between SNR of +12dB-+14dB. A plateau is reached around a SNR of +18dB.
- A_3 high-quality text-to-speech system is characterized by a less steep gradient. IG varies from 8%/dB to 5%/dB between SNR of +4dB to +12dB, then increases slowly by 2.5%/dB till the highest SNR tested. Plausibly a plateau could appear around a SNR of 19dB or 20dB, and an intelligibility of 85% could be reached.
- Evolution of IG for the low-cost synthetic system A_4 is quite similar to that of A_3. The gain is rather strong at first (9%/dB to 7%/dB for a SNR range of +8dB to +14dB). It then decreases to 4%-2%/dB for a +14dB to +18dB SNR range. Around a SNR of +18dB, the slope flattens out.

The more striking feature of Figure 1 lies in the clear contrast between natural and synthetic speech intelligibility. Comparison between A_1 and A_2 shows clearly that intelligibility of LPC coded speech decreases as noise gets louder. Discrepancy between the two systems is maximum for a SNR of +4dB and reaches a 75% loss of intelligibility. This loss is then reduced to about 45%, but remains high even for a rather weak noise. Thus, in the best listening condition, coded speech intelligibility does not exceed 90%. This result agrees with the hypothesis bearing on the specificity of synthetic speech, compared with natural speech.

Secondly, asymmetry of A_2 and A_3 intelligibility curves gives some interesting information pertaining to listener's strategies. The resistance to noise of A_3 is rather good for a loud noise. Intelligibility loss is indeed of 40% with regard to natural speech, but only of 10% compared with LPC speech. On the contrary, when SNR increases as noise becomes weaker, the gap between the systems A_2 and A_3 widens out. Assumption will be made that listeners adjust their identification strategy not only to the system, but also to the listening condition. When noise is very loud, they rely mainly on acoustic information. So very useful cues are given by the text-to-speech system. As a matter of fact, A_3 is characterized by clear segmentation cues, as for example prosodic cues i.e. F_0 movements and syllabic lengthening which are cues to word boundaries in French. As listening conditions get better, listeners can adopt another strategy, and give more attention to the sentence as a whole. This global comprehension strategy greatly improves the responses for a rather redundant speech as A_1 or even A_2, but it does not find a sufficient ground in impoverished synthetic speech to really succeed in A_4 and even A_3. That is perhaps why guessing or backward restoration very often fail when listeners are working with the two text-to-speech systems.

CONCLUSIONS

Impoverishment of speech by a pink noise varied mainly as a function of the system from which signal was generated. Relative weakness and lack of stability of sentence effect suggest that perceptual processing, in this experiment, has borne mainly on acoustic-phonetic cues, and secondly on prosodic segmentation cues. Listeners relied more on acoustic than on specifically linguistic information. Higher-order information was used, as demonstrated by the occurrence of backward lexical identification mechanisms; but its effect depends on the main effect of the quality of the system. Our results agree with the conclusion of Nusbaum and Pisoni /7/: "the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals" (p. 239). However, unlike them, we found that linguistic context becomes more important as the quality either of speech or of listening gets better, as is the case when one examines error restoration as well as identification thresholds. Furthermore, acoustic information is all the more processed as either speech quality or SNR are worse. In such bad listening conditions, subjects process the signal in a step-by-step fashion, more clearly so for synthetic speech than for natural speech.
Dissymmetry between responses is sufficient to rule out the hypothesis that synthetic speech is equivalent to natural speech degraded by noise. On the contrary, our results agree with the definition of synthetic speech as "impoverished speech" /7/, different in its nature from natural speech. They support the conclusion that the differences of intelligibility between natural and synthetic speech are related to the characteristics of speech signal. Different generating systems offer different patterns of cues to listeners. So listeners must construct and process several "perceptual spaces". The three synthetic speech systems generally present the same kind of confusion errors, more or less frequent depending on the quality of the system. Furthermore, two kinds of processing strategies can be hypothesized: a step-by-step decoding strategy and a global comprehension strategy. But further research is needed to better understand how perceptual spaces are built, what their consistency is, and how their processing can be improved.

REFERENCES

/1/ P. Combescure, 20 listes de dix phrases phonétiquement équilibrées, "Revue d'acoustique", 56, 34-38, 1981.
/2/ S.L. Greenspan, H.C. Nusbaum, D.B. Pisoni, "Perception of synthetic speech: Some effects of training and attentional limitations", Bloomington: Indiana University, Speech Research Laboratory, Progress Report, 387-413, 1985.
/3/ F. Grosjean, Spoken word recognition processes and the gating paradigm, "Perception and Psychophysics", 28, 267-283, 1980.
/4/ F. Grosjean, The recognition of words after their acoustic offset: Evidence and implications,"Perception and Psychophysics", 38, 299-310, 1985.
/5/ D.N. Kalikow, K.N. Stevens, L.L. Elliott, Development of a test of speech intelligibility in noise using sentence materials with controlled word predictibility, "J. of the Acoust. Soc. of America", 61, 1337-1351, 1977
/6/ G. Noizet, S. Bleuchot, R. Henry, Influence de la structure syntaxique de phrases entendues sur les stratégies de leur décodage perceptif, "Cahiers de Psychologie", 16, 149-180, 1973.
/7/ H.C. Nusbaum, D.B. Pisoni, Constraints on the perception of synthetic speech generated by rule, "Behavior Research Methods, Instruments & Computers", 17, 235-242, 1985.
/8/ D.B. Pisoni, Perception of speech: The human listener as a cognitive interface, "Speech Technology", 1, 10-23, 1982.
/9/ A. Salasoo, D.B. Pisoni, Interaction of knowledge sources in spoken word identification, "Journal of memory and language", 24, 210-231, 1985.
/10/ C. Sorin, Evaluation de la contribution de F_0 à l'intelligibilité, "Recherches/Acoustique", CNET, Vol. VII, 141-154, 1982/1983.

ASSESSING THE INTELLIGIBILITY AND PROCESSING SPEED OF PROCESSED SPEECH

KERRIE MACKIE PHILLIP DERMODY RICHARD KATSCH

Speech Communication Research Section
National Acoustic Laboratories
126 Greville, St., Chatswood. N.S.W. 2067. AUSTRALIA.

ABSTRACT - The present study examines evaluation measures designed to assess the intelligibility, and speed of processing of natural speech with harmonic distortion. The results indicate that even for highly intelligible processed natural speech delays in processing time are a consequence of poor acoustic phonetic information. The results also indicate the value of including more sensitive tests of speech intelligibility in evaluation protocols for speech transmission evaluation.

INTRODUCTION

In the development of voice communication devices,listening tests have always been an important part of evaluation procedures. The primary focus of these evaluation procedures has been the assessment of the intelligibility and quality of the speech through the device compared to some arbitrary standard. However, with continuing improvements in speech transmission systems the aim of the assessment procedure has changed to one which compares the output of speech transmission systems to listening results for natural speech.

The improvement in the quality and intelligibility of modern communications systems has produced a need for more sensitive assessment procedures which are appropriate to the evaluation of devices such as hearing aids where intelligibility is typically high, to the evaluation of synthetic speech produced by text to speech systems where intelligibility has a wide range of adequacy. The speech assessment of hearing aids is an active area of work and the evaluation of synthetic speech has produced some improved measures for speech evaluation(1).

In the present study we investigated measures of intelligibility and of processing speed to determine their relationship. Pisoni, et al (1) have applied a range of intelligibility and processing measures to the assessment of synthetic speech. They report that listeners have slower response times to synthetic speech compared to natural speech and they concluded that the increase in processing time is due to the poorer segmental intelligibility of the synthetic speech stimuli compared to natural speech. That is, difficulties at the acoustic phonetic level for synthetic speech underlie later processing time increases. In the present study we investigate measures of intelligibility and processing speed to determine their relationship for natural speech stimuli which have harmonic distortion.

EXPERIMENT 1

The first experiment was designed to demonstrate that the harmonic distortion of the stimuli increased the processing time for listeners in a manner similar to that reported for synthetic speech (1). The processing speed task chosen was the auditory lexical decision task. In this task the listener hears a stimulus and must decide as quickly as possible whether it is a word or a non word. Stimuli consisted of monosyllabic English words or pronounceable non-words. Pisoni, et al (1) reported that the lexical decision task showed slower reaction times for synthetic speech relative to natural speech, although for both synthetic and natural speech the relationship between words and non-word reaction times was similar. Pisoni, et al (1) concluded that synthetic speech is processed in a similar manner to natural speech at the lexical level, but that the impoverished acoustic-phonetic structure of synthetic speech led to its longer processing time overall.

STIMULI

The speech stimuli were recorded by an adult male speaker onto a computer speech storage/editing system using a 12 bit analogue to digital converter at a sampling rate of 36K samples per second. The stimuli were copied into three separate disc files which were then separately processed using an algorithm based on Schroeder (2) in which noise is added to the digitally sampled speech randomly over a specified time to produce harmonic distortion of the original waveform. The speech produced can be expressed as a change in signal-to-noise (S/N) ratio compared to natural speech. The S/N ratio is determined by the amount of distortion which is added per sample. In each file the speech was processed to give a S/N ratio of either 0, +3, or +6dB. The speech in each file was highly intelligible. This is attributable to the high sampling rate of the speech and the fact that the distortion technique is based on random noise addition per sample. When lower sampling rates are used, the result is considerably more degradation of the sampled speech for the same signal to noise ratio (2).

The listening tests were carried out on subjects seated in an audiometric test booth and speech was presented to them binaurally via headphones (type TDH49). The computer randomly selected the speech stimuli from the disc files and presented the stimuli, recorded the subject's responses and the reaction time in milliseconds for each stimulus.

RESULTS

Figure 1 presents the results for the lexical decision task and shows that in all conditions listeners respond faster to words than non words. The difference between the processed speech and the natural speech are reflected in reaction time differences only. That is, the lexical decision task reveals processing time differences between the different speech conditions and natural speech but not qualitative differences in processing the words and non-words. These results are similar to the results of Pisoni, et al (1) for synthetic speech compared to natural speech.

EXPERIMENT 2

In order to explore the relationship between intelligibility and processing speed we investigated intelligibility and other processing speed tasks. In these tasks the stimuli used were limited to a closed set of six highly confusable CV syllables (stop consonants plus the vowel /a/). This limited set was chosen because it provided a demanding discrimination task for listeners and would therefore be sensitive to both intelligibility and processing time differences. These stimuli were processed to give three levels of distortion, as in experiment 1.

To establish that the stimuli in each of the conditions were all highly intelligible subjects were presented with the stimuli from the three S/N conditions in a randomised order at 60 dB SPL. There were ten presentations of each stimulus and the subject's task was to press one of six buttons corresponding to the speech sound they heard. Response time as well as accuracy were recorded. The accuracy results showed that the speech stimuli in each S/N condition were intelligible with all scores approaching 100 percent (99.3%;99.7% & 98.2% for the 0,3 & 6 dB S/N conditions respectively). The reaction time data showed slight differences between the conditions with the slowest reaction time for the most distorted speech (the 0 dB S/N condition) lending some support to the claim by Pisoni,et al (1) that measures of processing speed can often show differences between speech signals which are not differentiated on suprathreshold intelligibility tests because of ceiling effects. However, the effects here are very slight and do not conclusively distinguish the test conditions.

These same stimuli were also used in two processing speed tasks including a two alternative forced choice task (2AFC) and a forced choice comparison task.

In the 2AFC task the subject heard a stimulus and chose from two response alternatives which stimulus was presented. The emphasis in the task is on speed of response although accuracy as well as reaction time are recorded. The results indicate a small but significant difference between the three S/N conditions, showing an increase in processing time with an increase in the amount of distortion. Accuracy results again indicate high intelligibility and no differences between the different conditions.

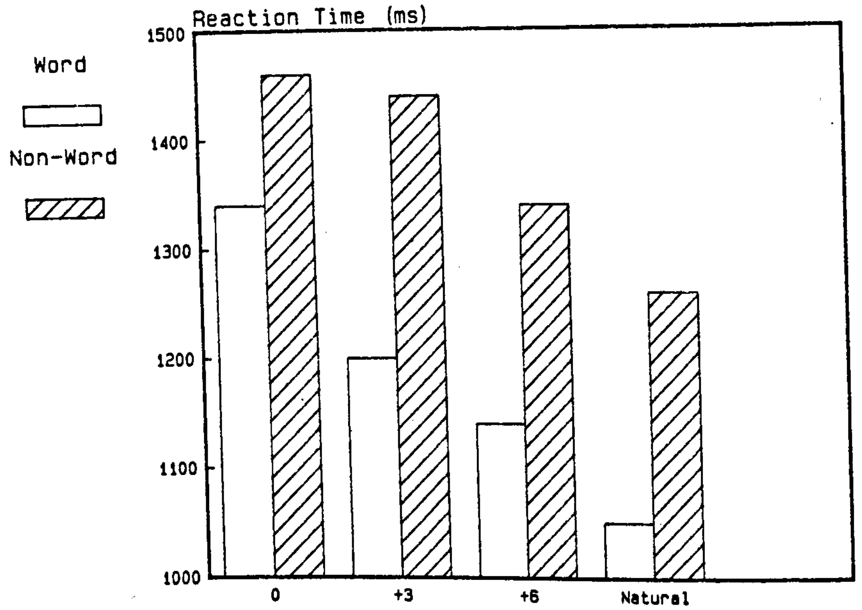


Figure 1
Results of the lexical decision task for processed and natural speech.

In the forced choice comparison task, the subject heard an undistorted CV syllable followed one second later by a tone and another CV syllable which was distorted. The subject's task was to make a yes/no judgement as to whether the second syllable was the same as the first syllable. Subjects were required to make these judgements as quickly as possible. The results indicated that for all conditions there was no difference in accuracy judgements but that subjects required longer processing times with the more distorted stimuli.

These data from the processing measures show consistent processing speed differences between the stimulus conditions in the absence of intelligibility differences, suggesting that the increased processing time is the result of higher level cognitive factors rather than difficulties at the representational level,as found in synthetic speech by Pisoni, et al (1).

To investigate this discrepancy we carried out two further intelligibility measures to ensure that the segmental intelligibility of the stimuli in each of the conditions was in fact the same. These measures included an adaptive speech test using the PEST procedure (3) and a stimulus repetition task (4).

In the PEST procedure the subject was required to press one of two buttons in front of them to indicate which stimulus was presented. The response alternatives changed on each trial and were displayed on a screen. The subject's responses were monitored for proportion correct and if this fell below a specified criterion, then the stimulus level was increased. If it fell above the specified criterion then the stimulus level was lowered. In this way, the presentation level of each stimulus was changed depending on the performance of the subject. The testing was continued until a specified criterion of performance was achieved. The results for the PEST procedure are expressed as the dB level at which the speech recognition threshold was achieved. The results show significant differences between the thresholds for the three conditions. The recognition threshold for the least distorted conditon (+6) is 2 dB better than for the +3 condition, which in turn is about 2dB better than the 0 condition.

The second intelligibility measure contrasted the subject's performance when the stimulus was presented once per trial compared to three repetitions of the stimulus before a response was required. Clark, Dermody & Palethorpe (4) found this procedure differentiated between synthetic and natural speech, with natural speech showing a significant increase in intelligibility with three repetitions while synthetic speech did not improve. In the present study the speech stimuli were presented near the 50% correct level (based on the OdB condition). Subjects were presented with the single repetition or the three repetition condition in a counterbalanced design. The results indicate that there is a repetition effect in each test condition and that the least distorted condition produces the greatest repetition effect. That is, speech in each of the test conditions is processed in a similar way to natural speech in the Clark,et al (4) study, but with a slightly reduced effect because the speech stimuli were more distorted. These results are similar to the results for the lexical decision task which also showed that the distorted stimuli behaved in a similar manner to natural speech in that case with longer processing time.

CONCLUSIONS

The results from experiment 2 suggest that even when processed natural speech is highly intelligible at suprathreshold levels, it can still produce slow processing times compared to natural speech. The results of a sensitive speech intelligibility task using the PEST procedure indicates that there are slight but significant differences for recognition of the processed speech which produce the slower processing times. This result is consistent with the notion that poorer acoustic phonetic processing will slow processing time for synthetic speech which is impoverished compared to natural speech as suggested by Pisoni, et al (1). The present study extends this finding to natural speech that has had noise added.

The present results suggest i) that high intelligibility at suprathreshold levels should not be used as a sole criterion for speech transmission if comparision with natural speech is intended and ii) that sensitive measures of both intelligibility and processing time can be used to differentiate processed natural speech from natural speech in listening performance when suprathreshold intelligibility of the processed speech is equivalent to natural speech.

REFERENCES

(1) PISONI,D.,NUSBAUM,H.& GREENE,B. (1985)"Perception of synthetic speech generated by rule",Proceedings of IEEE, 73,1665-1676.

(2) SCHROEDER,M.(1968)."Reference signal for signal quality studies",Journal of Acoustical Society of America, 44,1735-1736.

(3) TAYLOR,M.& CREELMAN,C.(1967)"PEST: efficient estimates on probability functions",Journal of Acoustical Society of America,41,782-787. for the three distortion

(4) CLARK,J.,DERMODY,P.& PALETHORPE,S.(1985)"Cue enhancement by stimulus repetition: natural and synthetic speech comparisons",Journal of Acoustical Society of America, 78, 458-462.

LIKENESS FUNCTIONS OF THE ACOUSTIC PATTERNS AS AN INDEX FOR OBJECTIVE ESTIMATION OF SPEECH TRANSMISSION QUALITY

Czesław Bąsztura

Institute of Telecommunication and Acoustics Technical University
of Wrocław ul.B.Prusa 53,55 50-317 Wrocław POLAND

ABSTRACT

The purpose of the paper is a presentation of a new objective measure for estimation of speech transmission quality and to perform a preliminary evaluation of conformity between the results obtained by means of the proposed method and subjectively measured speech intelligibility. The new method uses likeness functions of the acoustic patterns as an index for the evaluation of speech transmission quality. Eight likeness functions as distance or proximity meanures, i.e. Hamming, Euclidean, Minkowski, Chebyshev, Camber, Chi-square, Tanimoto and derectional cos, were investigated. As the test signal three key phrases of natural speech were used. The preliminary results indicate the possibility of good estimation of speech transmission quality by measuring and counting the likeness functions, especially by means of Hamming, Euclidean, Minkovski and Chebyshev distance measures.

INTRODUCTION

Speech intelligibility as a measure of speech transmission quality may be classified as either subjective or objective. The subjective measurement is a procedure for determining the communication channels intelligibility using a predetermined vocabulary and selected speakers and listeners panel.
Subjective measurement techniques generaly attempt to determine intelligibility for an information presented in one of the following three forms:
a) nonsense syllables (logatoms) list
b) limited list of words,
c) list of sentences.
In subjective measurement methods the inteligibility is determined by the ability of the listeners to identify spoken (or recorded) syllables, logatoms, words or sentences. A number of subjective methods have been devised with the desirable results. However, the reguirements for listeners panels greatly restricts the utility of these methods, and a long-sought goal is to replace the subjective scoring with objective measurements.. An objective measure for the fidelity of a speech communication system is a measure that is computed from data which contain no human subjective response.

There is a hypothesis that it is possible to design a relatively compact objective measures which are in a good correlation with subjective results over a subset of distortions and disturbances introduced by speech transmission channels [1,2,5,6]. Over the years some number of papers contained informations about evaluation of speech transmission systems by objective measures [1,2,7,8].

These measures include signal to noise ratios, arithmetic and geometric spectral distance measures (Viswanathan et al. [7], cepstral distance measures (Barnwell et al. [1,2]), various parametric distance measures such as pseudo-area functions and log are a fuctions from LPC analysis (Gray and Markel [5]), MTF (Steeneken and Houtgast [6]), and many more [7,8].

The task of comparing and contrasting the validity of such measures is immense. To check the validity of a particular objective measure over a given class of distortions and disturbances, a researcher must create a data base of distorted speech and corresponding data base of subjective results.

The essential features of computation of a relationship between subjective and obje--ctive measures are illustrated in Fig. 1

METHOD

In further analysis a following basic definitions and presumptions were mode:

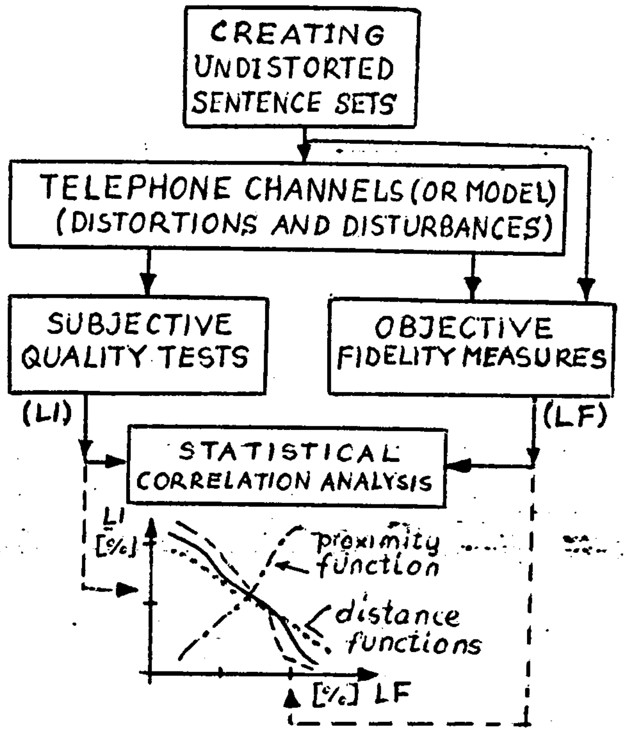

CREATING UNDISTORTED SENTENCE SETS
TELEPHONE CHANNELS (OR MODEL) (DISTORTIONS AND DISTURBANCES)
SUBJECTIVE QUALITY TESTS
OBJECTIVE FIDELITY MEASURES
(LI)
(LF)
STATISTICAL CORRELATION ANALYSIS
LI [%]
proximity function
distance functions
[%] LF

Fig.1 The computation of a relationship between objective and subjective quality measure.

a) The typical telephone channels, represented by an adjustable model of the the telephone channel as an investigated object, are used.

b) The criterion reference for objective measure is subjectively measured logatoms speech intelligibility.

c) The measurement conditions for both subjective and objective procedures are the same.

d) As a test signal for objective measurements three following Polish key phrases were used:

1. ALO (part of word "Hello"),

2. JUTRO BĘDZIE ŁADNY DZIEN ("Tomorrow will be a fine day")

3. SPRAWDZENIE PRZYDATNOSCI FUNKCJI PODOBIENSTWA DO OCENY JAKOSCI TRANSMISJI SYGNAŁU MOWY.
("Verification of the likeness functions usefulness to evaluation of speech signal transmission quality".)

e) As an objective measure, i.e. distance and proximity measures, eight likeness functions: Hamming, Euclidean, Minkovski, Chebyshev, Camber, Chi-square, Tanimoto and directional cos were examined.

The likeness function have a form :

$$d^{MIN}(X,Y)=\left[\sum_{p=1}^{P}|X_p-Y_p|^r\right]^{\frac{1}{r}} \quad r\geqslant 1 \qquad 1/$$

were: p=1,2,.. P p – dimensionality of vector parameters from speech signal.

Xp– p– th element of reference vector (from the undistorted speech signal)

Yp– p– th element of vector (from a distorded speech signal)

For r=1 d^{MIN} – Hamming distance
r=2 d^{MIN} – Euclidean distance.

Chebyshev distance:

$$d^{CZE}(X,Y) = \max_{p}\,(X_p-Y_p) \qquad 2/$$

Camber distance:

$$d^{CAM}(X,Y) = \sum_{p=1}^{P}\frac{|X_p-Y_p|}{|X_p+Y_p|} \qquad 3/$$

Chi-square distance:

$$d^{CHI}(X,Y)=\sum_{p=1}^{P}\frac{1}{X_p+Y_p}\left[\frac{X_p}{\sum_p X_p}-\frac{Y_p}{\sum_p Y_p}\right] \qquad 4/$$

Directional cos proximity:

$$b^{COS}(X,Y)= \frac{XY^{Tr}}{|X||Y|} \qquad 5/$$

Tanimoto proximity:

$$b^{TAN}(X,Y)=\frac{XY^{Tr}}{XX^{Tr}+YY^{Tr}-XY^{Tr}} \qquad 6/$$

EXPERIMENT AND RESULTS

First step in the experiment is a choise of an adequate test signals. Second problem relies on finding an effective set of parameters representing the test signal and presence of distortions and disturbances in these signals.

Next problems depend on the assumed method (Compare "METHOD").

The analysis of the previous investigations [3, 4] shows a fairly large effectiveness of representing the voice and speech features by parameter set with being the distribution of the time intervals between the zero - crossing of a speech signal [Fig.2]

$$Y \overset{col}{=} \{Y_1, Y_2, \cdots Y_p \cdots Y_P\} \qquad 7/$$

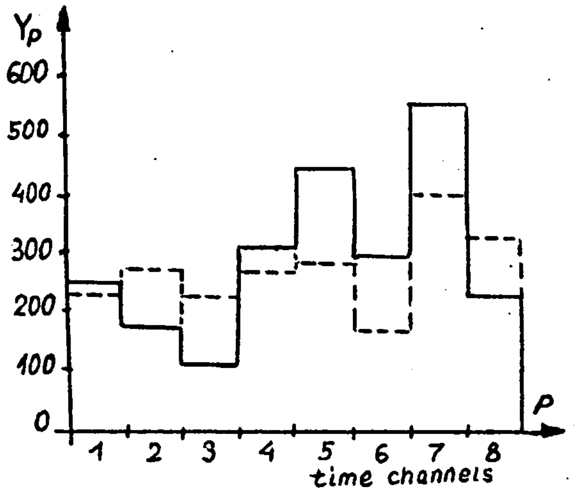


Fig.2 An example of the time intervals distribution between the zero-crossing of a speech signal (for two voices, the same text), P=8).

An especial computer program counted likeness functions (as objective measures) and made statistical correlation analysis of figure of merit LI=function of (LF).
Some of the results are shown in Fig. 3.

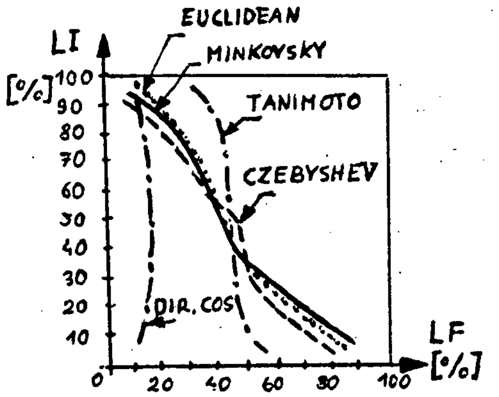


Fig.3 Examples of the statistical dependences of LI and LF.

Comparison of the results of the experiments on the 40 different telephone channels implemented by a physical model of telephone channel permits the following observations and conclusions to be given:

a) Hamming distance gives a good correlation LI with LF for all of the test signals (especially for 3 degrees of polynomial regression)

For example:

$$LI = 92{,}91 - 0.0121\, d^{HA} + $$
$$+ 0{,}151 \cdot 10^{-5} (d^{HA})^2 - 0.631 \cdot 10^{-10} (d^{HA})^3 \qquad 8/$$

b) Euclidean and Minkovski (r=3) distance (likewise the Hamming distance) give a good correlation for third key phrase. For example:

$$LI = 90{,}2 - 0{,}0184\, d^{EU} + $$
$$+ 0{,}364 \cdot 10^{-5} (d^{EU})^2 - 0{,}24 \cdot 10^{-9} (d^{EU})^3 \qquad 9/$$

$$LI = 89{,}6 - 0{,}0203\, d^{MIN} + $$
$$+ 0{,}448 \cdot 10^{-5} (d^{MIN})^2 - 0{,}33 \cdot 10^{-9} (d^{MIN})^3 \qquad 10/$$

c) Chebyshev distance gives a good correlation (especjally for third key phrase and for higer degree of polynomial regression).

d) Camber distance and Tanimoto and directional cos proximity did not give satisfactory results.

The future investigations will concentrate mainly on the selection of new test signals and other parameters of the speech signal.

LITERATURE

1. Barnwell T.P." Objective measures for speech quality testing", J.A.S.A. 66 (6), Dec. 1979,pp.1658-1663.

2. Barnvell T.P. " Correlation analysis of subjective and objective measure for speech quality". Proc. ICASSP 80,New york, 1980.

3. Basztura Cz., Majewski W. " The application of long term analysis of the zero-crossing of a speech signal in automatic speaker identification " Archives of Acoustic, 3.1,1978,pp.3-15.

4. Basztura Cz., Jurkiewicz J., " The zero - crossing analysis of a speech signal in the short term method of automatic speaker identification " Archives of Acoustic, 3, 3, 1978, pp. 185-195.

5. Gray A.H., Markel J.D., "Distance measures for speech processing" IEEE Trans. on Acoustic, speech and Signal Process, ASSP - 24, No.5, 1976,pp.380-391.

6. Steeneken H.J.M., Houtgast T.," A physical method for measuring speech transmission quality ", JASA, 67 (1). Jan. 1980,pp. 318-326.

7. Viswanathan R., " Objective speech quality evaluation of narrowband LPC vocoders" Proc. IEEE, No.6, 1978, pp. 591-594.

8. Voiers W.D., " Diagnostic acceptability measure for speech communication systems". Proc. IEEE ICAPP. 1977,pp. 204-207.

9. Second Draft Proposal ISO/DP 4870 for Acoustics - Recommended method for the construction and calibration of speech intelligibility tests -October. 1976.

SPEAKER RECOGNITION FROM PHONATED VS. WHISPERED VOWELS UNDER DIFFERENT FILTERING CONDITIONS

WIM A. VAN DOMMELEN

Institut für Phonetik und
digitale Sprachverarbeitung
Universität Kiel
2300 Kiel, FRG

ABSTRACT

The perceptual contribution of glottal source and vocal tract characteristics to speaker recognition was investigated in two listening tests. A group of eight female speakers produced sustained /e/ and /o/ vowels in isolation, both whispered and phonated. 500 ms portions of these vowels were used as stimuli under different filtering conditions (0-1, 1-2, 2-5 and 0-5 kHz). The results indicate that neither these filtering conditions nor vowel quality exert systematic influence upon speaker identification. Glottal source information, however, proved to be of considerable perceptual importance.

INTRODUCTION

Relatively little is known about those perceptual cues in the acoustic speech signal that contribute to the recognition of speakers by the human listener. Following up investigations reported in the literature, e.g. /1, 2/, this paper examines the role of glottal source and vocal tract information via a direct comparison of speaker identification rates for phonated vs. whispered vowels. By taking isolated vowels spoken on a monotone speaker-specific supraglottal timing characteristics and pitch movements are ruled out as possible cues. At the same time, the question as to whether there are specific frequency domains of special importance was investigated by band-pass filtering and by the use of two vowels with different spectral composition.

PROCEDURE

A group of eight female German speakers (students of phonetics at Kiel University) produced sustained /e/ and /o/ vowels in isolation, both whispered and phonated. They were instructed to approximate vowel durations of about 1-2 seconds, a condition which was fulfilled with ease by all subjects. A second requirement concerned the phonated vowels, which had to be produced on a monotone. The pitch level, however, could be freely chosen at an individually comfortable level. Auditory examination as well as an analysis of the fundamental frequency showed that the vowels were produced with only slight perturbations, which were very unlikely to contribute to the recognition of the individual speakers.

After 5 kHz low-pass filtering (12 dB/octave) the speech material was digitized at a sampling rate of 10 kHz and manipulated in the following way: From the middle of each vowel a 500 ms portion was taken to serve as the raw material for a stimulus. Subsequently, loudness differences between the vowel portions were auditorily equalized by amplitude manipulation. After digital-to-analogue conversion the speech samples were filtered with three different bandpasses (0-1 kHz, 1-2 kHz and 2-5 kHz; 24 dB/octave) and re-digitized. Starting from this material, following a second digital-to-analogue conversion two stimulus tapes were constructed, containing whispered and phonated vowel portions respectively. Each tape comprised 64 stimuli (8 speakers x 2 vowel qualities x 4 filtering conditions; the fourth condition being 0-5 kHz) in a randomized order. Each stimulus was composed of a 100 ms warning tone and subsequent 0.5 sec pause followed by a vowel portion, repeated four times at intervals of 1.5 sec, and an ensueing 6 sec response pause.

The group of speakers, who were all well-acquainted with one another, also acted as listeners, being presented with the stimuli over a loudspeaker in a quiet room. The two experiments (whispered and phonated vowels respectively) were performed in two sessions separated by a week. The listeners indicated their answers by writing down the perceived identity of the speakers on a prepared answer sheet.

RESULTS

For clarity of presentation we will deal firstly with the effect of the different filtering conditions upon the speaker identification, secondly with the influence of vowel quality and, finally, with the role of glottal source. The statistical significance of the identification results was tested by means of Wilcoxon matched-pairs signed-rank tests.

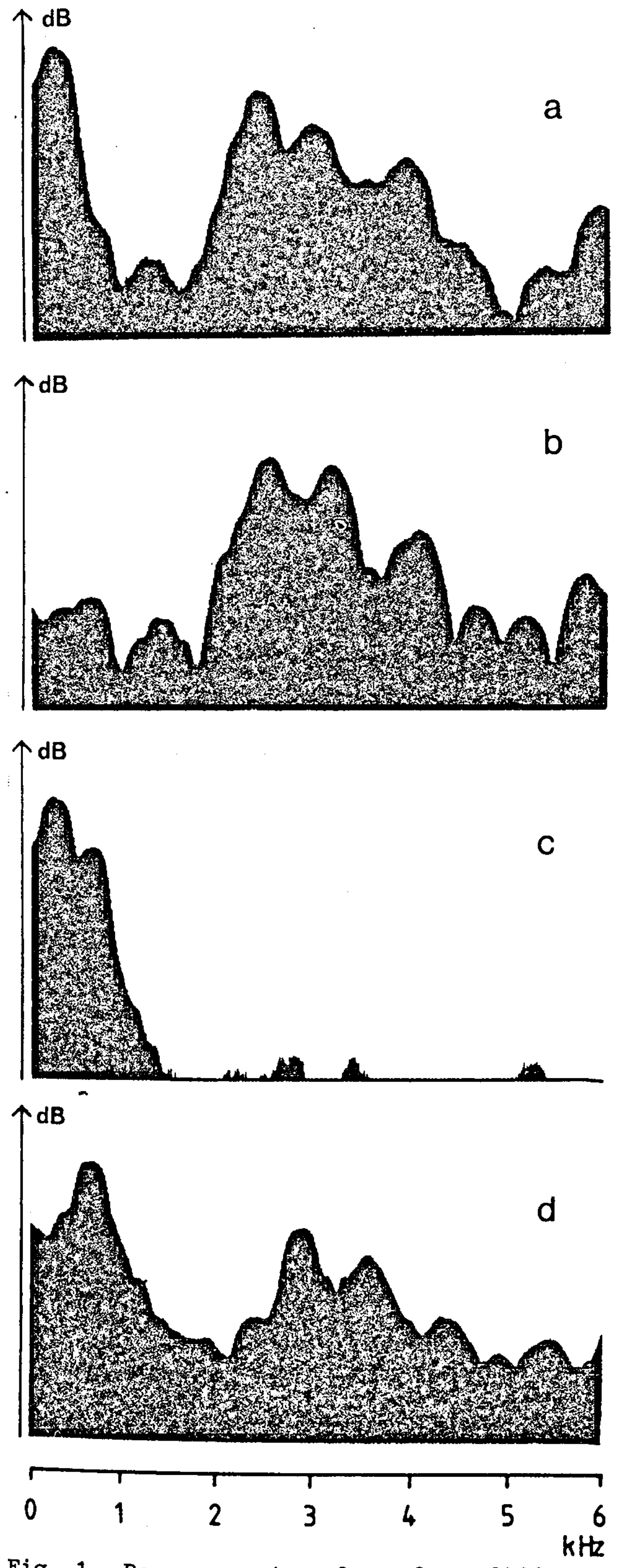


Fig. 1. Power spectra from four different vowels (speaker 2)
/e/: phonated (a) and whispered (b)
/o/: phonated (c) and whispered (d)

Filtering conditions

The overall spectra of the vowel portions are differently shaped (cf. Fig. 1), firstly due to the different positions of the first four or five formants in /e/ vs. /o/ (vowel quality) and, secondly, due to differences between the glottal spectra in phonated vs. whispered vowels (glottal source). Therefore, it might seem plausible to expect an effect of different filtering conditions upon the recognition rates (cf. Table I). Although in both cases there were sizeable differences between the best and the worst scores (16% and 13% respectively), they failed to reach statistical significance.

Table I
Overall identification scores
(percent correct)

Phonated vowels				
0-1	1-2	2-5	0-5	kHz
29	35	35	45	%

Whispered vowels				
0-1	1-2	2-5	0-5	kHz
13	21	10	23	%

Vowel quality

The overall recognition rates for /e/ vs. /o/ amounted to 34% vs. 38% in the case of phonated vowels and to 17% vs. 16% for the whispered vowels. In view of the small differences between the two conditions it is not surprising that they were statistically insignificant. This insignificance also holds when the individual filtering conditions are treated separately.

Glottal source

In contrast to both factors discussed above, the identification rates were affected by the glottal source parameter in a consistent way. At 36% the overall correct identification rate for phonated vowels lies significantly above that of 17% for the whispered ones (1% level). The effect holds for both vowel qualities (over all four filtering conditions) as well as for the filtering conditions (over the two vowel qualities), except for 1-2 kHz.

There are two aspects of the glottal source that might be responsible for the consistently higher identification scores in the case of the phonated vowels. In the first place, it is thinkable that speaker-specific pitch height was used as a primary cue in the identification task.

Alternatively, the spectrum of the glottal excitation is a possible candidate. With both possibilities in mind the data were examined further. Unlike the glottal spectrum, on which no data were available, mean fundamental frequency could be calculated for each speaker and turned out to vary between ca. 180 Hz and 250 Hz. Subsequently, a rank was given to each speaker, firstly according to their fundamental frequency (where a rank of 1 meant the speakers' own F0, a rank of 2 the next nearest pitch value etc.), and secondly according to their perceptual confusion with other speakers (over both vowel qualities and all filtering conditions; a rank of 1 standing for the highest recognition rate etc.). Calculation of Kendall correlation coefficients showed significant relationships (for one speaker at the 5% level; otherwise 1%) with values of r= 0.57, 0.67, 0.69, 0.72, 0.73, 0.76, 0.84 and 0.96. Obviously, speakers with similar fundamental frequencies are far more likely to be confused than speakers showing different pitch height. So it seems that the listeners relied upon the F0 factor to a varying, sometimes rather high degree in their identification of the various speakers.

Following a procedure similar to the one described above, correlations between perceptual confusions in the phonated vs. the whispered condition were calculated. Since the information present in whispered vowels is almost exclusively vocal tract information, it was postulated that high correlation rates might indicate a high perceptual value of such information. These correlation rates turned out to be significant for only two speakers (r= 0.64 and 0.81 respectively, at the 5% and 1% level respectively). Overall recognition rates for these speakers happened to be the highest ones (48% and 47% respectively for phonated vowels as against 34% and 28% for whispered vowels). Therefore, it seems likely that vocal tract information served as a perceptual cue in these two cases in addition to the glottal source parameter.

DISCUSSION

Of the three factors investigated in this paper, various filtering conditions, vowel quality and glottal source, only the latter turned out to have a systematic influence upon the speaker identification scores. The enhancing effect of glottal source information on identification can probably be accounted for by the speaker-specific pitch height, which is in line with the findings of Compton /1/.

Further, the results suggested a predominance of the glottal source parameter over vocal tract filtering characteristics. This confirms the findings of Lass, Hughes, Bowyer, Waters and Bourne /3/ for speaker sex identification. Possibly due to the use of synthetic stimuli instead of natural speech Lehiste and Meltzer /4/ arrived at the opposite conclusion, whilst LaRiviere /2/ found both factors to contribute about equally to speaker recognition. One should note, however, that in the present paper no data about the contribution of the glottal spectrum was available, so that its minor relevance had to be inferred from the data on fundamental frequency and from the mostly weak correlations between perceptual confusions for phonated vs. whispered vowels.

The fact that the four filtering conditions failed to influence the listeners' identification judgements may be due to there being only 16 stimuli in the sample (8 speakers x 2 vowel qualities per filtering condition; cf. the clearer effect of 24% for 1020 Hz low-pass vs. 1020 Hz high-pass found by Compton /1/ using considerably more stimuli). However, with vowel quality the sample size (32) is twice as big (8 speakers x 4 filtering conditions per vowel quality); this increases the likelyhood of the vowel quality results being representative.

REFERENCES

/1/ A. Compton, "Effects of filtering and vocal duration upon the identification of speakers, Aurally", Journ. of the Acoustical Society of America 35, pp. 1748-1752, 1963.

/2/ C. LaRiviere, "Some acoustic and perceptual correlates of speaker identification", Proc. of the Seventh Int. Congr. of Phon. Sci. (Rigault, A.; Charbonneau, R., eds.); Mouton: The Hague/Paris, pp. 558-564, 1972.

/3/ N.J. Lass, K.R. Hughes, M.D. Bowyer, L.T. Waters, & V.T. Bourne, "Speaker sex identification from voiced, whispered, and filtered isolated vowels", Journ. of the Acoustical Society of America 59, pp. 675-678, 1976.

/4/ I. Lehiste, & D. Meltzer, "Vowel and speaker identification in natural and synthetic speech", Language and Speech 16, pp. 356-364, 1973.

SPEAKER RECOGNITION BY MEANS OF SHORT SPEECH SEGMENTS ANALYSIS USING TIME-VARYING LINEAR PREDICTION IN LATTICE FORMULATION

JANUSZ ZALEWSKI
Technical University of Wroclaw,50-370 Wroclaw,Poland

Abstract - This paper presents the method of speaker recognition.In this technique the reflection coefficients obtained from short speech segments by means of time-varying linear prediction in lattice formulation procedure was utilized as the identification parameters and the minimum of time-average spectral difference between the corresponding short speech segments was the recognition criterion. The results of the recognition task using this method has been compared with others.

INTRODUCTION

The procedure utilized in any approach to speaker identification could substantially influence the resulting level of the ultimate identification accuracy of the used technique.In this regard,two distinctly separate operational phases may be identified for any approach of this type. First,the identification parameters and associated measurement technique must be chosen.Secondly,statistical distance measurement and associatted decision criterion must be identified and evaluated.

In the research we have previously reported /1,2,3/ the speaker has been represented by some phonemes,or short segments of speech regarded as the reference samples. The minimum cumulated distance measure between corresponding test and reference samples was the decision criterion.The method we have presented may be succesfully used as a procedure for identifying individuals from their speech.- at last under laboratory conditions.The parameter sets,chosen for speech waveforms parametrisation was the predictor coefficients and the cepstrum obtained via parametric analysis of speech signals,using an autoregressive model.For linear predictive coding,it is asumed that the signal is stationary over the time of analysis,and therefore the coefficients given in this model are constants.However speech signal to be modeled,even in short segments as are the phonemes, is not stationary.Therefore it seems to be reasonable to use an autoregressive signal modelling in which the coefficients are time-varying i.e. each coefficient in the model is allowed to change in time,by assuming it is a

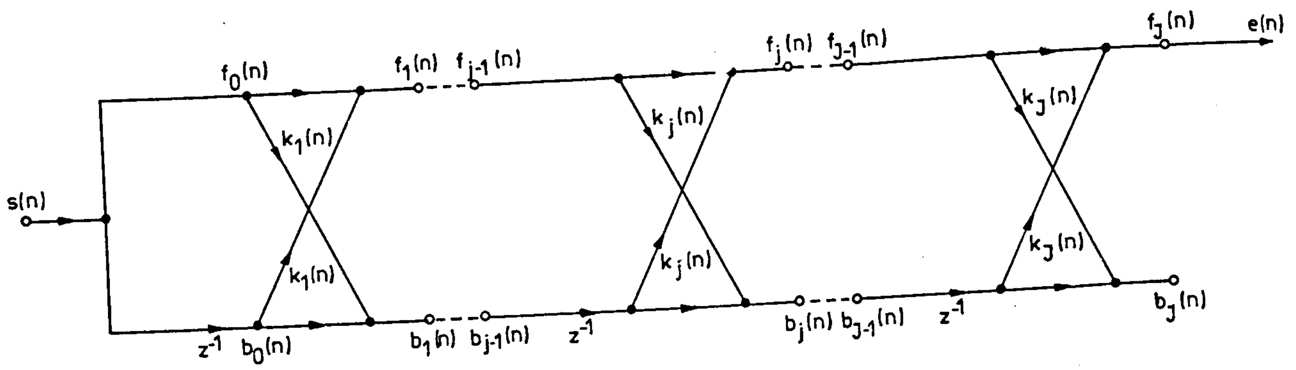


Fig. 1. The inverse filter in lattice form

linear combination of some set of known time functions.This model allows for continuously changing behavior of the signal,such propriety should enable the model to have possible better accuracy and allows for the analysis over longer data windows.

THEORETICAL BASES

The fundamental works on linear prediction of time-varying signals was done by Liporace /4/, Hall/5/, Hall et al. /6/ Turner and Dickinson/7/,and Jurkiewicz/8/. In present research it was utilized time-varying linear predictor in lattice formulation done by Jurkiewicz, who has reformulated the linear predictive technique to estimate the variable parameters $k_j(n)$ of the inverse filter in lattice form,as depicted in Fig.1, rather than in direct form. That is the inverse filter is in lattice form, and its parameters $k_j(n)$ are estimated by minimizing the given (after Burg) MSE norm /8/.

$$D_J = \sum_{n=o}^{L} (f_J^{2}(n) + b_J^{2} (n)) \qquad (1)$$

where

$$f_j(n) = f_{j-1} (n) + k_j(n).b_{j-1}(n-1), \qquad (2)$$
$$b_j(n) = b_{j-1} (n-1) + k_j(n).f_{j-1}(n) \qquad (3)$$
$$f_o(n) = b_o (n) = s (n) \qquad (4)$$
$$k_j(n) = \sum_{l=o}^{N-1} q_{lj} . u_l(n) \qquad (5)$$

and $u_l(n)$ are the time series (eg trigonometric functions as in Fourier series)

Denoting

$$f_j^{\prime} (n) = f_{j-1} (n) \qquad (6)$$
$$t_j^{\prime} (n) = b_{j-1} (n-1) \qquad (7)$$

and omitting the subscript j to simplificate notation,equations (1) - (5) become

$$D = \sum_{n=o}^{L} (f^2 (n) + b^2 (n)) \qquad (8)$$

$$f(n) = f^{\prime}(n) + k(n).b^{\prime}(n) \qquad (9)$$
$$b(n) = b^{\prime}(n) + k(n).f^{\prime}(n) \qquad (10)$$

$$k (n) = \sum_{l=o}^{N-1} q_l . u_l (n) \qquad (11)$$

Minimizing the error D with respect to each coefficient q_l by setting

$$\frac{\partial D}{\partial q} = 0 \quad ,l=o,1,..N-1 \qquad (12)$$

yelds the linear normal equations

$$q_l . R_{il} = S_i ,i= 0,1, ..N-1 \qquad (13)$$

where $$R_{i,l} = \sum_{n=o}^{L} u_i (n). u_l(n).d(n) \qquad (14)$$

$$S_i = \sum_{n=o}^{L} u_i (n) . c(n) \qquad (15)$$
$$c(n) = - 2 f^{\prime}(n) . b^{\prime}(n) \qquad (16)$$
$$d(n) = f^{\prime 2}(n) + b^{\prime 2} (n) \qquad (17)$$

The coefficients q are specified by the equation (13) , or in matrix form

$$R \times Q = S \qquad (18)$$

Below is the complete algorithm for described time-varying linear prediction in lattice formulation:

In each j-th step of analysis (i.e. in j-th section of the filter:

-the matrix R and the vector S are computed from equations (14),(15),(16) and(17)

- the set of equations (13) or (17) are solved,

•the signals $f^{\prime}(n)$ and $b^{\prime}(n)$ are filtered according eq (9) and (1o) in the lattice system (Fig.1)

This set of operations is repeated in each succeding step j, for j=1,2 to J.

In the experiments described in this paper each of 1o reflection coefficients $k_j(n)$ was evaluated, according eq (9), as the linear combination of 3 or 5 time functions.

$$u_i(n) = \begin{cases} 1 & i=o \\ \cos(n(i+1)\pi/m), & i \text{ odd} \\ \sin(n\, i\, \pi / m), & i \text{ even} \end{cases} \qquad (19)$$

$i= o,1,..., N_T$,

$N_T= \Omega_c M /T$,

8 Ω_c = digital cut-off frequency of $k_i(n)$ spectrum,

M = period of the $u_i(n)$ functions set.

For each sample,from the filter parameters trajectories $k_i(n)$,the 10 sets of 40 cepstrum coefficients was evaluated; each set at one of 10 equidistant time instants. From the cepstrum coefficients of the reference sample c_i, and those of the test sample c_i, the time average spectral difference (i.e. the time-average Euclidean distance of c_i and c_i sets,multiplied by 10/ln 10) was computed. The time-average spectral difference is

$$d = 10\cdot(\log e)\cdot(L^{-1} \sum_{l=1}^{L} 2 \sum_{k=1}^{K} (c_k-c_k)^2)^{1/2}$$

where $c_k = c_k(l)$ are the cepstral coefficients of the test sample,$c_k^{\prime} = c_k$ l) are the cepstral coefficients of the reference sample, l - succeding time instant at which the cepstra are evaluated,L - number of time instants at which the cepstra are evaluated (here 1o), K number of cepstral coefficients representing the sample (here 4o)

EXPERIMENTAL PROCEDURE

Subjects were the same 20 male speakers as in speaker recognition experiments /3/,where speakers have been represented by some phonemes,and the parameter set for speech waveforms parametrisation,was the predictor coefficients,obtained using autoregressive model with constant coefficients.The speech material consisted of 240 utterances,including 2 repetitions of 6 Polish vowels /a, o,e,i,u,y/each spoken in two contexts.The speech signal was manually segmented,to detach the vowels,pre-emphasized 6 dB per octave,low -pass filtered with 5 kHz cut-off frequency,sampled at a rate of 10 ksamples per second and converted into digital form by means of 8-bit A/D converter.The segments of 100 ms duration was processed to obtain 10 time-varying reflection coefficients trajectories. To compare test and reference samples, the average spectral differences between them was computed. In the first speaker recognition experiment the speakers were represented by a single phoneme,in the second by pairs (15 combinations),in the third by three (20 comb.)and in the 4-th - by four phonemes (15 comb.). The minimum distance criterion was used as the decision rule,i.e. the m-th test sample was considered to be identical with the n-th reference,if for j=1 to 20 and j≠n,d_{mj} d_{mn} where d_{mn} denote the distance measure (average spectral difference) between the m-th test and the n-th reference sample.

RESULTS AND CONCLUSIONS

The detail results of all 112 recognition experiments will be presented at the Conress.Hereafter are presentedsome typical results obtained in two experiments,first where the speakers were represented by phoneme "i" and second where the representation included phonemes "i" and "a".The results are compared with results of experiments with parametrization obtained using constant model. In table 1, the average recognition errors are shown; subscript i denotes the first experiment,subscript a,i denotes the second,subscript v - variable model, subscript c - constant model.

TABLE 1. RECOGNITION ERRORS

$E_{i,v}$	$E_{i,c}$	$E_{a,i,v}$	$E_{a,i,c}$
0.050	0.183	0.000	0.008

Several conclusions can be drawn from the result of this research.First it may be stated that representation of speakers by short speech segments and comparison of corresponding segments may be succesfullly used in a procedure for identifying individuals from their speech. Second, the time-varying linear prediction procedure in lattice formulation is a convenient form of the parame trisation procedure.Finally,it is shown that augmenting the number of speech segments representing the speaker,could possibly result in an even more powerful identification process.

REFERENCES

/1/ J.Zalewski et al.,"Speaker recognition by means of linear predictor coefficients", Proc 9ICA,Madrid,1977,I-32,438

/2/ j.Zalewski et al.,"An application of the Itakura distance measure for the estimation of the predictive coded pattern similarity" /in Polish/,Proc XXIV Open Acoust. Seminar. ,Gdańsk,1077,Part 1 ,pp380,381

/3/ J.Zalewski,"A comparison of the effectiveness of some distance measures in spe speaker recognition experiments", Paper on the Speaker Recognition Working Group on theTenth International Congress of Phonetic Sciences, Utrecht, 1983 (also Reports of Techn.Univ.of Wroclaw,I-28/PRE-033/183 Wroclaw 1983)

/4/L.A. Liporace,"Linear Estimation of Nonstationary Signals", J.Acoust.Soc.Am., vol 58,no 6,December 1975,pp 1268-1295

/5/ M.G. Hall,"Time-Varying Linear Predictive Coding of Speech Signals,S.M. Thesis, Dept of Electrical Engineering and Computer Science, Mass.Inst.of Techn.,Cambridge, Massachusettes,August 1977

/6/ M.G.Hall et al. "Time-Varying parametric modelling of speech.", Proc. of the 1977 IEEE Conf. on Decision and Control, New Orlean, Dec 1977, pp 1095 - 1091

/7/ J.Turner E.Dickinson,"Linear prediction applied to time-varying all-pole signals",Proc.1977 ieee Int.Conf.on Acoust. Speech and Sign. Proc.,Hartford,Conn., 1977,pp 750 - 753

/8/ J.Jurkiewicz " Time-varying Linear Prediction in Lattice Formulation for Speech Analysis",Ph.D.Dissertation,Inst. of Telecommunication and Acoustics,Techn. University of Wroclaw,Wroclaw 1984.

A METHOD OF AUTOMATIC SPEAKER RECOGNITION IN OPEN SETS

WOJCIECH MAJEWSKI, CZESŁAW BASZTURA, JERZY JURKIEWICZ

Institute of Telecommunication and Acoustics
Technical University of Wrocław, Poland

ABSTRACT

The paper presents a method of automatic speaker recognition in open sets ensuring a good effectiveness of elimination of strangers' voices, i.e. the voices that do not belong to a given set of known speakers. The applied procedure is discussed and description of speaker recognition experiments based on this procedure presented. The results obtained for a test material consisting of speech samples produced by 10 known speakers and 10 other speakers are very promissing /99 % of correct elimination of strangers' voices/ and confirming the pertinence of theoretical assumptions.

INTRODUCTION

In tasks of automatic speaker recognition such situations may occur in which it cannot be assumed that an unknown voice to be recognized belongs to a known set of classes of voices /closed set/. Thus, a problem arises to work out an algorithm of recognition that could operate in open sets of speakers, i.e. with no assumption that a speech sample of an unknown speaker must belong to one speaker from a given set of speakers. The idea of such approach to the problem of automatic speaker recognition was presented to the 10 ICPhS [1] . The present paper contains the analysis of this problem taking as a basis the classical Bayes's decision criterion. One of the main purposes of this study was to perform the analysis of probability of error and risk connected with a decision-making process in open sets with regard to the selection of discrimination threshold and the manner of approximation of conditional distribution of strangers' voices.

THEORETICAL BASES

In automatic voice recognition speech samples are represented by their patterns, i.e. multidimensional vectors of parameters in observation space X^K /K - space dimension/. The vectors x extracted from speech samples of particular speakers form distributions characterized by densities of conditional probabilities $Q(x|m)$, where m is a speaker number or generally a class. It may be assumed that these distributions are normal distributions expressed by the formula /Fig.1/:

$$Q(x|m)=(2\pi)^{-\frac{K}{2}}|B_m|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-W_m)^{Tr}B_m^{-1}(x-W_m)\right\} \quad /1/$$

where B_m - covariance matrix for a class

W_m - mean vector for a class

$$W_m = \frac{1}{M}\sum_{i=1}^{I_m} x_{m,i}$$

$m=1,2,..M$ M - number of classes

$i=1,2,..I_m$ I_m- number of utterance repetitions for a class

Tr - sign of vector transposition

In recognition process the classical Bayes's decision criterion considers a probability $P(m|y)$ with which a test pattern y represents the class m.

$$P(m|y) = \frac{Q(y|m)P_m}{\sum_{l=1}^{M} Q(y|l)P_l} \quad /2/$$

where P_m - probability of appearances of patterns from a given class

The classical approach to recognition problem relies on finding a minimal risk $R_m(y)$ connected with assigning the pattern y to the class m.

$$R_m(y) = \sum_{l=1}^{M} C_{m,l} Q(y|l) P_l \qquad /3/$$

where $C_{m,l}$ - element of decision matrix representing the cost of decision resulting from assigning the pattern from the class l as belonging to the class m [2].

In case of speaker recognition in open sets the set of classes consists of M known classes /closed set/ and one multiobject class corresponding to all other voices that do not belong to the set M. These voices constitute so called "ground" or strangers' voices class /m = 0/. The conditional distribution $Q(x|0)$ of ground is in general case a multimodal distribution with parameters that are not known.

Considering these assumptions the recognition procedure in open sets may be presented as consisting of two stages:

1. Identification in the closed set, i.e. finding m^* for which

$$R_{m^*}(y) = \min_{m} R_m(y) \qquad /4/$$

what means a temporary assigning a test pattern y to the class m^*.

2. Verification, i.e. checking the condition

$$R_{m^*}(y) < R_o(y) \qquad /5/$$

If the condition /5/ is fulfilled, the pattern y belongs to the class m^* ; in the opposite case it belongs to the class m = 0, i.e. the ground.

It is to observe that the formula /4/ permits to devide the parameter space into M subspaces X_m^K. Similarly, the inequality /5/ defines areas X_{Wm}^K in subspaces X_m^K /Fig.2/. It follows that it is not necessary to know the total distribution of $Q(x|0)$, but only the limits of areas X_{Wm}^K or $Q(x|0)$ distribution in the vicinity of these limits. Thus, the $Q(x|0)$ distribution may be approximated by means of M planes, one by one for each subspace X_m^K.

$$Q(x|0) = G_m(x) \qquad m : \quad x \in X_m^K \qquad /6/$$

$$\text{where } G_m(x) = g_{m,0} + \sum_{k=1}^{K} g_{m,k} + x_k \qquad /7/$$

is the equation of plane m in subspace X_m^K /Fig.2/. Discontinuities of such approximation at the borders of subspaces X_m^K are insignificant for the verification process.

RECOGNITION ERRORS

The information about errors is contained in the statistics of identification and verification shown in Fig.3. In this figure particular symbols have the meaning:

N - number of voice patterns
W - patterns belonging to the closed set
O - patterns from beyond the closed set
P - initially correctly recognized
B - initially incorrectly recognized
A - accepted by verification
E - eliminated by verification

For example N_{WPE} indicates the number of patterns from the closed set, correctly recognized by the classifier, but next rejected in the verification process.

Within the closed set the statistics of incorrect recognitions is represented by:

$$\delta = \frac{N_{WB}}{N_W} \qquad /8/$$

Verification procedure devides this error into two components:

$$\delta_A = \frac{N_{WBA}}{N_W} \quad \text{and} \quad \delta_E = \frac{N_{WBE}}{N_W} \qquad /9/$$

and introduces verification errors: the error of incorrect rejection expressed as

$$\alpha' = \frac{N_{WPE}}{N_W} \qquad /10/$$

$$\text{or as} \quad \alpha'' = \frac{N_{WPE} + N_{WB}}{N_W} = \alpha' + \delta \qquad /11/$$

$$\alpha''' = \frac{N_{WPE} + N_{WBE}}{N_W} = \alpha' + \delta_E \qquad /12/$$

and - in relation to the open set - also the error of false acceptance:

$$\beta = \frac{N_{OA}}{N_O} \qquad /13/$$

EXPERIMENTAL PROCEDURE

Speaker recognition experiments in open set were performed in the following conditions:

a/ A specific cue material was used. It was a Polish sentence "Jutro będzie ładny dzień" /Tomorrow it'll be a fine day/. Distributions of time intervals between zero-crossings [3] were extracted from this sentence and used as vectors of parameters. The dimension of observation space was K = 4 /the parameters of the largest discrimination power were selected/.

b/ The learning sequence consisted of 100 vectors $x_{m,i}$ obtained from I_m = 10 repetitions of the utterance by M = 10 speakers.

c/ The testing sequence consisted of 10 other repetitions of the utterance by 10 speakers from the closed set and 10 repetitions by 10 speakers from beyond the closed set. Thus, the open set contained 200 vectors $y_{m,i}$ obtained from 20 speakers.

Since the main concern of this study was verification procedure for a fixed measurement set-up, the experiments were arranged in such a way that first speaker identification procedure was applied to the total testing sequence and next verification procedure was utilized for different values of verification parameters.

In the recognition process $m^*_{m,i}$, $R_{m^*}(y_{m,i})$ and $R_o(y_{m,i})$ were calculated for each pair m,i /see formulas 3,4 and 5/, assuming that the elements $C_{m,l}$ of matrix C are equal 1 in case of incorrect decisions or 0 in case of correct decisions.

From the possible ways of $Q(x|0)$ approximation /eq. 6 and 7/ two simple cases were investigated in the experiments /see Fig.2/:

$$1^o \quad Q(x|0) = H \qquad H = \text{const} \qquad /14/$$

$$2^o \quad Q(x|0) = H_m \qquad m : x \in X_m^K \qquad /15/$$

For the first case the decision threshold was defined as

$$H = \gamma Q_{av} \qquad Q_{av} = \frac{1}{M}\sum_{m=1}^{M} Q(W_m|m) \qquad /16/$$

where γ - coefficient /experiment parameter/

For the second case two versions of defining individual thresholds H_m were distinguished:

$$H_m = \gamma Q(W_m|m) \qquad /17/$$

$$\text{and} \qquad H_m = \gamma_m Q(W_m|m) \qquad /18/$$

where γ_m - coefficient selected individually to minimize verification risk for a given class on the basis of α' and β errors.

RESULTS AND CONCLUSIONS

The results of the experiments are set together in Table 1 which presents the errors for different approximations of $Q(x|0)$ and γ values that minimize the verification risk. The influence of γ coefficient on α' and β errors for the case nr 1 /eq. 14 and 16/ is shown as example in Fig. 4. Analyzing the data presented in Table 1 it may be noticed that the speaker recognition scores are very little differentiated in the examined cases. This may be the result of very effective discriminating power of the vectors applied and/or too small size of the test set. It is, however, necessary to emphasize that the results obtained confirmed the pertinence of methodological assumptions what was the main purpose of this study.

The methodological considerations permit to state that the proposed method of voice recognition in open sets is very elastic and it enables to adjust the global characteristics, i.e. α and β errors, to adopted strategy of recognition system. For a given set of patterns describing the voices it is always possible to optimize the recognition by a proper selection of approximation of the ground class distribution, i.e. by proper selection of decision threshold. It is a basic advantage of the presented method of speaker recognition in open sets verified experimentally for a test population of 20 speakers.

Table 1. Recognition errors

Case	1°-eq16 γ=2.10^{-5}	2°-eq17 γ=2.10^{-4}	2°-eq17 γ=3.10^{-4}	2°-eq18 indγ_m
Error	%	%	%	%
δ	8	8	8	8
δ_A	6	6	6	6
δ_E	2	2	2	2
α'	1	2	1	1
α''	9	10	9	9
α'''	3	4	3	3
β	1	1	2	1

REFERENCES

1. W.Majewski, Cz.Basztura, Speaker recognition in open sets, Proceedings of the Tenth International Congress of Phonetic Sciences /M.P.R. Van den Broecke and A.Cohen eds./, Foris Publications, Dordrecht, 1984, 322-325.

2. J.Z.Cypkin, Podstawy teorii układów uczacych sie, WN-T, Warszawa, 1973.

3. Cz.Basztura, W.Majewski, The application of long-term analysis of the zero-crossing of a speech signal in automatic speaker identification, Archives of Acoustics, 3, 1, 1978, 3-15.

FIGURES

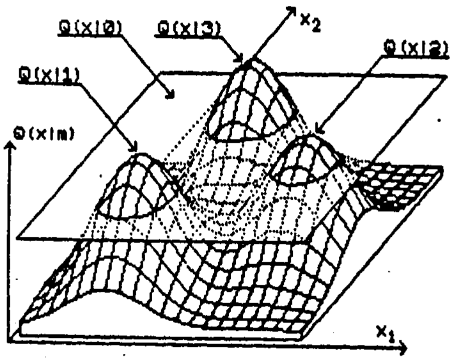


Fig.1. Examples of Q(x|m) distributions in case of two dimensional space /K = 2/.

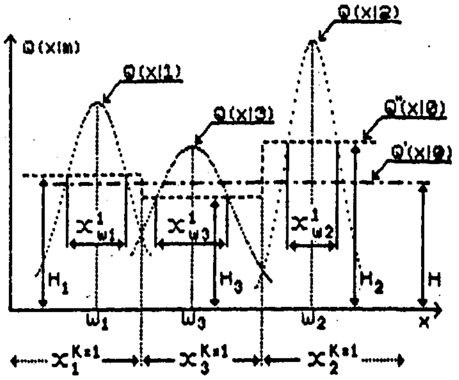


Fig.2. Illustration of approximation of Q(x|0) and determination of decision threshold H for one dimensional space.
Q'(x|0) - case nr 1 /eq.14/
Q"(x|0) - case nr 2 /eq.15/

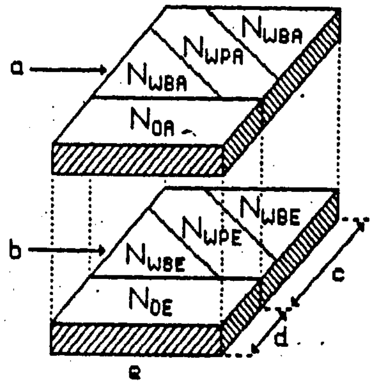


Fig.3. Statistics of recognitions in open sets; a - recognitions accepted by verification, b - recognitions rejected by verification, c - patterns from the closed set, d - patterns from beyond the closed set, e - classes.

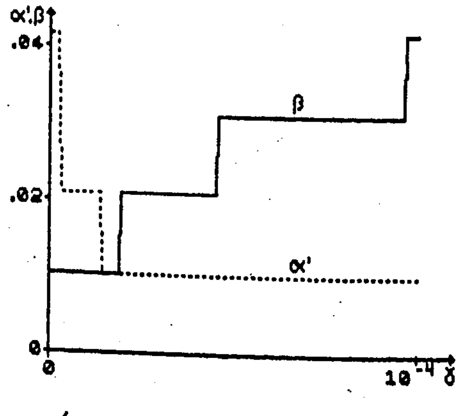


Fig.4. α and β errors in the function of γ for the case nr 1 /eq.14 and 16/.

AN EXPERIMENT IN INTER-LANGUAGES SPEAKER RECOGNITION USING THE SDDD INDEX

Bernard HARMEGNIES Albert LANDERCY Marielle BRUYNINCKX

Université de l'Etat à Mons - Département de Phonétique et Psycho-acoustique - Avenue du Champ de Mars, B-7000 MONS - BELGIQUE

ABSTRACT

The voices of 10 Belgian bilingual (Dutch-French) subjects were analysed by means of a high resolution frequency analyser (400 channels FFT). Long Term Average Spectra (LTAS) of the subjects'voices were computed both on the basis of French and of Dutch utterances (balanced texts). The SDDD index was used in crder to compare these LTAS. Its discriminating ability in an inter-language speaker recognition task was evaluated by means of the Receiver Operating Characteristics (ROC) curves for all the comparison conditions under investigation and revealed to be greater than the one of the cross-correlation coefficient.

INTRODUCTION

Although Long Term Average Spectra (LTAS) have been used in various contexts and are usually considered as good acoustical cues to voice quality, several of their properties are not yet well known. Among others, the question of the LTAS resistance to changes in the languages used by speakers is still controversial.

On the one hand, several experiments suggest that languages exert strong effects on LTAS. KIUKAANNIEMI and MATTILA report differences between Finnish and English data [1]. HALLE, de BOYSSON - BARDIES and SAGARD suggest that even LTAS from 8 and 10 month old babies can be influenced by the language of the social group they belong to [2]. MAJEWSKI and HOLLIEN [3] and ZALEWSKI, MAJEWSKI and HOLLIEN [4] obtain recognition rates different for Americans and Poles; this seems to suggest language-related LTAS differences. On the other hand, some authors consider that LTAS are language-independant to some extent. BYRNE [5] notices that LTAS he drew from English texts uttered by Australian look much like those from ANIANSSON [6] drawn from Swedish speech. On the basis of an experiment involving Piamontes, Italian and French, TOSI [7] concludes that each speaker possesses "relative" LTAS invariance, irrespective of the language spoken. HARMEGNIES and LANDERCY [8] report few differences between LTAS drawn from Dutch and French utterances produced by bilingual subjects. As NOLAN remarks [9], there is a conflict between these research trends and it is unclear whether LTAS can be considered as language-independent cues to voices quality.

In this paper, which constitutes a contribution to this problem, we will study to what extent inter-language speaker recognition based on LTAS is possible. Because its discriminatory ability is supposed to overcome those of classical indices, a new dissimilarity index, SDDD [10, 11] will be used for the purpose of comparing spectra; its power will be assessed by comparison with the correlation coefficient.

EXPERIMENT

Experimental setting

The speakers were 10 bilingual Belgian subjects, between 18 and 21 years old. Each of them uttered two texts ten times in succession : a phonetically balanced French text and a phonetically balanced Dutch text. Both texts were about 18 seconds long. The recording sessions took place in a sound-proof room. The subjects were sitting in front of the microphone, placed at a constant 40 cm distance from their lips. All texts were recorded on a NAGRA IV S recorder, by means of a KM 84 NEUMANN microphone.

Acoustical analysis

The acoustical analyses were performed later by means of a 400-channels 2033 Brüel Kjaer FFT analyser (BK 2033). Its sampling frequency was set to 12.8 kHz, in order to obtain a 0-5 kHz frequency span. With this setting, the spectra presented a 12.5 Hz resolution over the whole frequency band under investigation. The BK 2033 built-in linear averaging process was used in order to compute LTAS. The 200 (10 subjects x 2 languages x 10 utterances) so-obtained LTAS were then transmitted from the analyser to a 4341

IBM computer via a personal computer, for storage and further computations.

Comparison procedure

Inter- and intra-language comparisons were performed. For intra-language comparisons, the same procedure was used both for the Dutch and the French LTAS : 1. (intra-speaker comparisons) for each of the 10 speakers, one comparison was performed for each possible non-redundant pair of his 10 LTAS (i.e. 45 comparisons); 2. (inter-speaker comparisons) for each possible non-redundant pair of different speakers (i.e. 45 pairs), all possible comparisons of their respective 10 LTAS were performed (i.e. 100 comparisons for one pair). For each language, 450 intra-speaker and 4500 inter-speaker comparisons were therefore performed.

For inter-language comparisons, all the French LTAS were compared with all the Dutch LTAS; 1000 intra-speaker and 9000 inter-speaker comparisons were therefore performed.

For each comparison, both a similarity (R) and a dissimilarity (SDDD) index were computed.

Indices for the comparison of LTAS

In order to define the indices, it is convenient to consider each LTAS as a K-dimensional vector, with k being the total number of frequency channels taken into account in the spectrum. Therefore, spectrum S may be defined as :

$$S = (S_1, \ldots, S_i, \ldots, S_k) \qquad (1)$$

with S_i, the level of the ith frequency component. In this paper as well as in most previous ones [1,4,8,10,12], the S_i values will be expressed in decibels.

The Bravais-Pearson cross-correlation coefficient (R) can be used as a similarity index for the comparison of LTAS. It expresses the tendency of the s_i values to covary with the S_i values and it ranges, in absolute values from 0 (complete independance of the S_i and S'_i variabilities) to 1 (perfect correlation of the S_i and S_i values). R can be defined as :

$$R_{SS'} = \frac{1}{K} \frac{\sum_{i=1}^{K} (S_i - M_S)(S'_i - M'_S)}{\sigma_S \sigma'_S} \qquad (2)$$

Where M_S and M'_S are the means for all S_i and S'_i values, respectively, and σ_S and σ'_S are the corresponding standard deviations. If the spectra beeing compared are identical, the correlation between the S_i and S_i values is 1. On the contrary, a weak correlation indicates a lack of similarity of the spectral shapes. R is usually considered as one of the best indices because : 1. it exhibits a discriminating ability in the same range than the one of other classical indices (e.g. the euclidean distance) [4]; 2. unlike other classical indices, R is insensitive to changes in the overall levels of the spectra and, therefore, does not require any intensity normalization.

The Standard Deviation of the interspectral Differences Distribution (SDDD) has been recently introduced [10, 11]. SDDD measures the variability of the $S_i - S'_i$ differences. It is defined as :

$$SDDD_{SS'} = \sqrt{\frac{1}{K} \sum_{i=1}^{K} \left[S_i - S'_i - MD \right]^2} \qquad (3)$$

where MD is the average of the $S_i - S'_i$ differences. If the shapes of the spectra compared are highly similar, the differences values are almost invariant and tend to concentrate around a given central tendency influenced only by the between-spectra overall level difference. If the shapes are different, one can find large level differences in certain frequency channels and small ones in others; the standard deviation of the differences increases. SDDD can therefore be used as a dissimilarity index for LTAS. Like R, SDDD is insensitive to changes in the levels of the spectra; moreover, in recent intra-language speaker recognition experiments [10, 11], SDDD has revealed to be more discriminative than R.

RESULTS

The distributions characteristics of the SDDD values drawn from all kinds of comparisons are presented in table 1. This table shows that, in the case of intra-language comparisons, the intra-speaker distributions are well separated from the inter-speaker distributions (French : mean SDDD intra = 2.912 against mean SDDD inter = 4.66; Dutch : mean SDDD intra = 3.162 against mean SDDD inter = 5.138). Nevertheless the separation of inter- and intra-speaker distributions is less important in the case of inter-language comparisons (mean SDDD intra = 3.941 against mean SDDD inter = 5.221). Similar observations can be drawn from table 2, about the distributions of the correlation index.

	Mean	Standard deviation	Extreme Values
French/French			
Intra spk.	2.912	.421	1.9-4.7
Inter spk.	4.966	.805	3.0-8.3
Dutch/Dutch			
Intra spk.	3.162	.510	2.0-4.9
Inter spk.	5.138	.789	3.0-7.8
French/Dutch			
Intra spk.	3.941	.514	2.6-5.6
Inter spk.	5.221	.866	2.9-8.7

Table 1 : Characteristics of the inter- and intra-speaker distributions of the SDDD values drawn from intra-language and inter-language comparisons.

In order to study more accurately the relationships between these distributions, we decided to plot the corresponding Relative Operating Characteristic (ROC) curves. For each comparison condition (French/French, Dutch/Dutch and French/Dutch), a series of values across the entire range of variation of each index were successively considered as rejection thresholds for a recognition task. The corresponding false alarm- and correct recognition rates were drawn from the observed distributions and considered as couples of coordinates in the ROC space. Six (2 indices x 3 comparison conditions) ROC curves were plotted this way (see fig. 1).

	Mean	Standard deviation	Extreme Values
French/French			
Intra spk.	.932	.002	.84-.97
Inter spk.	.805	.007	.52-.94
Dutch/Dutch			
Intra spk.	.928	.002	.81-.97
Inter spk.	.817	.006	.57-.94
French/Dutch			
Intra spk.	.886	.003	.78-.96
Inter spk.	.798	.007	.43-.94

Table 2 : Characteristics of the inter- and intra-speaker distributions of the R values drawn from intra-language and inter-language comparisons.

It is well known that the area enclosed in the entire ROC space beneath a ROC curve is a distribution-free measure of sensitivity [13]. It is therefore very easy, even from simple direct examination of figure 1, to perform a ranking of the six curves on the basis of the corresponding discriminative powers. In order of decreasing discriminating ability, this ranking is : 1. SDDD, intra-language comparisons (French/French); 2. SDDD, intra-language comparisons (Dutch/Dutch); 3. R, intra-language comparisons (French/French); 4. R, intra-language comparisons (Dutch/Dutch); 5. SDDD, inter-language comparisons; 6. R, inter-language comparisons.

In order to obtain more informative figures than this ranking, we also evaluated the surfaces beneath the curves. For this purpose, each curve was fitted by a polynomial function thanks to polynomial regression techniques. Polynomes of the sixth and seventh order were used and a good fitting was achieved in every case (residual sum of squares was in the .01 - .001 range). The polynomial functions were thereafter integrated. Table 3 gives the values of the surfaces thus obtained (taking into consideration that a unit-surface would mean perfect discrimination and a .5 surface would mean random recognition).

DISCUSSION

Both indices under investigation lead to the same main conclusion, i.e., LTAS-based speaker recognition is language dependant : the ROC curves clearly show that the inter-language comparisons are less speaker-discriminant than the intra-language comparisons. This finding seems to plead for the idea that languages exert some effects on LTAS. Moreover, both indices reveal better performances with LTAS drawn from French utterances than with those drawn from Dutch texts. This is in agreement with Majewski and Hollien's suggestion that the power of long term spectrum as an identification tool might be somewhat language dependant [3]. It should be noticed, however, that although LTAS turn out to be language-dependant, it still convey enough cues to the speaker's personnality to make inter language speaker recognition possible : our ROC curves demonstrate that the power of the inter-language recognition process is still far better than chance. In this sense, we can agree with Tosi's conclusion [7] that LTAS possess "relative" invariances, irrespective of the language they come from.

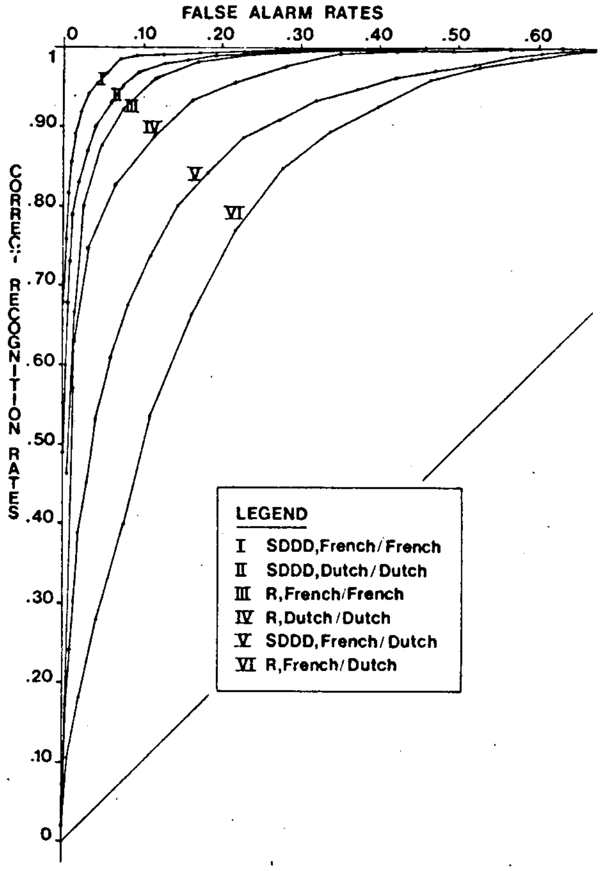


Figure 1 : Receiver Operating Characteristics curves for both indices in each comparison condition.

In other words, our general conclusion as to the LTAS resistance to changes in languages could be : long term spectra are influenced by the language spoken (at least when bilingual Dutch/French subjects read 18 seconds long texts), but the speaker influence is greater; LTAS-based inter-language recognition is therefore less safe than intra-language recognition, but quite possible.

This conclusion leads to the question of the relative power of inter-language recognition. Firstly, it is quite obvious, from figure 1 and table 3, that SDDD is more powerfull in all comparison situations. In a case where one suspects that the comparison situation could lower the discriminative ability of the comparison procedure (e.g. in the case of inter-language recognition), SDDD should therefore be prefered. Furthermore, if the ROC surfaces listed in table 3 are measures of the corresponding discriminative abilities, their ratios can inform about the relative powers of the indices in each situation. One can compute, this way, that the power of SDDD in inter-language recognition is about 92 % of its own power for intra-language recognition, although the power of R in inter-language recognition is only about 88 % of its own power for intra-language recognition. Thus, not only SDDD is more speaker-discriminant than R, it is moreover less sensitive to changes in the comparison conditions (at least language variations).

	SDDD	R
French/French	.994	.983
Dutch/Dutch	.987	.963
French/Dutch	.910	.856

Table 3 : Areas of the entire ROC space beneath each ROC curve.

As a final remark, we must emphazise that our data were collected on a restricted number of subjects only : any overgeneralization would therefore be hazardous. We nevertheless think that they convey structures strong enough to consider our findings at least as firm working hypotheses for our future research.

REFERENCES

[1] KIUKAANNIEMI, H.J. and P. MATTILA, "Long-term speech spectra : a computerized method of measurement and a comparative study of Finnish and English data", Scandinavian Audiology, 9, 1980, 67-72.

[2] HALLE, P., de BOYSSON-BARDIES and L. SAGARD, "Utilisation des spectres à long terme pour dégager des propriétés acoustiques des langues : étude comparative et développementale", Actes des 13èmes J.E.P., GALF, Bruxelles, 1984, 147.

[3] W. MAJEWSKI, and H. HOLLIEN, "Euclidean distances between long-term speech spectra as a criterion for speaker recognition" Speech perception and automatic recognition, vol. 3 (Communication seminar), Stockholm, Almqvist and Wiksell, 1974, 303-310.

[4] J. ZALEWSKI, W. MAJEWSKI, and H. HOLLIEN, "Cross correlation of long-term speech spectra as a speaker identification technique", Acustica, 34, 1975, 20-24.

[5] BYRNE, D., "The speech spectrum - some aspects of its significance for hearing aid selection and evaluation", British Journal of Audiology, 11, 1977, 40-46.

[6] ANIANNSON, G., "Speech discrimination predicted from tone audiometry and articulation index", Acta Otolaryngogica, Suppl. 320, 1974, 36-43.

[7] TOSI, O., Voice Identification, University Park Press, Baltimore, 1979.

[8] HARMEGNIES, B., and A. LANDERCY, "Language features in the long-term average spectrum", Revue de Phonétique Appliquée, 73-74-75, 1985, 51-68.

[9] NOLAN, F. The phonetic bases of speaker recognition, Cambridge University Press, Cambridge, 1983.

[10] HARMEGNIES, B. and A. LANDERCY, "Comparison of spectral similarity indices for speaker recognition", Proceedings of the 12th International Congress on Acoustics, vol. I, A1-4, Toronto, 1986.

[11] HARMEGNIES, B., "SDDD : a new dissimilarity index for the comparison of speech spectra" (submitted).

[12] HOLLIEN H. and W. MAJEWSKI, Speaker identification by long-term spectra under normal and distorted speech conditions, J. Acoust. Soc. Am., 62, 4, (1977), 975-80.

[13] SWETS, J.A., "The relative operating characteristic in psychology", Science, 182, 1973, 990-1000.

QUANTIFICATION OF A MULTI-SPEAKER DATABASE OF SPOKEN AUSTRALIAN ENGLISH

J.B.Millar

Department of Engineering Physics
Australian National University
Canberra, Australia, 2601

ABSTRACT

Measurements for the quantification of the speaker dimension of a database of spoken language are presented. These measurements are structured in two dimensions - utterance extent and acoustic feature. Utterance extent ranges from the macro-acoustics of long-term statistics to the micro-acoustics of the organisation of the syllable, while the acoustic features cover the phonetically motivated areas of Energy, Timing, Excitation, and spectral Colour. An example is given of the application of a small sub-set of these measurements to extract representative speakers from a 33-speaker database of spoken Australian English for the development of robust speech processing for a cochlear implant project.

INTRODUCTION

The acoustic-phonetic description of the speaker dimension of speech has received some solid attention in recent years [2,5]. These studies have approached the issue from the standpoint of phonetics with acoustics used to give quantitative backup to the phonetic judgements. In this paper, the starting point is acoustics, and the aim is to present a hierarchy of measurements that can specify speaker characteristics with increasing refinement, and to apply a subset of these measurements to a specific practical task.

Much work in acoustical speaker characterisation has in view high resolution speaker discrimination for a variety of purposes. One route to this goal is the direct route of looking for discriminating cues. This route however begs the question of the density of the speaker space from which test speakers are selected. My route is to first of all map out the speaker space and, with progressive refinement, plot out the trajectory of speakers within that space as they use language in different ways and over periods of time. This approach is motivated by the belief that speech technology needs an adequate model of speaker characteristics which is compatible with, complements, and interacts with linguistic models of speech.

I have selected four phonetically motivated dimensions which reflect the roles of diverse organs of the human body that are involved in the production of speech. Variance in speech timing, energy, excitation, and spectral colour can be uniquely and independently individual. Each of these dimensions may be examined at more than one level of utterance extent (Figure.1). As speech is implemented by sets of short-term articulatory gestures superimposed on longer term breath resources and even longer term patterns of articualtory status, my analysis strategy aims to see the short-term against the background of the long-term, and to use the long-term to constrain the short-term descriptions. Analysis across all the four feature domains and the timescales is in progress but only a subset are presented by results of its application to the speakers in a 33-speaker database of spoken Australian English.

	LONG TERM	BREATH GROUP	SYLLABLE
ENERGY	Long-term Energy	Multisyllable Pattern	Contour
TIMING	Overall Duration, and Breath Group Structure	Inter-syllabic Intervals	Duration
EXCITATION	Long-term Excitation Frequency Distribution	Intonation Pattern	Contour
COLOUR	Long-term Spectrum		Contour

Figure 1. Analysis structure for Quantification of of Speakers (see text).

SPEECH TIMING

The timing of speech is influenced by many factors including the cognitive process of assembling the message, and the musculature of respiration and articulation. The way in which these factors influence speech timing can be measured in different ways ranging from the macro-timing of overall duration of utterances, to the micro-timing of individual articulatory gestures. The overall duration of a lengthy utterance involving many respiratory cycles will encompass the summation of many different timing strategies. Measurement of overall duration is technically trivial but provides a foundation on which to build the timing picture for a speaker. The direct contribution of pauses for inspiration (respiratory) may be removed by the decomposition of longer utterances into single respiratory cycles or breath groups. Durations of breath groups and their variation reveal temporal aspects of speech planning related to the semantics and syntax of the utterance, and to the management of breath resources. This individually determined mapping of linguistic content onto the time course of a breath carrier will to some degree determine the repertoire of prosodic expression which may be used - each with its own breath resource penalty in addition to prosodic pattern and pattern continuity constraints. The complete picture of prosody management is undeniably complex but is here identified as a speaker variable begging quantification. At the simplest level syllabic rate and intersyllabic interval within breath groups provide a more accurate timing description than overall syllabic rate. Accurate temporal identification of the syllable provides the foundation for syllabic modelling. Sub-syllabic temporal segments are not attempted as this area is best catered for by coarticulatory models of the syllable.

SPEECH ENERGY

The acoustic energy inherent in speech sounds is controlled also by a number of organs. The movement of exhalatory muscles and the variation of constrictions at the larynx and other vocal tract sites influence the overall energy of a voice and the dynamic range of energy that contributes cues for the perception of stress and certain phonetic distinctions. The assessment of what constitutes significant energy to declare that speech is present, or of what constitutes significant change in energy to declare that a particular speech sound transition is in progress, underlies many speech measurements. Energy range characteristics have been shown to discriminate between a small sample of Australian and North American adult male voices [4], and the energy distribution of speech, non-speech, vowels, nasals, and fricatives have been shown to be distinctive [7].

In the current study a full-band long-term energy histogram has been routinely produced for each one-minute speech passage analysed. As speech comprises periods of silence, inhalation of breath, frication, and phonation, the energies of all these components are included in the histogram. The major features for most speakers are peaks at energies roughly corresponding to silence and phonation. The ratio of these varies from speaker to speaker as does the additional contribution of inhalation, often indistinguishable from silence energy, and frication, which most often supplements the low energy skirt of the phonation energy distribution. For all practical purposes we have a two-peak histogram (Figure 2). An initial interpretation of this histogram relates the overall energy of the voice to the difference between the two peak positions and may be considered a speaker characteristic if all recordings were made in the same environment with a constant ambient noise. Further speaker characteristics may be attributed to the shape and relative size of the energy peaks but have proved to be difficult to extract with any reliability.

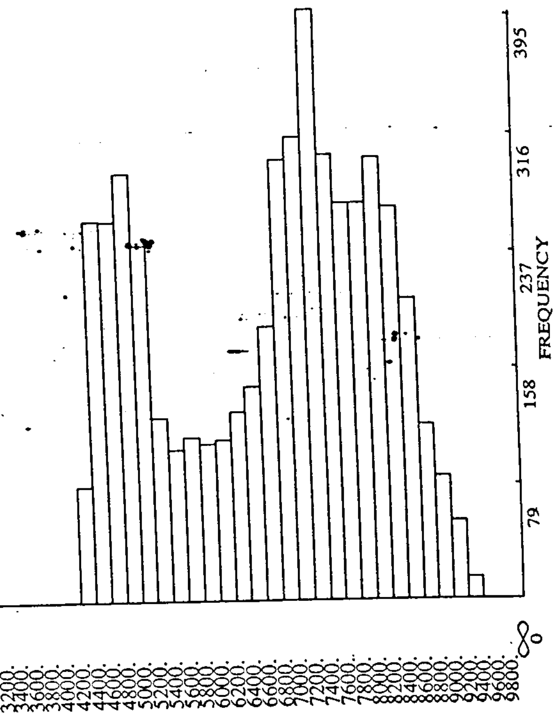


Figure 2. Distribution of occurences of energy levels in One-minute passage.

SPEECH EXCITATION

Speech excitation may be of several forms, a variety of kinds of phonation, frication caused by turbulence, or a mixture of frication and some kind of phonation. In the current analysis only phonation was measured by detecting excitation via transglottal electrical impedance. Instants of glottal closure were determined by a simple algorithm acting on the impedance waveform. Histograms of time intervals between closures were produced. The first order model was to assume uni-modal symmetrical distributions, and to parameterise this model in terms of the mean and standard deviation of the distribution (Figure 3).

SPEECH COLOUR

The acoustic wave at the termination of the vocal tract reflects the "colouring" of the excitation spectrum by the selective absorption of energy in the tract. Anatomical and habitual constraints on the shape of the vocal tract will give rise to long-term effects, whereas specific muscle gestures will characterise the articulation of individual sounds.

Speech colour has been measured at two levels. Firstly, we measured the long-term spectrum of one-minute passages of speech and viewed the results at three levels of resolution [3]. Secondly, we modelled the resonance patterns in syllables in order to express, in a speaker-specific way, the expected resonance trajectories between specific consonants and vowels [1].

THE DATABASE

Our database of spoken Australian English [6] comprises 15 male and 18 female speakers whose accent may be broadly classified as General Australian. In this paper we focus on the analysis of three reading passages, each designed to last one minute, which were recorded over a period of approximately 6 weeks from all 33 speakers. The recordings from each speaker were separated by at least 2 weeks in most cases. Passage A is an expository discourse from a popular scientific text (with some scient~~ific terms~~ omitted), passage B is a narrative from a childrens' book including a fair amount of dialogue, and passage C comprises two short passages often used in speech research - the fable of the "North wind and the Sun" (C1), and the "Rainbow" passage (C2).

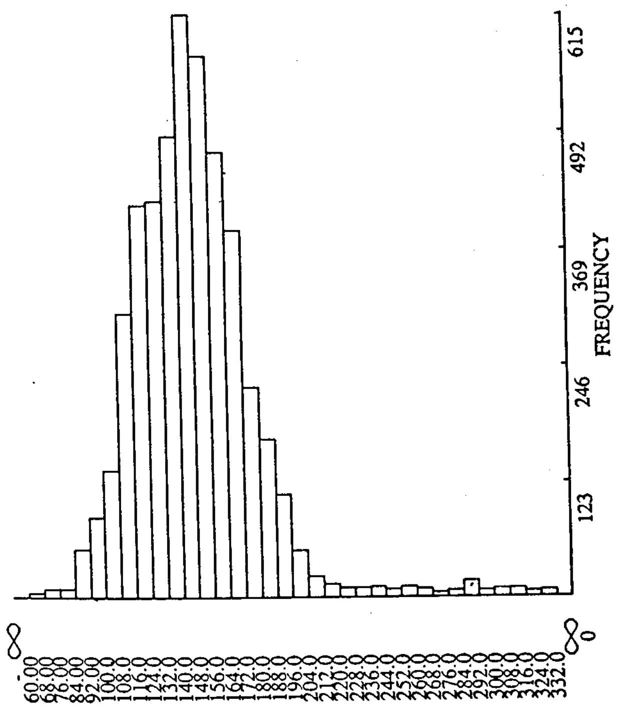


Figure 3. Distribution of intervals between glottal closures in a One-minute passage

POPULATION CHARACTERISTICS

Before looking at individual speaker variation across the reading passages, we gauge the overall population variance in timing and excitation.

The overall durational analysis showed strong consistency over all passages. The standard deviation of durations across all speakers was approximately 10% in nearly all cases. This result holds separately for the male and female sub-populations, and there was no significant difference between the overall durational characteristics of male and female sub-populations. The raw data are given in table 1.

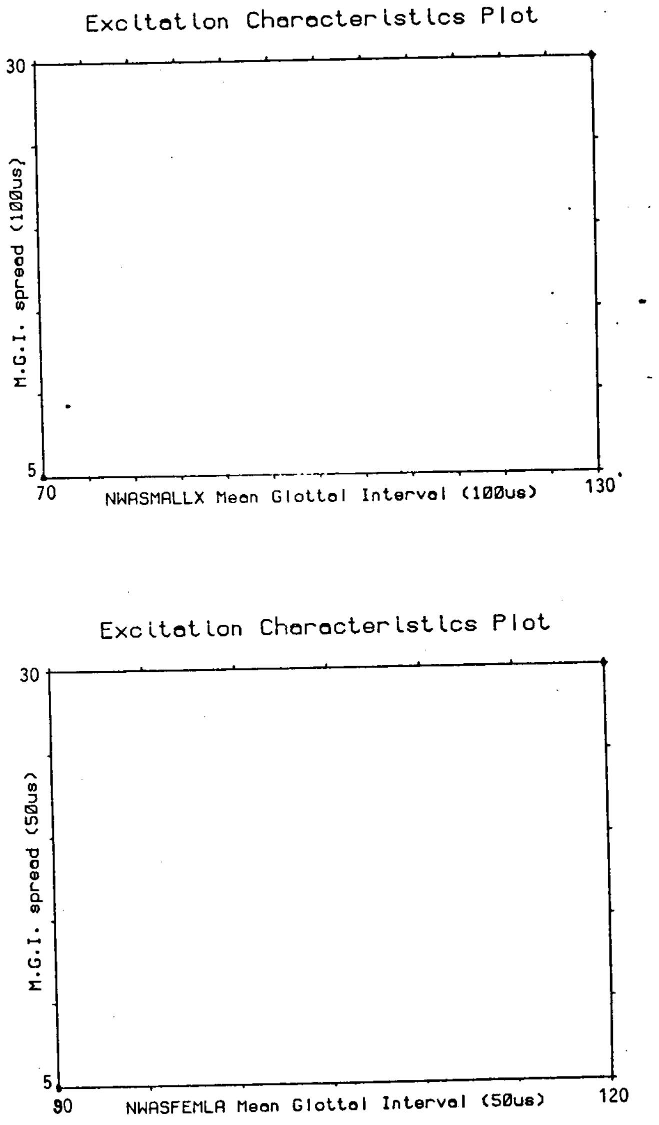


Figure 4. Scatter plots of mean glottal interval vs spread of glottal intervals in a 30s passage for (a) males, and (b) females.

Macro-timing analysis revealed several different types of speakers - those who are consistently slow (>1.2sd above mean), medium (within 0.5sd of mean), or fast (>1.2sd below mean) speakers, and those who vary their speed in a way that is influenced by the material.

Table 1. Overall Durations of reading passages.

Passage	MALE Mean	St.Dev.	FEMALE Mean	St.Dev.
A	66.24 sec	9.1%	69.44 sec	10.3%
B	69.57 sec	9.6%	70.78 sec	12.5%
C1	32.74 sec	9.5%	33.28 sec	8.6%
C2	31.19 sec	10.4%	31.85 sec	10.3%

The excitation analysis was performed on all those speakers for whom impedance waveforms could be measured. The population-wide results for the residual male and female subpopulations are given in table 2. This analysis has revealed speakers who exhibit diverse combinations of mean level of voice pitch and range of intonation. Within our sample the females tend to have a proportional relationship between pitch and intonation range (Figure 4a). The males however, show a more complex relationship with the most quantitatively monotonous voices being in the middle of the mean pitch range (Figure 4b). The mean standard deviation of inter-glottal closure intervals for males and females over all passages is in the range 15-20%. This is equivalent to a total range of less than 1 octave.

Table 2. Overall excitation characteristics in terms of parameters of the distribution of intervals between glottal closures.

	MALE				FEMALE			
	Median (ms)	Mean (ms)	SDof Mean (ms)	Mean ofSD (ms)	Median (ms)	Mean (ms)	SDof Mean (ms)	Mean ofSD (ms)
A	9.5	9.6	1.64	1.6	5.0	5.1	0.36	0.9
B	8.9	9.0	1.27	1.8	5.1	5.2	0.37	0.9
C1	9.4	9.5		1.6	5.2	5.2		0.8
C2	9.4	9.5		1.7	5.2	5.2		0.8

CLASSIFICATION OF SPEAKERS

Macro-acoustic analyses in the domains of timing and excitation have been used to select four speakers whose speech can be used for testing speech processing algorithms for the Australian Cochlear Implant project at Melbourne University. It is necessary to thoroughly test algorithms for such speech processing on natural data that is representative of the kind of speech that the implantee is likely to receive.

The selected speakers conform to the following criteria:

1. As a group they represent both male and female speakers having mean fundamental frequencies approximately one standard deviation above and below the male and female population averages.

2. They have a speaking rate that is below the population average.

3. They have the highest range of fundamental frequencies consistent with the above requirements.

These criteria provide speech samples which may be used directly or in segmented form for perceptual experiments (2), which represents a representative range of fundamental frequencies (1), and in which there is plenty of speech dynamics (3). It was felt that these basic criteria provided, within the scope of four voices, a reasonable spread of speaker variance typical of the normal population that is relevant to this specific application. Other factors such as phonation type and vocal tract length may well be applied as analysis of the database proceeds.

SUMMARY AND DISCUSSION

The use of a structured and phonetically motivated set of measurements for quantifying speakers has been suggested. The scheme has been applied in part to a 33-speaker database of spoken Australian English, and preliminary results have been outlined. It has also been shown how such a technique can provide a small number of representative speakers for the evaluation of new developments in speech technology.

Such a hierarchy of measurements for progressively refined quantification of the speaker-space are most important for the development of speech technology which admits the use of multiple speakers. Speaker-independent automatic speech recognition needs to be tested against speaker groups which are evenly spread, or which have defined clustering characteristics. Only in this way will comparisons between techniques be valid. Conversely, many distinct voices in multi-voice speech synthesis may be produced when distributed evenly over the perceptual equivalent of the analytic speaker-space.

REFERENCES

[1] CLERMONT,F., MILLAR,J.B. (1986) "Multi-speaker validation of coarticulation models of syllabic nuclei", Proc. ICASSP-86, 2671-2674.

[2] LAVER, J.D.M. (1980) "The Phonetic Description of Voice Quality", Cambridge University Press: Cambridge.

[3] MILLAR, J.B. (1982) "Analysis of continuous speech for speaker characteristics", In J.E.Clark (Ed), "Collected papers on normal aspects of speech and language", Speech & Language Research Centre, Occasional Papers, Macquarie University.

[4] MILLAR,J.B., WAGNER,M. (1983) "The Automatic Analysis of acoustic variance in speech", Language and Speech, 26, 145-158.

[5] NOLAN, F.J. (198) "The Phonetic Bases of Speaker Recognition", Cambridge University Press: Cambridge.

[6] O'KANE,M., MILLAR,J.B., BRYANT,P. (1982) "A database of spoken Australian English: Design and Collection", Technical Note No.6, School of Information Sciences, Canberra College of Advanced Education.

[7] WAGNER.M. (1978) "The application of a learning technique for the identification of speaker characteristics in continuous speech", Unpublished Ph.D. Thesis, Australian National University.

EXPERIMENTAL EVIDENCE FOR PHONOLOGICAL UNITS

BRUCE L. DERWING MAUREEN L. DOW TERRANCE M. NEAREY

Department of Linguistics, University of Alberta, Edmonton, Canada T6G 0Z1

ABSTRACT

Experimental evidence is discussed that bears on the psychological status of three kinds of hypothesized units in phonology: the **syllable**, the **phoneme** (or **segment**) and a variety of **intermediate units** that have been proposed (viz., the **onset**, **rhyme**, **nucleus**, **coda**, **head**, and **margin**). Some of these units are seen to be more viable than others, though in all cases more cross-linguistic evidence is required.

INTRODUCTION: THE SYLLABLE AND THE PHONEME

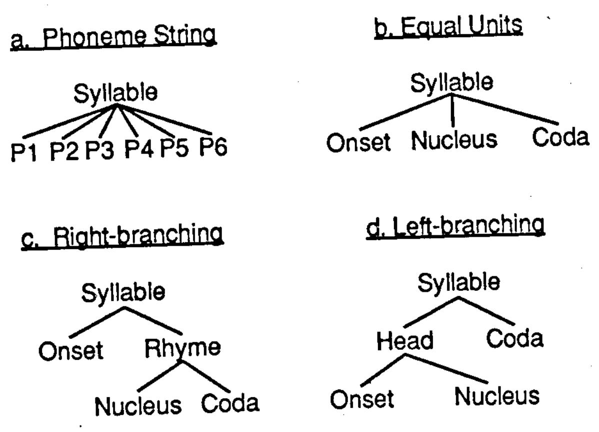
A variety of theoretical units have been proposed for the phonological description of languages and for the internal lexicon, but few experimental tests have as yet been carried out to determine which, if any, of these hypothesized units are psychologically real for speakers. Least controversial of the proposed units, perhaps, is the **syllable**, for, although much work certainly remains to be done on the question of actual syllable boundaries (not to mention cross-linguistic work on languages of different canonical types), it is nonetheless clear that something very close to the traditional notion of the syllable is a viable psychological unit for speakers of languages like English and Russian from the earliest linguistic stages of childhood (see especially [1,2,3] for evidence that even young children can do quite well at counting the number of syllables in words and can manipulate whole syllables in a variety of different experimental tasks).

Somewhat more controversial, however, is the notion of the **phoneme**, which corresponds in scope to the traditional idea of the individual consonant or vowel **segment**. Although most theorists in both phonetics and phonology have tacitly assumed that something akin to the phoneme is the basic unit of speech segmentation by speakers, experimental evidence in support of this position is neither abundant in quantity nor unequivocal in its interpretation. To begin with, for example, all three of the studies already cited above suggest that pre-literate children, at least, are much less successful at counting or otherwise manipulating individual phonemes than they are in dealing with whole syllables; and even after literacy is achieved, in fact, English-speaking children still exhibit some confusion in dealing with words for which the grapheme and hypothesized phoneme counts may differ (as in examples like *pitch* and *judge* [4]). Furthermore, though by now quite a variety of different controlled experimental techniques have been employed in the attempt to assess the psychological status of the phoneme (including concept-formation [5], string similarity judgments [6,7,8], and discrimination tasks [7,8], in addition to segment counting [8,9,10]), it is apparent that orthographic effects have contaminated the results of all of these studies at least to some extent, so that no clear answer has yet emerged on the phoneme issue (see especially [8] for a full discussion and review).

INTERMEDIATE UNITS

In addition to the syllable and the phoneme/segment, a number of units have also been proposed that are intermediate in scope between the two. Such **intermediate units** are highly controversial in two senses: (a) there is much theoretical disagreement on such fundamental matters as the actual number of such units and how they are organized in relation to one another, and (b) because, until the relatively recent advent of the so-called "metrical" approach to phonology (see [11] for an overview), few theorists subscribed to the idea that such units even existed. The classical view of the syllable, in other words, regarded it as a simple **linear** sequence of phonemes, with no internal structure beyond this (see Fig. 1a).

Figure 1. Alternative Models of Syllable Structure



Contemporary **hierarchical** models, however, assign to the syllable a kind of constituent structure not unlike that found in syntax. In one version of this model, for example, the syllable is broken down into the three intermediate units of **onset**, **nucleus** and **coda**, each of equal status (Fig. 1b). In another version of this model, the nucleus and coda are linked as part of an additional constituent called the **rhyme** (or **rime**), yielding the right-branching structure shown in Fig. 1c. In a third version of this model, a left-branching structure is proposed by grouping the nucleus with the onset, rather than with the coda, thus supplanting the rhyme constituent in favor of what is called here the **head** (Fig. 1d). Finally, a third model, less prominent in the literature, provides for significant groupings of segments that are linked, not in a hierarchical way, but rather in terms of "affinity bonds" of varying strengths [12]. Such a **bonding** model allows for the introduction of significant units such as the **margin** (equivalent to Vennemann's "shell", i.e., onset + coda) that would involve cutting across constituent boundaries in the hierarchical framework.

Though research on this question has been quite limited to date, there is a modicum of evidence that attests to the viability of at least some intermediate units of the kinds described. Hierarchical models seem to account better for speech error data than does the simple linear model, for example [13], and some reading research has further suggested that even graphemes may be read preferentially in terms of onset and coda units, rather than as arbitrary letter pairs or triples [14]. More recently, in a series of experiments involving the use of novel word games, Treiman found that rules involving onsets, rhymes, and, to some extent, codas were easier to learn than rules that involved breaking up these units [15]. Such units are part of all three versions of the hierarchical model, however, not to mention the bonding model, so a new series of experiments was recently performed in our laboratory in the hope of further clarifying the situation (see [16] and especially [17] for further details).

Experiment 1 was a simple unit-counting task performed by young (K and G1) children and by high school students. As expected from prior studies, the former performed best on the syllable counts (61% correct) and worst on the phoneme counts (12%); a new finding was that performance on the intermediate units was significantly better than on the segments, particularly under the onset + rhyme analysis (36%). The HS students performed almost perfectly on the syllables (96%) and were able to count both segments and intermediate units at about half of this level of accuracy.

Experiments 2 and 3 both involved variants of a highly flexible new experimental technique that we call the substitution-by-analogy task. This technique involves the aural presentation of two pairs of monosyllables (real or possible words), both of which illustrate a common modification pattern; the subject is then required to modify a test item "on analogy" with the examples. For instance, the pairs *beam-stream* and *cling-string* both illustrate a common change in the onset to /str-/ (from /b-/ and /kl-/, respectively), keeping the rest of the syllable (the rhyme) intact. In Experiment 2 a variant was used in which the manipulated unit was replaced by null (i.e., deleted), and the only three types of units involved were onsets, codas, and incomplete units (in which individual segments were deleted out of either an onset or a coda, as in *sprang-rang*). Subjects in this experiment were the same as in Experiment 1. The results were that the children performed significantly better in deleting onsets (42%) than either codas (12%) or incomplete units (7%), while the high schoolers deleted onsets and codas with about the same high rate of success (89% vs. 94%), significantly better than when either type of unit was broken up (63%). These findings attest to the integrity of both onsets and codas as phonological units for mature subjects, but with the former appearing earlier in development than the latter.

In Experiment 3 a variety of non-null patterns were employed for all six of the hypothesized intermediate units discussed above, with controls introduced for the number of segments affected, for spelling matches vs. mismatches, and for real vs. nonsense words. The results were scored in terms of both proportion correct and response latencies (for correct responses only), and both analyses revealed an almost identical response pattern, with rhymes and onsets the easiest/fastest to manipulate, nuclei and codas next, and heads and margins by far the most difficult/slowest. An analysis of error types also showed a strong tendency for errors to move in the direction of changes in the onsets or rhymes, with very few errors of the other types.

CONCLUSIONS

Taken together, these results not only serve to confirm the earlier indications that a hierarchical model is preferable to the linear one (which implies that all single-segment changes ought to be of equal difficulty, demonstrably not the case), but, by revealing the prominence of the rhyme constituent, they also indicate that the version of the hierarchical model shown in Fig. 1c is preferable to either of those shown in Figs. 1b or 1d. A fourth experiment is now in progress, utilizing a phonetic similarity judgment technique, that we hope will help in the proper interpretation of the consistently poor performance on the heads and margins: are they not units at all, or are the "bonds" merely weak? It is also important to establish whether or not the superior performance on the onsets and rhymes will extend to languages of diverse canonical types and with orthographic and poetic traditions that are different from English.

REFERENCES

[1] L. Zhurova. Razvitiye zvukovogo analiza slov u detey doshkol'nogo vozrasta. *Voprosi Psikhologii* 9.20-32, 1963.

[2] I. Liberman, D. Schankweiler, F. Fisher, B. Carter. Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology* 18.201-212, 1974.

[3] B. Fox, D. Routh. Analyzing spoken language into words, syllables, and phonemes: A developmental study. *Journal of Psycholinguistic Research* 4.331-342, 1975.

[4] L. Ehri, L. Wilce. The influence of orthography on readers' conception of the phonemic structure of words. *Applied Psycholinguistics* 1.371-385, 1980.

[5] J. Jaeger, Testing the psychological reality of phonemes. *Language and Speech* 23.233-253, 1980.
[6] P. Vitz, B. Winkler. Predicting the judged 'similarity of sound' of English words. *Journal of Verbal Learning and Verbal Behavior* 12.373-388, 1973.
[7] B. Derwing, T. Nearey. Experimental phonology at the University of Alberta. In J. Ohala & J. Jaeger (eds.), *Experimental phonology*. Academic Press, 1986.
[8] B. Derwing, T. Nearey, M. Dow. On the phoneme as the unit of the 'second articulation'. *Phonology Yearbook* 3.45-69, 1986.
[9] M. Dow. *On the role of orthography in experimental phonology*. M.Sc. thesis, University of Alberta, 1981.
[10] B. Derwing, M. Dow. Orthographic effects on lexical representations. In S. DeLancey & R. Tomlin (eds.), *Proceedings of the Second Annual Meeting of the Pacific Linguistics Conference*, University of Oregon, 1987.
[11] H. Van der Hulst, N. Smith. An overview of autosegmental and metrical phonology. In H. Van der Hulst & N. Smith (eds.), *The structure of phonological representations (Part I)*, Foris, 1982.
[12] T. Vennemann. The rule dependence of syllable structure. Unpublished ms..
[13] S. Shattuck-Hufnagel. Sublexical units and suprasegmental structure in speech production planning. In P. MacNeilage (ed.), *The production of speech.* Springer, 1983.
[14] J. Santa, C. Santa, E. Smith. Units of word recognition: Evidence for the use of multiple units. *Perception and Psychophysics* 22.585-591, 1977.
[15] R. Treiman. The structure of spoken syllables: Evidence from novel word games. *Cognition* 15.49-74, 1983.
[16] M. Dow, B. Derwing. Experimental evidence for syllable-internal structure. Paper presented at the UWM Linguistics Symposium on Linguistic Categorization, Milwaukee, WI, April, 1987.
[17] M. Dow. *On the psychological reality of sub-syllabic units*. Ph.D. dissertation, University of Alberta, 1987.

TOWARDS THE PHONOLOGICAL MODEL FOR CONTRASTIVE ANALYSIS

STEFAN GRZYBOWSKI

Russian Dept., Pedagogical University
Bydgoszcz, Poland 85-064

ABSTRACT

Two so far contradictory approaches to contrastive analysis of phonic phenomena, viz. structural (taxonomic) phonemics and generative (derivational) phonology, in fact can be regarded as complementary in the global phonological analysis used in the comparison of two different languages.

INTRODUCTION

Contrastive analysis, sometimes called, perhaps more appropriately, confrontative, in my opinion, does not consist in mere juxtaposition of two languages. Its aim should be defined rather as looking for equivalence between linguistic phenomena of languages under comparison. With that in mind I will consider the two phonological models mentioned above with respect to their usefulness for such an analysis.

Generative phonologists who refuse to regard phonemic level as relevant in the explanation of derivational processes (rules) transforming abstract underlying (phonological) representation directly into surface phonetic representation, insist that all differences between compared languages can be accounted for by phonological rules /1/. They seldom suggest a comparison of the phonetic systematic level /2/, which in such a case, however, does not have a definite theoretic status, rather it is viewed as a final result of derivation. On the other hand, different structural approaches assert in general that a comparison of the phonic shapes of two languages, especially for educational purposes,should be confined to relevant phonemic features extracted by any kind of distributional analysis /3/.Sometimes they insist on taking into account the phonetic reality of sounds as a necessary component in a contrastive phonological analysis /4/.

There are both theoretical and empirical arguments for the two approaches apart from contrastive linguistics. Let us, however, confine ourselves only to argumentation taken from the latter in evaluating below the usefulness of each of the models in contrastive phonology.

UNSTRESSED VOWELS

As regards the problem of unstressed vowels, Polish and Russian differ from the point of view of both models: the differences, however, are not of the same kind and value. Generative phonology seems to expose a more essential difference, since it sees the difference in both directions of equivalence between the languages; either from Russian to Polish or from Polish to Russian. This is because generative phonology claims that Polish lacks rules of so-called "unstressed vowel reduction" which is inseparable part of the Russian phonology.

From the point of view of many models of structural phonemics only the direction from Russian to Polish seems to be unsatisfactory. It is because,depending on interpretation, both languages can be considered as consisting of the same pattern of 5 or 6 vowels, which can be viewed as interchangeable in the course of using the opposite language provided that we do not intend to eliminate "foreign accent", but only to confine ourselves to minimal orthoepic correctness. In spite of that, in the direction at issue the difference ensues from distribution of the vowels, since in unstressed syllables Russian does not use [o] and sometimes [e] after non-palatalized sounds, and [a],[o],[e] after palatalized consonants, distinguishing only at least 3 vowels in the former position and 2 in the latter. This results in underdifferentiation when Russians use Polish, since in the latter all vowels are used in unstressed position as well.

The opposite direction of equivalence does not seem to cause any phonemic difficulties, since 5 or 6 Polish vowels can fit the same amount of Russian sounds in stressed position and from 2 to 3 in unstressed position. The only question here seems to be the choice of a proper vowel for a given semantic item. Thus, Polish /a/ should be a satisfactory substitute for Russian unstressed vowels represented by letter o in:

/1/ górod, gorodá, zamók, molokó etc.,

and it should be Polish /i/ which can replace Russian unstressed [i]-like sounds spelled by letter ya or a in:

/2/ tianú, vzialá, yazýk, chasý etc.

These and similar replacements form an evidence for the use of the structural approach in contrastive analysis, and that is the appropriate framework which is able to provide such a solution.

Thus, from the point of view of taxonomy Polish appears to have enough phonic means for Russian unstressed vowels, and consequently Poles should not have many difficulties in acquiring these Russian sounds. However, this is not the case, since one of the greatest difficulties of Poles learning Russian consists in"okanie" instead of "akanie" and "yakanie" instead of "ikanie". Those errors are usually blamed on Russian orthography, because unstressed vowels can be spelled as it has been shown in the examples /1/ and /2/. The influence of spelling is not to be neglected, yet there are examples proving that such errors are caused by other factors as well. For instance, the independence of phonic shapes of words from spelling and vice versa can be demonstrated both by very frequent Polish pronunciation of the Russian pronoun on as *[an] in stressed position and by refusing to write letter a for stressed /a/ in items like zarabátyvat'. Therefore we should look for another explanation, or,at least, for partial explanation of the cause of such errors. It is the absence of vowel alternation in Polish depending on stress and consonantal enviroment, which seems to cause such an inability of Poles to put proper vowel in unstressed and even in stressed syllables.

In the case mentioned above there appears to be more appropriate explanation based on generative phonology. It may be developed as follows. Since in Russian there are forms [ana], [anó], [an'í] containing the [a]-like vowel in the first syllable, and since Polish does not have the rule deriving an unstressed [a] from stressed [o] and thereby relating them to each other, then on the grounds of correct pronunciation of the forms mentioned, it is the [a]-like sound that is generalized and regarded as underlying one, and then it appears in stressed position as well, in spite of the proper spelling. In the case of improper "okanie" in unstressed syllables, the proper underlying vowel is introduced, yet it is not changed into an [a]-like sound because of the lack of a corresponding rule in Polish. Such is the case of *[vodá] instead of [vadá], where on the basis of stressed vowel and the spelling, [o] is regarded as underlying and as unchangeable. And such is also the cause of "yakanie" in the example /2/.

What follows is a conclusion that generative phonology should be included in contrastive analysis of Russian and Polish due to its capacity to explain real differences and thereby real equivalences with respect to unstressed vowels.

THE PROBLEM OF STRESS

The conclusion should be more obvious when we proceed to the problem of stress in Russian and Polish. It is common knowledge that both languages differ considerably in this respect. However, if we put aside morphological and lexical determination of it, then within the framework of taxonomy,two possible solutions are available concerning the differences, viz. Russian differs from Polish either in that its stress is no determined by any position of the word,while in Polish it is determined by the end of the word, or in that in Russian there are two different sets of vowels, stressed and unstressed /5/, with a relative freedom of appearance in the word, while Polish has the same set for both positions. Closer scrutiny at the solutions, indeed, leads to the conclusion af a less categorical nature. For the former we should admit that in Polish multisyllabic words, stress can be established on four last vowels e.g. PZU [pezetú] ,zabáwa, matemátyka, dálibyśmy etc. , and then the difference against Russian consists in three syllables, because in the latter the stress can select each of the seven last vowels. It does not offer very much for equivalence between Russian and Polish; it would be enough to say that Poles have to acquire three more syllables for stress when using Russian. If we accept the second solution we should go back to the question of what causes Polish nonpenultimate stress, and keeping to the same procedure we must establish two analogical sets of vowels as well. The only difference between the languages at issue would be confined to the fact that in Polish there would be no other distinction between the sets save the stress, while in Russian there should be different sets of vowels, i.e. 5 or 6 stressed-vowel pattern and 2-3 unstressed-vowel pattern.

Now we should recall that similar statement has been inferred in the framework of taxonomic model for the unstressed vowels. The problems of stress and unstressed vowels appear to be interdependent on that level of explanation, and were established independently of each other,which confirms the way of reasoning and forms a second justification for that level. That methodological justification together with some other observations concerning differences between Russian and Polish, e.g.that of distinctive function of the Russian stress against Polish, are satisfactory premises for accepting the structural model as a part of contrastive phonological analysis. However, as in the case of unstressed vowels it is not satisfactory in respect of the whole problem of stress for many reasons. Firstly, such a solution is not able to explain the changeablity of Russian stress in the course of inflexion and word-formation. Secondly, it cannot describe the stress as a suprasegmental phenomenon which can influence vowels. Finally, it does not provide an account of a crucial difference between Russian and Polish stress; it gives only an explanation which exposes merely different degrees in displaying the phenomenon of the same kind. There are, indeed, many indications to the opposite, i.e. that Russian and Polish stress are of different kind.

First of all, Polish stress does not have any connection with any particular morpheme, while Russian stress is attributed to many of them, e.g. cases of fixed stress or stressed affixes. There exist quite convincing arguments to treat Russian stress as morpheme stress /6/. And therefore any phonological model neglecting the morpheme and its phonic properties is not able to provide a satisfactory explanation of Russian stress and its difference from Polish.

Such an explanation is available within the framework of generative phonology, as it has been shown in the dissertation of H.S.Coats, Jr. /7/, who has inferred Russian word stress from accentual properties of underlying morphemes and has demonstrated that stress rules are the earliest in the course of derivation, placed just after word-formation rules. This is in accordance with the fact that Russian stress is of morphological origin, and this exposes the crucial difference from Polish stress, because the latter, as governed by the latest of phonological rules, is of different kind, being of phonetic origin /8/. On these grounds, the problem of interdependence, or rather dependence of Russian unstressed vowels upon the stress can attain better solution as well; stress rules are placed before any other phonological processes and therefore the stress can affect vowels. No such conclusion may be drawn from Polish, where the stressing is situated at the end of derivation, when all the vowels are established. Thereby stress and vowels are not interdependent. And that is why Russian vowels, for which Polish seems to have a sufficient number of surface sounds, cause such difficulties for Poles.

CONCLUSION

In the course of analyzing stress and vowels, it has been shown that both structural and generative models are useful and necessary in contrastive phonology, however, on a different level of explanation. Structural approach seems to provide explanations of direct equivalences of phonic phenomena, while generative phonology explains the way of forming phonic shapes of semantic items, as well as hierarchy and interdependencies of different linguistic phenomena, enabling thereby to get a better understanding of equivalences among languages under comparison. Thus, they can be regarded as complementary in the global phonological analysis needed in contrastive linguistics.

REFERENCES

/1/ E.Gussmann, "Contrastive Polish-English Consonantal Phonology", Warszawa: PWN 1978, p.148ff.

/2/ J.Fisiak, Generative Phonological Contrastive Studies, "Kwartalnik Neofilologiczny",33, 1976, p.120-121.

/3/ C.James, "Contrastive Analysis" Longman, 1980, p.80-85.

/4/ G.Hentschel, On the Relevance of Phonetic, Phonological, and Morphonological Levels in Contrastive Phonology, "Papers and Studies in Contrastive Linguistics", 20, 1985, p.32-33 .

/5/ Cf. M.Romportl, "Studies in Phonetics", Prague: Academia, 1973, p.39.

/6/ E.L.Ginzburg, Udarenie morfemy? In: "Fonetika. Fonologija. Grammatika", Moskwa 1971, p. 106-113.

/7/ H.S.Coats, Jr., "Word Stress Assignment in a Generative Grammar of Russian", Urbana-Champaign: University of Illinois, 1970.

/8/ L.Ossowski, O pričinach trudnostej, suščestvujuscich u poljakov pri obučenii russkomu udareniju, In: "Osnovnye doklady i soobščenija pol'skoj delegacii. III Meždunarodnyj kongress MAPRJaL", Warszawa:PWN 1976, p. 140-141.

Acknowledgements

I wish to express my gratitude to Dr. Boh Donnorummo for his help in acquiring fonds for my stay at the University of Pittsburgh, Pensylvania, which made it possible for me to study the problems discussed above. I also want to thank Dr. Aleksander Szwedek for helping me to check the English version of the paper.

PROBLEME DER PHONEMVARIANTEN IN DER GEGENWÄRTIGEN PHONOLOGIE

WEISALOW FACHRADDIN JADIGAR OGLY

Aserbaidshanisches Fremdspracheninstitut
Baku, 370055

ABSTRACT

This report contains the results of the investigation in the field of the phoneme variation based on the theoretical conception of L.V.Sherba and his followers. According to their presumption it's necessary to distinguish three levels of the phonological analyses: invariants, variants and sound levels. Each level is characterized by its own units and rules of their combination.

Fast ein Jahrhundert lang ist die Sprachforschung von der Frage nach Aufstellung des Sprachlautsystems und Herausstellung der Varianz und Invarianz der einzelnen Sprachlaute beherrscht/1/ und die Erfolge sind dabei so suggestiv, daß kein anderer Gesichtspunkt Anspruch auf solch ein wissenschaftliches Interesse erheben kann. Diese Fragestellung wurde besonders weitreichend inspiriert durch die grundlegenden Werke von I.A.Baudouin de Courtenay /2/ und Ferdinand de Sausseure/3/. Diese zwei Gelehrten stehen am Steuer der sich im letzten Jahrhundert stark entwickelnden Strömungen in der Sprachwissenschaft, auf deren Basis sich die Phonem- und Variantentheorie herausbildete.

Nach Ferdinand de Sausseure, der von der Dichotomie Sprache(langue)-Sprechen(parole) ausgeht, wird die Sprache als System von Zeichen bestimmt, in dem die Verbindung von Sinn und Lautzeichen wesentlich ist. Während die Sprache sich durch sozialen und konstanten Charakter kennzeichnet, ist das Sprechen immer individuell und euphemerisch. Sprache und Sprechen treten nach dem schweizerischen Gelehrten als zwei gegenüberstehende Polarstufen einer einheitlichen Erscheinung auf. Sie bilden zusammen die menschliche Rede/4/.

Auf der Dichotomie beruht eigentlich der Gedanke N.S.Trubetzkoys über die scharfe Trennung der Phonologie von der Phonetik, indem die Phoneme Einheiten der Sprache, die Varianten aber außer den Fällen,in denen sie eine delimitative Funktion ausüben, Einheiten des Sprechens sind.

Eine Hauptschwäche der sausseureanischen Dichotomie besteht darin, daß sie die Varianzebene aus der Sprachtheorie völlig ausschließt. In Wirklichkeit aber kommen in der zwischenmenschlichen Kommunikation solche Erscheinungen vor, die im Sprachsystem als solche nicht existieren und auch nicht als physikalische Dimensionen aufgefaßt werden können, aber deren Berücksichtigung für eine aufschlußreiche phonologische Theorie von großer Bedeutung ist. Es handelt sich hier um Eigenschaften, die als Ergebnis der segmentellen und suprasegmentellen Wechselwirkung an der Lautgrenze zutage treten und die die Varianz der sprachlichen Einheiten voraussetzen.

Die gegenwärtige deutsche Aussprachenorm fordert die starke Behauchung der stimmlosen Verschlußsprengelaute /p,t,k/ im Anlaut vor betonten Vokalen, sowie im absoluten Auslaut der Wörter und Sätze. Außerdem ist die Realisation der deutschen Vokale im Morphem-und Lexemanlaut mit einem starken Einsatz(Knacklaut) aussprachenormgerecht. Vgl.:

∠ p^{h}ákən, p^{h}ʏ́pə , t^{h}a:k^h , ẓánt^h , ʼá:bənt^h ,

ʼé:Rdə_/ usw.

Die Nichteinhaltung dieser Bedingungen wird als Abweichung von der gegenwärtigen Hochlautung wahrgenommen. Die Behauchung von /p,t,k/ im Anlaut vor betonten Vokalen und im Auslaut einerseits und die Aussprache der Vokale mit Knacklaut im Morphem- und Lexemanlaut andererseits haben im phonologischen System keinen funktionellen Wert. Sie sind phonetische Eigenschaften des Gesprochenen und obligatorisch für die Aufrechterhaltung der orthophonischen Norm. Diese Eigenschaften sind aber nicht im Sprachsystem integriert und daher treten sie im Deutschen nie als Differenzmerkmal auf. Es gibt im deutschen

Phonemsystem keine gegenüberstehenden Phonempaare, in denen sich das eine vom anderen durch das Vorhandensein oder Nichtvorhandensein von Behauchung unterscheiden würde. Im Gegensatz dazu ist die Behauchung im grichischen Dialekt Zakonisch, im Georgischen, Amcharischen, Tibetischen u.a. ein relevantes Merkmal/5/. Daraus ergibt sich, daß die materielle Beschaffenheit nicht frei ist und ihre Gebundenheit an Relevanz und Redundanz vollkommen von den Systemverhältnissen abhängt. Daher muß sie für jede Sprache isoliert betrachtet werden. Es gibt keine Sprache, deren Einheiten immer als solche fungieren,ohne daß sie auf bestimmte Variierungen hinweisen. Allerdings ist die Varianz von der Funktion her eine wichtige Ebene,denn die Systemverhältnisse und Einheiten der Sprache werden in der Varianzebene am deutlichsten ausgedrückt. Obwohl die Varianzeigenschaften als redundante Merkmale auftreten, sind sie doch wichtig für die Spracherkennung.

Im Gegensatz zu Ferdinand de Sausseure geht die sowjetische Sprachwissenschaft, insbesondere die Schule L.V.Schtscherbas, von der trichotomischen Sprachgliederung aus und betont die Wichtigkeit der dreiteiligen Gegenüberstellung für die Sprachtheorie und Sprachpraxis. Bei der trichotomischen Betrachtung lassen sich drei Aspekte aussondern:1.Das sprachliche System; 2.Das sprachliche Material als Gesamtheit von Gesprochenem und Wahrgenommenem(Text) unter Berücksichtigung der Begriffe von Zeit, Ort und Realisierungsbedingungen; 3. Die sprachliche Tätigkeit als Prozeß des Sprechens und Verstehens /6/. Diese Dreiteilung bezieht sich auf alle Ebenen der Sprachstruktur.

Die Varianten der Spracheinheiten können in dieser Gliederung zum sprachlichen Material gezählt werden, sie lassen sich gerade auf dieser Ebene am deutlichsten erkennen. Im sprachlichen Material(Text) greifen alle Faktoren der segmentellen und suprasegmentellen, sowie auch der außersprachlichen Einwirkungen auf die Phonemrealisation ineinander. Dementsprechend kann man in einer Sprachtheorie drei Ebenen auseinanderhalten, von denen die funktionelle Ebene oder Invarianzebene und die Sprechaktebene oder Lautebene zwei Polarstufen sind, zwischen denen die Normebene oder Varianzebene als Zwischenstufe existiert. Jede von diesen Ebenen kennzeichnet sich durch ihre Einheiten und Merkmale.

Im Prozeß des Sprechens und Verstehens treten die Laute(Phone oder Lautexemplare) als einzelne physikalische Erscheinungen auf, die sich aufgrund der Realisierungsbedingungen(innersprachlichen und außersprachlichen) in bestimmte Klassen gruppieren lassen und deren allophonische Abstrahierung sich auf der Norm-oder Varianzebene vollzieht. Sie werden durch verschiedene Schulen verschiedenartig bezeichnet: Allophone(nach den amerikanischen Deskriptivisten), Phonemschattierung oder Phonemvariante(nach Schtscherba und seinen Nachfolgern), Phonemvariante(nach N.S.Trubetzkoy u.a.) usw. Hier kann man auf eine tiefgreifende und ausführliche Analyse terminologischer und konzeptueller Grundlagen von den erwähnten und auch anderen Schulen verzichten. Wir gebrauchen Allophone und Variante als Synonyme.

Die Variante tritt als Besondere auf, weil jede ihrer Erscheinungen als Verallgemeinerung unter Berücksichtigung besonderer Bedingungen(Position, Kombination, Einfluß der suprasegmentellen Einheiten usw.) aufgefaßt werden kann.

Die Varianten als Besondere lassen sich auf der nächsten Abstraktionsebene in Phoneme verallgemeinern. Auf dieser Ebene bilden die Phoneme als Allgemeine ein funktionelles System, in dem ein Phonem dem anderen durch Oppositionsverhältnisse gegenübersteht. All diese drei Ebenen stehen einerseits miteinander in enger Beziehung und Wechselbeziehung, andererseits bewahrt jede von ihnen ihre autonome Selbständigkeit durch die im Sprachsystem integrierten Merkmale.

Die Laute sind solche akustischen Phänomene, die vom Sprechenden erzeugt werden,um dem Gesprächspartner etwas mitzuteilen.Der Gesprächspartner dekodiert das vom Sprechenden Übermittelte nach den akustischen Eigenschaften, die die Laute haben.

Von der Varianzebene her stehen alle möglichen Varianten nebeneinander und bilden ein Subsystem, in dem eine Variante im Vergleiche mit anderen nicht hervorgehoben werden kann. Nur von der Frequenz her kann man vom Vorzug dieser oder jener Variante sprechen. Das heißt, daß irgendeine Variante in einer höheren Gebrauchsfrequenz auftreten kann und somit der Häufigkeit des Auftretens nach statisch den Vorzug haben kann. Ansonsten sind alle Varianten für den Mechanismus der Sprache gleichwichtig und die Varianten eines jeden Phonems bilden gemeinsam die Gesamtheit der Regeln, nach denen die Phoneme der gegebenen Sprache zu gebrauchen sind.

Es gibt aber Varianten, die mehr oder weniger abhängig sind. Dazu gehören die isoliert ausgesprochenen Varianten. Dasselbe beobachten wir in den Fällen, wenn das Wort ein einziges Phonem hat. Vgl.:/O/-"er" im Aserbaidshanischen oder /ae/-"Ei" im Deutschen.

Was das Phonem anbetrifft, so besteht seine Funktion in der Konstruierung und Differenzierung von Wörtern und Wortformen. Ein Phonem ist so abstrakt wie ein Morphem oder ein Lexem und tritt im Text immer in seinen Varianten auf. Einem Phonem entspricht auf dem akustisch-artikulatorischen Niveau die Gesamtheit von akustischen Eindrücken und artikulatorischen Bewegungen. Ein Laut ist immer Repräsentant irgendeiner Variante und diese ist Vertreter irgendeines Phonems. Dieser Gedanke kann auch umgekehrt formuliert werden:ein Phonem ist durch Varianten, die letzteren aber sind durch Laute zu realisieren. Zusammenfassend kann man sagen, daß die Laute, Varianten und Phoneme sich zueinander so verhalten, wie sich das Einzelne zum Besonderen und das Besondere zum Allgemeinen verhält/7/.

Die hier dargelegte Variantenauffassung steht in Übereinstimmung mit der Lehre des dialektischen Materialismus über das Einzelne, Besondere und Allgemeine.

In der realen Wirklichkeit hat jedes Phonem mehr Varianten, als das von der Sprachbeschreibung zugegeben wird. Die Varianten eines Phonems bilden ein ununterbrochenes Skala und durch die Entwicklung neuer Untersuchungsmethoden werden immer mehr Typen von Varianten gefunden.

Die Varianten eines Phonems werden als solche nicht aufgrund der akustisch- artikulatorischen Verwandtschaft identifiziert, sondern durch eine linguistische Analyse, wenn es auch wichtig ist zu betonen, daß die Sachverhalte nur dann linguistisch unterschiedlich identifiziert werden, wenn akustisch-artikulatorische Unterschiede vorhanden sind. Akustisch- artikulatorisch verschiedene Laute können auch als Varianten eines Phonems zusammengefaßt werden, wenn diese Varianten gleiche Differenzmerkmale aufweisen und wenn die Integralmerkmale bei ihnen unterschiedlich sind. Laute mit absolut gleichen akustisch-artikulatorischen Besonderheiten aber können nie als zwei verschiedene Phoneme aufgefaßt werden. Die Integralmerkmale hängen völlig mit Position,Kombination und anderen suprasegmentellen Eigenschaften zusammen.Die enge Wechselbeziehung der akustisch- artikulatorischen Korrelate und der linguistischen Bewertung von lautlichen Eigenschaften ist in der von L.R.Zinder erarbeiteten These deutlich ausgedrückt. Das Wesen dieser These besteht darin, daß die phonetischen Unterschiede unbedingt eine Voraussetzung für die linguistischen oder phonologischen Unterschiede darstellen, aber nicht alle phonetischen Unterschiede unbedingt zu phonologischen oder linguistischen Unterschieden führen/8/. Dieses grundlegende Kriterium hat eine außerordentlich wichtige Bedeutung für die Aufdeckung der inneren Beziehungen zwischen Form und Substanz einerseits und zwischen Form und Funktion der lautlichen Erscheinungen andererseits.

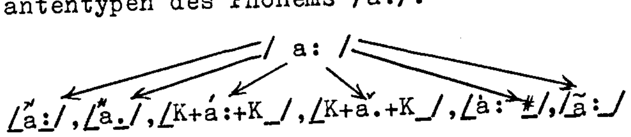
Ausgehend von der Theorie der Autonomie der lautlichen Seite der Sprache muß bei der phonologischen Analyse zuerst das Phoneminventar, danach das Phonemsystem mit möglichen Gegenüberstellungen nach Differenzmerkmalen festgestellt werden. Der nächste Schritt muß die Beantwortung der Frage sein, wie dieses oder jenes Phonem gebraucht wird, wobei durch die eingeführten Begriffe Position und Distribution Phonemvarianten näher präzisiert werden. Es ist wichtig zu betonen, daß diese Prozedur nicht durch mechanisch ausgedachte Minimalpaare und auch nicht nach der Distinktivität, die durch mechanische Gegenüberstellung von Quasihomonymen herausgefunden ist, durchgeführt wird, sondern sich auf das morphologische Kriterium stützt, dessen Wesen in der Aussonderung und Identifikation der Segmenteinheiten durch ihre Verbindung mit der minimalen bedeutungstragenden Einheit- dem Morphem (unmittelbar oder potentiell) besteht.

Einen besonderen Platz in der allgemeinen Phonemtheorie nimmt das Problem der Variantenbeschreibung ein. Die Untersuchungen der letzten Jahrzehnte weisen darauf hin, daß die sogenannten stationären Gebiete an und für sich nicht hinreichend für die Phonemidentifikation sind. Für die Phonemwahrnehmung braucht der Sprachträger außerdem die Information, die in Übergangsgebieten angegeben ist. Diese Stellungnahme führte zu tiefgreifenden Forschungen der Phonemvariierung.

Die Formulierung, nach der der Einfachheit halber unter dem Phonem im weiteren gerade die isoliert ausgesprochenen, sogenannten typischen Varianten verstanden werden(L.V.Schtscherba), rief später eine lebhafte Diskussion hervor. Man unterstellte L.V.Schtscherba die Vernachlässigung des dialektischen Zusammenhangs zwischen Phonem und Variante. Es erweist sich die Grundlosigkeit dieser Vorwürfe, da jede Realisierung des Phonems als das Phonem selbst bestimmt werden muß. Der Weg von der Variante zum Phonem und vom Phonem zu jeder seiner Varianten ist direkt, nicht aber über andere Varianten. Diese Fragestellung muß aber nicht zu Mißverständnissen führen, als sei Schtscherba in seiner Phonemtheorie von der Dialektik zurück getreten, indem er den Phonembegriff mit dem isoliert ausgesprochenen Laut gleichsetzt und ihn als typische Variante anführt.

Die Erarbeitung der Variantentheorie hat außerordentliche theoretische Bedeutung, vor allem für die Sprachperzeption, aber auch für die Sprachvermittlung. Beide Aspekte können sich erfolgreich entwickeln, wenn die Probleme der Phonemvarianten gelöst sind. Selbstverständlich haben die meisten Sprachen heute eine Beschreibung ihres Phonembestandes, manche Sprachen besitzen sogar eine ausführliche Beschreibung der einzelnen Phoneme, des Phonemsystems. Dessen ungeachtet sind die wichtigsten Probleme der allgemeinen, angewandten, konfrontativen und kontrastiven Phonetik heutzutage nicht lösbar, weil die Variantenbeschreibung stark zurückbleibt:

Unter vielen strittigen Fragen der Phonemvarianten mögen hier einige erwähnt werden. Es handelt sich vor allem um Variantendefinition, Variantenarten, Beziehungen zwischen Phonem, Variante und Laut. Eine umfangreichende distributive Analyse der Phonemrealisationen ermöglicht, das Subsystem der Varianten aufzustellen,ihre Grenzwerte und Arten festzustellen. Von den traditionell zu unterscheidenden obligatorischen,stilistischen und fakultativen Varianten sind die obligatorischen für die Sprachfunktion sehr wichtig, da die Phoneme der Sprache nicht isoliert fungieren, sondern sich zueinander in enger Beziehung befinden. Daher lassen sie sich den Einflüssen der Position und Kombination und auch der suprasegmentellen Einheiten unterwerfen. Aus der distributiven Analyse ergeben sich folgende Variantentypen des Phonems /a:/.



usw. Vgl.: Abend, aber, fragen, fatal, sah, mahnen u.a.
Die Zahl der Varianten kann man vermehren, wenn alle Faktoren der Phonemvariierung berücksichtigt werden.
Die Variante ist nicht begabt, Wörter und Wortformen zu differenzieren, sie steht einerseits den anderen Varianten des Phonems, zu dem sie selbst gehört,gegenüber, andererseits den Varianten anderer Phoneme durch das Phonem selbst. Vgl.:

/ a:/ - /a̋:_/ - /ǎ._/, /á:_/-/ǎ._/ u.a.
 #   #   #   #   #
/ e:/ - /e̋:_/ - /ě._/, /ẹ́:_/-/ě._/ u.a.

Die Varianten eines Phonems können miteinander durch die Einheiten höherer Stufe verbunden sein./e:_/ und /e._/ stehen miteinander durch ihre Realisation in Allolexemen /lé:bən/-/le.bɛ́n/(diç) in enger Wechselbeziehung. Bei der Lösung der Frage über die Zugehörigkeit der Variante zu diesem oder jenem Phonem spielt das Kriterium der komplementären Verteilung eine wichtige Rolle.
Die Klassifikation der Varianten und die Bestimmung des Variierungsdiapasons sind in der Wirklichkeit von der Analyse der positionellen und kombinatorischen Realisationsbedingungen abhängig. Im folgenden sind einige Regeln dieser Analyse zu erwähnen:
1. Gleiche Phonemumgebung, aber verschiedene phonetische Realisationsbedingungen, die durch die Wirkung der koartikulatorischen Einflüsse zu erklären sind. Vgl.: /u:/, /o:/, /a:/ in den Wörtern /tú:t/, /tó:t/, /tá:t/ u.a.
2. Gleiche Phonemumgebung und identische Realisationsbedingungen, aber verschiedene Situationen, Kontexte und Sprechakte. Vgl.: individuelle, situative und kontextuelle Varianten in folgenden Belegen.

Situation A: / zi.‿lí:st ae̯n‿b̋u:x //

Situation B: / zi.‿lí:st ae̯n‿b̋u:x //(niçt

ae̯nə‿tsáe̯tʊŋ) // u.a.

3. Gleiche Phonemumgebung, aber verschiedene phonetische Realisationsbedingungen, die durch suprasegmentelle Eigenschaften hervorgerufen sind. Vgl.: akzentuierte und nicht akzentuierte Aussprache der Vokale, ihre verschiedenen Stellen in der Struktur des Wortes: /la.bó:R/- /la.bo.ránt/,/dɔ́k-to.R/ - /dɔktó:Rən/ u.a.
Gleiche Phonemumgebung und daraus resultierende verschiedene phonetische Realisationsbedingungen bei gleichen suprasegmentellen Faktoren. Vgl.:/ ,a:,o:,u:/ in den Wörtern /hʊndə , bá:dən, ló:nən, tú:x/usw. Die angeführten Regeln verdeutlichen die Wechselbeziehung von Phonem und Variante. Es gibt natürlich noch andere Regelmäßigkeiten, die sich auf Phonemvariierung beziehen. Deren Formulierung stellen weitere Aufgaben der syntagmatischen Phonologie.

Literatur

/1/ E.Sievers, "Grundzüge der Phonetik", Leipzig, 1901.
/2/ I.A.Baudouin de Courtenay, "Izbrannyje trudy po obšemu jazykoznaniju";M.,1963.
/3/ F.de Sausseure, "Grundlagen der allgemeinen Sprachwissenschaft",Göttingen,1967.
/4/ N.S.Trubetzkoy,"Grundzüge der Phonologie", Prague, 1939.
/5/ Ibd, S. 46.
/6/ L.V.Schtscherba, "O trojakom aspekte jazykovyx javlenij i ob eksperimente v jazykoznanii", v kn.:Jazykovaja sistema i rečevaja dejatel'nost', Leningrad, 1974.
/7/ F.Vejsalov, "Die phonetische Wissenschaft in der UdSSR und einige Probleme der gegenwärtigen Phonologie", Wissenschaftliche Zeitschrift der Humboldt -Universität zu Berlin, Ges.-Sprachw. Reihe, 1978, Jg. XXVII H.3.
/8/ L.R.Zinder,'"Obšaja fonetika",M.,1979, S.73.

DIE FORSCHUNGEN DER PHONEMREALISATION UND DIE BEDEUTUNGEN DER WORTFORM

JÓZEF WIERZCHOWSKI

Uniwersytet Warszawski
Warszawa

Zusammenfassung

Mit der Entwicklung von Kriterien, die zur Aussonderung von Bedeutungen innerhalb der Wortformen dienen, und für den ganzen Lexikbestand einheitlich sind, entsteht die Möglichkeit die Realisationen der Phonemen samt einzelnen Bedeutungen zu untersuchen. Die Unterschiede in der Realisation sind besonders in schwachen Stellen der Sprachstruktur zu beobachten. Ansehnlich sind in der gepflegter Version der Umgangssprache. Z.B. poln. wskazówka in der Bedeutung 'Zeiger' oft skazufka, in der Bedeutung 'Weisung des Vorgesetzten' immer mit deutlichem f am Anfang.

In der Forschungen zur Klanggestalt der Sprache schenkte man im 20. Jh. die Aufmerksamkeit vor allem I. Phonemen als Bestandteilen des Sprachsystems, II. Phonemen unter dem Aspekt der phonemischen Umgebungen, sowie III. Phonemen als Bestandteilen jeweils bestimmter Wortformen bzw. bestimmter Morpheme (und auch anderen Abhängigkeiten von der Stelle, an der ein Phonem in der Wortform vorkommt). Beim Untersuchen der phonetischen Realisationen von Wortformen wurde man ebenfalls aufmerksam auf IV. stilistisch bedingte Unterschiede, die in prosodischen Werten, aber auch in unterschiedlichen phonemischen Werten zum Ausdruck kommen.

In den Bemerkungen zu Lautveränderungen bedient man sich gewöhnlich des Begriffes eines mehr oder weniger rigoros augefaßten Lautgesetzes oder des Begriffes der phonetischen Tendenzen. Man wurde darauf aufmerksam, daß manche phonetische Veränderungen deutlich mit einer bestimmten Wortform, oder mit einem bestimmten Morphem zusammenhängen, und nicht mit anderen, was an der Ausschließlichkeit der aus der phonetischen Umgebung sich ergebenden Bedingungen zweifeln läßt.

Sowohl die Untersuchung der Phonemrealisationen, als auch Beobachtungen zu diachronischen Lautveränderungen bezogen sich gewöhnlich auf die ganze sprachliche Einheit, oder auf einen Teil dieser Einheit, die samt aller ihrer Bedeutungen die Bezeichnung Wort trägt. In phonologischen Forschungen wurde die Bedeutung ausschließlich in Kategorien der Distinktivität berücksichtigt. Das phonologische Wort war gewöhnlich mehrdeutig. Es reichte bereits aus festzustellen, daß die Wortform A in Bezug auf die Wortform B hinsichtlich der Form und des Inhalts distinktiv war. Beide Wortformen konnten mehrdeutig sein. Der Stand der semantischen Forschungen mit dessen charakteristischer Willkür hinsichtlich der Grundlagen der Bedeutungsunterscheidung innerhalb der Wortformen verhinderte das Unternehmen jeglicher Versuche, die zur Aufdeckung solcher Unterschiede führen könnten, die mit Bedeutungsunterschieden (innerhalb der untersuchten mehrdeutigen Wortformen) verbunden sind. Für derartige Untersuchungen war auch die bekannte These von Hjelmslev nicht förderlich, daß es zwischen der Ausdrucksseite und der Inhaltsseite keinen Gleichgewicht gibt. Ebensowenig förderlich war die Entwicklung der distributionellen Methoden, die prinzipiellen Nachdruck der spezifisch aufgefaßten phonemischen Nachbarschaft verleihen.

Eine ganz andere Situation schaffen die tiefgreifenden Veränderungen in der Auffassung von der Struktur der Sprache. Es geht um die Betrachtung der Sprache als eine monoseme Struktur, die nicht aus mehrdeutigen Wortformen aufgebaut ist, sondern aus bedeutungseinheitlichen Spracheinheiten, d.h. aus bedeutungseinheitlichen sprachlichen Zeichen. Die entscheidende Rolle kommt dabei der Entwicklung von Kriterien zu, die zur Aussonderung von Bedeutungen innerhalb der Wortformen dienen, und die für den ganzen Lexikbestand einheitlich sein sollen.

Die Forderung, bei der Sprachbeschreibung mit dem Begriff des bedeutungseinheitlichen sprachlichen Zeichens zu operieren, ist nicht neu. Aber es war der Mangel an einheitlichen Kriterien zur Aussonderung so aufgefaßter Zeichen, der sich auch auf dià phonetische Erforschung der Zusammenhänge zwischen den Unte rschieden der Phonemrealisation und den einzelnen Bedeutungen der Wortformen hemmend auswirkte.

Durch die Gewinnung solcher Prinzipien und deren deutliche Bestimmung kommt es zur Abschaffung der Willkür in der Bedeutungsaussonderung, und somit entsteht die Möglichkeit, die Untersuchungen der phonischen Gestalt der Sprachzeichen mit der Inhaltsseite dieser Zeichen zu verbinden. Es entsteht die Möglichkeit der Verbindung von Bedeutungsunterschieden innerhalb der Wortformen mit den Realisationsunterschieden der Phoneme als Bestandteile einzelner homophoner Zeichen. Es wird dann ebenfalls möglich, die Forschungsergebnisse zu vergleichen.

Die für den ganzen Bestand der Lexik einheitlichen Kriterien zur Aussonderung von Bedeutungen innerhalb der Wortformen, Kriterien zur Gewinnung von bedeutungseinheitlichen, homophonen sprachlichen Zeichen ergeben sich aus der Erforschung von Relationen, die zwischen den Zeichen den Zeichen der Sprache bestehen. Wesentlich ist dabei der Begriff der Benennungsrelation. Die Benennungsrelation besteht zwischen den in ihrem Stammteil ähnlichen Zeichen, dia auch in den Paraphrasen dieser Zeichen auftreten können. Diese Paraphrasen tragen den Namen der Verbindungsphrasen. Die Gesamtheit der in Benennungsrelatin zueinander stehenden Zeichen bildet die Gruppe der Benennungsrelationen.

Homophone sprachliche Zeichen werden durch Unterschiede der Benennungsrelationen voneinander unterschieden, vor allem durch die Unterschiede in der Anzahl und im phonemischen Bau der Zeichen, zu denen sie in Benennungsrelation stehen. Die Benennungsrelation ist somit dasjenige, was die sog. abgeleitete Formen miteinander verbindet. Von der für die Ableitungsanalyse typischen binären Betrachtung und von dem Begriff der Ableitung wird hier jedoch abgesehen. Nachdruck wird hier auf die Berücksichtigung aller Zeichen gelegt, die zur Gruppe der Benennungsrelationen gehören, sowie darauf, daß sie in der Sprache koexistieren, und nicht abgeleitet sind.

Die Benennungsrelationsgruppe enthält alle Formen, die in Benennungsrelation zueinander stehen und gesonderte, sich sich voneinander unterscheidende Zeichen darstellen, ungeachtet dessen, ob sie zu verschiedenen Redeteilen gehören oder zu demselben Redeteil gehören und sich nur durch ihre Flexionsformen voneinander unterscheiden. In der Gruppe der Benennungsrelation befinden sich die Zeichen ohne Rücksicht auf deren grammatischen Wert. Dieser Typ der Analyse kann als prägrammatische Sprachanalyse bezeichnet werden. Die Aussonderung der monosemen Zeichen ist durch eine Analyse bedingt, die sich nicht auf grammatischen Relationen gründet, sondern auf der Benennugsrelation, also af einer prägrammatischen Relation zwischen den sprachlichen Zeichen, deren sprachliche Exponenten die Verbindungsphrasen sind.

Bei den Bedeutungsunterschieden der so ausgesonderten Zeichen handelt es sich um solche, die sich sprachlich gefestigt und in der Sprachstruktur eingeprägt haben. Monoseme Zeichen werden aus der innersprachlichen Relationen gewonnen, und nicht aus der außersprachlichen Welt. Deshalb sind für die Struktur der Sprache vor allem diejenigen Bestandteile wesentlich, die sich in dieser Struktur eingeprägt haben, und nicht solche, die sich direkt aus der außersprachlichen, mit sprachlichen Formen mitteilbaren Welt ergeben. Die außersprachliche Welt widerspiegelt sich in der Sprache durch das Prisma bereits existierender sprachlicher Formen. Die Bedeutungseinheitlichkeit der sprachlichen Zeichen ergibt sich daraus, daß die Sprachstruk tur autonom ist, und nicht aaus einfachen und leicht wahrnehmbaren Unterschieden in der außersprachlichen Welt. Einfache Unterschiede, die zwischen verschiedenen Dingen bestehen, können keine gute Grundlage zur Aussonderung von sprachlichen Bedeutungen sein. Auch die verschiedenartig aus der Meinung der Muttersprachler gewonnenen und somit durch die außersprachliche Welt bedingten Distinktionen scheinen nicht genügend begründet zu sein.

Wesentlich ist die Untersuchung von Phonemrealisationen in Zusammenhang mit den Bedeutungen, die sich aus den innersprachlichen Relationen zwischen den Zeichen ergeben. Die Möglichkeit, dieselben Kriterien der Bedeutungsunterscheidung auf den ganzen Bestand der Lexik anzuwenden, bietet auch die volle Möglichkeit, die Ergebnisse zu vergleichen und statistische Methoden zu verwenden. So wie klar bestimmte Prinzipien der Aussonderung der Phoneme als Bestandteile des Phoneminventars es ermöglichen, vergleichbare phonologische Untersuchungen durchzuführen, so ermöglichen es die bedeutungseinheitlichen und nach einer, klar bestimmten Prozedur gewonnenen Zeichen, vergleichbare subphonologische Untersuchungen durchzuführen. Gemeint sind hier Untersuchungen der Phonemrealisationen in Zusammenhang mit den einzelnen Bedeutungen der phonologischen Worte.

Die subphonologischen Untersuchungen der Phonemrealisationen sind eingehender als jene, die sich mit Phonemrealisationen als Realisationen von Bestandteilen einzelner phonologischer Worte befassen. Diese subphonologischen Untersuchungen weisen darauf hin, daß Realistionsunterschiede, die mit Bedeutungsunterschieden, mit den einzelnen Bedeutungen der untersuchten Formen verbunden sein können, sehr schwierig von solchen Realisationsunterschieden zu trennen sind, die deutlich mit dem stilistischen Wert der realisierten Formen zusammenhängen. Diese Tatsache ist zu verzeichnen, obwohl der stilistische Wert auch als ein semantischer Wert angesehen werden kann. Für die Beobachtung von Realistionsunterschieden, die mit einzelnen Bedeutungen der Wortformen verbunden werden können, eignen sich besser solche Formen, in denen man leicht Bedeutungen aussondern kann, die sich durch ihren stilistischen Wert im wesentlichen nicht unterscheiden. Die Untersuchungsergebnisse der Realisationen des Phonems r sind auch etwas anders für die polnische form nerwy 'Nerven', die man oft in der Rede deutlich unterstreicht, als die Ergebnisse der Realisationen des r als Bestandteil des Phoneminventars. Realistionsunterschiede können bekanntlich auch milieubedingt sein. Selbst wenn von diesen und allen anderen Bedingtheiten abgesehen wird, erhält man oft unterschiedliche Untersuchungsergebnisse von Phonemrealisationen in den einzelnen Bedeutungen der untersuchten Wortformen. Es ist dann gut, nach eventuellen Bestätigungen dieser Unterschiede zu suchen, indem zunächst die Aussprache einzelner Personen untersucht wird, um erst dann die Aussprache, die Phonemrealisationen in verschiedenen Sprachmilieus zu untersuchen, und schließlich zu Generalisierungen zu übergehen, die sich auf die ganze Sprache beziehen.

Die Untersuchung von Realisaationsunterschieden der Phoneme als Bestandteile bedeutungseinheitlicher Zeichen mit gleicher phonemischer Form, also als Bestandteile homophoner Zeichen weisen darauf hin, daß sich diese Unterschiede vor allen Dingen an den sogenannten schwachen Stellen des Systems beobachten lassen. Es kann nicht behauptet werden, daß diese Unterschiede überall dort zu finden seien, ihr Vorkommen ist nur eine Möglichkeit. Jedoch ist dies eine Tatsache von weittragender Bedeutung sowohl für die Beschreibung der Sprachstruktur als auch für da s Verständnis historischer Wandlungen.

Zu ganz anderen Ergebnissen führt die Realisationsuntersuchung des Phonems ǫ in der polnischen Wortform wąż und in den beiden Einzelbedeutungen dieser Form. Der Nasal in vǫŝ in der Bedeutung 'Schlange' klingt voller, obwohl er immer biphonematisch und asynchron bleibt. In der Bedeutung 'Schlauch' unterliegt wesentlich häufiger der Zerlegung in Richtung ou und wird schwächer nasaliert. Man beobachtet das im Sprachgebrauch von Einzelpersonen und auch in größeren Sprachmilieus mit annährend gleichen phonetischen Merkmalen. Es scheint, daß dieselben Resultate zu gewinnen sind, wenn eine entsprechend große zahl von Personen untersucht wird, die die gepflegte Version der Umgangssprache sprechen.

Genaue Realistionsuntersuchungen des Anlauts ler polnischen Wortform wskazówka führen zu ganz anderen Ergebnissen, wenn man den Anlaut dieser Form getrennt in der Bedeutung 'Zeiger' und in der Bedeutung 'Weisung des Vorgesetzten' untersucht. In der erstgegenannten Bedeutung kann der Anlaut in umgangssprachlichen, wenig gepflegten Realisationen sogar auf s (skazufka) reduziert werden, oder er kann ein schwach ausgesprochenes bilabiales f enthalten. In der zweiten Bedeutung ist eine solche Realisation kaum möglich; das f wird fast immer ziemlich deutlich artikuliert.

Unvergleichlich häufiger kommt ein e statt eines y in der polnischen Wortform dyrekcją vor, wenn sie auf das Gebäude bezogen wird. Dieses e taucht nur ausnahmsweise auf, wenn es sich um ... unter der Leitung'... handelt.

Die Form porządnym je nach Bedeutung wird realisiert entweder pozonnym oder pozodnym. Ansehnlich ist die Zahl sprachlich fixierter Paare mit derselben Abstammung, wie z. B. na czole : na czele, dzialo : dzielo usw. Diese Fakten verpflichten dazu, die Lehrsätze der historischen Grammatik zu ergänzen, und zu erwägen, ob es wirklich die mehrdeutigen Wortformen sind, die historischen Wandlungsprozessen unterliegen, oder sind es vielleicht eher die bedeutungseinheitlichen Sprachzeichen.

THE METALANGUAGE OF PHONOLOGY

A. Yevdoshenko

Institute of Language and Literature
MSSR Academy of Sciences

ABSTRACT

The metalanguage of phonology unites a strictly determined number of notions, forming an hierarchy - phoneme, feature, category, system, connected by law-governed relations and necessary for a non-contradictory, comprehensive and economic, both synchronic and diachronic, as well as typologic description of languages.

Metalanguage is a notion of mathematical logics, opposed to a language as an object. We interpret these notions in the spirit of J.Baudouin de Courtenay, who anticipated them to a certain degree distinguishing between "linguistic categories", i.e. the metalanguage, and "language-category", i.e. a language-object. The first ones are "pure abstractions", wile the second ones are something "living in a language" /3/.

Phoneme. The first notion of phonology metalanguage, the phoneme, was introduced by J.Baudouin de Courtenay, who defined it in his last works as "the psychical equivalent of a sound" /3/. This phoneme definition can be evaluated in the light of materialistic dialectics, taking into account that a speech sound is something material, moving air particles, while a phoneme is something ideal, i.e. a peculiar reflection of the material in the consciousness of a language speaker. As K.Marx wrote, the ideal is "nothing else than the material, transplanted into a human head and transformed there" /I/. Elaborating his phoneme theory, N.S. Trubetzkoy underlined, that phonemes realised themselves in speech sounds, that phonemes and sounds "lie in different planes", thet is why a phoneme could not be considered as a family or a group of sounds, that we could not "begin with the sound" in defining a phoneme /4/. However, this true principles did not prevent him from criticizing the above-mentioned definition of the phoneme by J.Baudouin de Courtenay. Baudouin's phoneme is sensually concrete, it is the starting point in phonological cognition. The process of ascent from concrete to abstract cognition is ascending from a Baudouin's phoneme to a feature, from a feature to a category of features, from the latter to a system. A further ascent from abstract to mentally concrete, the deepest and the most content-bound cognition, consists in the definition of the phoneme as a totality of abstractions /5/, /6/.

A feature is a class (split into property and set) of phonemes. N.S. Trubetzkoy considered the feature either as a class of phonemes, when he determined the series as set of phonemes, characterised by the same feature (in logic it is a determination of set via property), or as an element of the phoneme, when he said, that "a phoneme is easily split into its phonological features" /4/, /5/. In the last case he hypostazed the feature, i.e. he transformed it from a class into an element.

Category. N.S.Trubetzkoy has introduced the notion of "coordinate" (category of features). He was right as he noted, that "every feature certainly belongs to a coordinate", but he also noted, that "a feature enters a phoneme composition" (!) /4/. Neither his followers nor like-minded persons paed any attention to his coordinate. L.Hjelmslev and later E. Benveniste reduced all phonological system to two notions, the first one - to taxeme and glosseme /7/, /8/, the second one - to phoneme and merisme /9/. A complex integral system, i.e. the phonological one was turned by them into a simple, summative system, based not on subordination but on coordination /5/, /10/, /II/. Meanwhile a category of features (a coordinate for N.S.Trubetzkoy) is a cardinal notion of phonology, that can be defined as a class of non-crossing features.

System. The most general notion of the phonology metalanguage is the system, that is a class of feature categories. It is in the system that every feature of one category intersects with every feature of its other categories. It is further, the feature intersection points that form the phonemes. Every phoneme is determined by the same feature number (one belonging to every category of a given system). According to N.S. Trubetzkoy, phonemes of the same system are determined by an unequal number of features. Sometimes he gave a phoneme "pure negative" definitions, that contradict elementary rules of logic.

Potentiality. A phonological system has a potential character /5/, /12/. The number of its possible phonemes is equal to the product of the numbers of different category features /5/. Thus, in the system of vowels there are 12 (i.e. 3 x 2 x 2 = 12) potential phonemes. The system is opposed to the inventory of actual phonemes, representing the realization of a certain number of potential phonemes. Thus, of the 12 potential vowels 3 are realized in Semitic languages, 5- in Russian, 6 - Bulgarian, 7 - in Italian, 9 - in Moldavian etc. /5/.

Inventory. In order to reveal the interrelation between actual (realized) and potential phonemes, it is important to distinguish the plane of expression from that of contents, as identified by L.Hjelmslev /7/.

A phonological structure does not belong completely to the plane of the language expression, as he supposed. It is only the inventory of actual phonemes which belongs to the plane of expression, while the potential system belongs to that of content.

Since only the inventory of realized (actual) phonemes is a part of the plane of expression, the problem of language levels becomes meaningless.

Opposition. As far as actual phonemes are concerned, it is necessary to note, that their opposition can not be the criterion of feature relvancy, as N.S.Trubetzkoy supposed. He exaggerated the importance of opposition, as he said, that "in phonology the main role is played not by phonemes, but by sense distinguishing oppositions" /4/. V.I.Lenin considered exaggerations like that of regarding one of the sides or verges of cognition as something absolute /2/, as idealism.

Sense. Besides that, neither features nor phonemes have direct relations to sense differentiation. Sense is expressed by words, not by phonemes, features or their oppositions.

System laws. The relevance of a feature, as well as that of a phoneme is determined by system laws, by data of language history, its dialectology and typology. E.g., for labial consonants it is not labiality that appears relevant. The nasality of French vowels belongs to the category of opening features /24/ etc. Let us exemplify it. Thus, nasality of French vowels proves to be a feature of the category of opening degree with the following features of its own: close, mid open, open and utmost (maximum) open, i.e. nasal, because according to the system laws /5/ nasality as a feature does not intersect with other features of its category, but it does intersect with all the features of other categories of a given system.

The interpretation of nasality as of a feature of the category of the vowel opening degree is explained by phonetical switching of nasal cavity as an additional resonator pronouncing open vowels.

The development of a feature as a phoneme class is observed also in many other cases in the process of investigation of phonological systems in historical (diachronical) plane.

Thus, all the class of actual backlingual consonants of vulgar Latin was transformed under certain conditions in the class of "labial" consonants in Moldavian (k>p; g>b; ŋ>m). And vice versa: all the class of labial consonants of Moldavian before i, j is transformed, in one of its dialects, in the class of soft mediopalatal consonants and then in one of soft prepalatal consonants, in its other dialect (p>k'>t'; b>g'>d'; m>n>ɲ'; f>š>s'; v>ž>z').

The movement of whole classes (features) of vowels allows us to admit, that what seems to be labial is not labial, but exactly its opposite, being more back than backlingual. Such a feature is apparently glottal, if we agree with M.Grammont's point of view, according to which labial consonants are characterized by a double occlusion, both glottal and labial /15/.

The glottal feature of the consonants explains why all the class of front vowels i,e,ɛ, preserved as such in Aromanian (Macedo - Romanian), shifted after labial consonants to the class of back vowels p, b, m, f, v in Romanian.

Consequently it is reasonable to admit that the formula of local features of the consonants p-t-k should be replaced by t-k-p.

Feature incompatibility. The class of soft mediopalatal consonants is incompatible with that of soft prepalatal consonants in the same language or dialect /5/.

Phonological feature vs. anthropophonic (phonetical) feature. An anthropophonic feature (e.g. a voiced or voiceless consonant) can be directly observed by a phonetist. Thus Russian word пруд is pronounced as прут . Sonority as a phonological feature of the phoneme "д" is the essence of the sound /д/, i.e. the abstract feature of the phoneme. Essence, as Hegel said "is preserved in its perfect purity" irrespective of the position, in which one or another allophone may turn out. The sound as such may also disappear completely without affecting the phoneme it represents (as for example in the Russian солнце where "л" is not pronounced).

Neutralization. It is not the phonological features but the anthropophonic properties, which are neutralized /14/. Y.K. Lekomtsev is right, in considering neutralization as "substantional noise", included in the circle of theoretical notions /16/, /17/.

Marking. According to phonology, marking opposition members (as admitted by N.S. Trubetzkoy) /4/ is not possible because it contradicts the logical equivalence both of phonemes and of phonological features (see System).

Constitutive and non-constitutive changes. We consider as non-constitutive the transformation of one inventory phoneme into another, e.g. l r, as in Latin dolere, Romanian durere. It was using only non-constitutive changes that W.Wartburg divided Romania into two language zones /18/

The transformation of potential phonemes , and others into actual (inventory) ones reflects minimum constitutive changes, i.e. those affecting only the inventory, but not the system.

The constitutive changes concerning a system can be devided into minimum and maximum ones. The former are nothing more than the appearance or disappearance of a given feature (e.g. in the development of French consonant system a class of affricates appeared, which later disappeared), an exemple of the latter is the disappearance of vowel length, a fact which we consider as the beginning of vulgar Latin between the 2-nd and the 3-rd centuries A.D.) /19/.

It is in this period that we see the division of Romania into three dialect zones with their specific vowel inventories /20/.

Taking into account the change hierarchy is essential for the determination of the language relation degree /21/, as well as for the language history periodization /22/ and for the dividing of a language into dialect units of several degrees: dialect, subdialect etc. /23/.

Phonology and Geometry (stereometry)

As N.S.Trubetzkoy has noted, "the order, achieved by phoneme division into parallel rows reflects the phonological reality" /4/. Features as phoneme classes (see Feature) parallel to each other within the same category and perpendicular to parallel features of other categories of the same system (see Category and System) form geometrical figures, that allow us to regard the theory of phoneme as a science not inferior to geometry in its exactitude /24/, /25/, /27/, /28/.

References

I. Маркс К. Капитал, т. I. - К.Маркс и Ф.Энгельс. Сочинения, т. 23.

2. Ленин В.И. Философские тетради. М., I973.

3. Бодуэн де Куртенэ И.А. Избранные труды по общему языкознанию. М., т. I, I963.

4. Трубецкой Н.С. Основы фонологии. М., I960.

5. Евдошенко А.П. Проблема структуры языка. Кишинев, I967.

6. Категории материалистической диалектики. М., I957.

7. Ельмслев Л. Пролегомены к теории языка. - Новое в лингвистике, вып. I. М., I960.

8. Yevdošenko (Evdoshenko) A.P. Glossématique? - Proceedings of the Eleventh International Congress of linguists. Bologna-Florence, 1972, II.

9. Бенвенист Э. Уровни лингвистического анализа. - Новое в лингвистике, вып. IУ. М., I965.

I0. Афанасьев В.Г. О принципах классификации целостных систем. - "Вопросы философии", I963, № 5.

II. Мамардашвили М.К. Процессы анализа и синтеза. - "Вопросы философии", I958, № 2.

I2. Евдошенко А.П. Стереометрическое моделирование и изоморфизм. - Студий де лимбэ молдовеняскэ. Кишинэу, I963.

I3. Евдошенко А.П. К обоснованию изоморфизма фонологии и морфологии. - Acta linguistica Academiae Scientiarum Hungaricae. Tomus 18 (I-2). Budapest, I968.

I4. Евдошенко А.П. Сопоставительная фонология и морфология молдавского и русского языков. Кишинев, I977.

I5. Grammont M. Traité de phonétique. Paris, 1939.

I6. Лекомцев Ю.К. О потребности в формальном метаязыке для фонологии. - Конференция по структурной лингвистике, посвященная базисным проблемам фонологии (тезисы доклада). М., I963.

I7. Евдошенко А.П. Сравнительная типология. - Сравнительно-типологическое изучение русского и молдавского языков. Кишинев, I986.

I8. Wartburg W. La fragmentation linguistique de la Romania. Paris, 1967.

I9. Евдошенко А.П. Ынтродучере ын студиул лимбилор романиче. Кишинэу, I987.

20. Evdochenko A.P. A propos de la fragmentation dialectale du latin vulgaire ... - Congresso internationale de linguistica e filologia romanza (Riassunti delle comunicazioni). Napoli, I974.

2I. Евдошенко А.П. Структурные критерии определения степени родства языков. - Происхождение аборигенов Сибири. Томск, I969.

22. Евдошенко А.П. О фонологическом критерии периодизации истории языка (на материале романских языков). - Лингвогеография, диалектология и история языка. Кишинев, I973.

23. Евдошенко А.П. Консидераций асупра системелор фоноложиче але граюрилор молдовенешть. - Диалектоложия молдовеняскэ. Кишинэу, I976.

24. Евдошенко А.П. Системул диференциал ал сунетелор балканоромaниче. - "Лимба ши литература молдовеняскэ", I960, № I.

25. Евдошенко А.П. К вопросу о применении стереометрической модели в области фо-

нологии. Сб. "Исследования по структурной типологии". Москва, I963.

26. Евдошенко А.П. Некоторые принципы диахронической фонологии. - Сб. **Omagiu lui Alexandru Rosetti.** Бухарест, I965.

27. Евдошенко А.П. Елементе де фоноложие диакроникэ. "Курс де граматикэ историкэ а лимбий молдовенешть". Кишинэу, I965.

28. Евдошенко А.П. Типологические основы русской грамматики для нерусских. - "Вопросы языкознания". I987, № 2.

TONOGENESIS IN NORTHERN MON-KHMER

Jan-Olof Svantesson

Lund University

ABSTRACT

Tonogenesis in the Northern Mon-Khmer languages Kammu, Hu and U is described. Each of these languages has acquired tones in its own way, and mechanisms other than those generally used to explain the origin of tones are involved. The fact that these languages have undergone different types of tonogenesis, while other closely related languages have not acquired tones shows that presence or absence of tones cannot be taken as an indicator of genetic relationship between languages.

INTRODUCTION

In this paper, tonogenesis in the three languages (Northern) Kammu, Hu and U will be described. These languages are spoken in northern Laos, Thailand and Burma, and in Southwest China. This area is dominated by tone languages belonging to the Tai and Sino-Tibetan language families, and there is a strong tendency for languages in this area to acquire tones.

The places of these languages within the Kammuic and Palaungic branches of Mon-Khmer are as follows (**boldface** indicates tone languages):

- Kammuic:
  - Kammu
    - **Northern Kammu**
    - Southern Kammu
  - Mlabri
  - Mal
  - .....

- Palaungic:
  - Lamet
  - Waic
    - Parauk
    - **Blang**
  - .....
  - Angkuic
    - **Hu**
    - **U**
    - Mok
  - .....
  - **Danaw**
  - **Riang**
  - Rumai

From this table it is obvious that tones have developed independently in several of these languages.

KAMMU

This language has two tones, which are rather level, and can be described as high (´) and low (`), although the difference between them is rather small. (See Gårding and Lindell 1977 and Svantesson 1983 for Kammu tones.)

Fundamental frequency contours of the two tones are shown in Figure 1.

Tonogenesis is simple: voiceless and voiced initial consonants have merged, and given rise to high and low tone, respectively. This type of tonogenesis is expected and phonetically motivated, since numerous investigations have shown that voiceless consonants increase and voiced consonants decrease F_0 in the following vowel. Nevertheless, this type of tonogenesis is not encountered very often in actual languages (see Hombert 1978:78).

The Kammu tone system is an innovation which has started in a central area, in northern Laos. Dialects to the south of this area have not developed tones, so Kammu is an example of a language in the process of acquiring tones. Examples of Kammu tonogenesis:

Kammu	S Kammu	
hntá?	hnta?	"tail"
hntà?	hnda?	"thin"
ráaŋ	r̥aaŋ	"tooth"
ràaŋ	raaŋ	"flower"

HU

This language (Svantesson forthc.a.) belongs to the Angkuic branch of Palaungic. Like Kammu it has a two-tone system, with high (´) and low (`) tones, illustrated in Figure 2.

In the Angkuic languages, including Hu and U, initial voiceless and voiced stops have not merged, but have been retained as aspirated and unaspirated voiceless stops, respectively. The tones do not depend on voicing in initial consonants. Instead, Hu has combined two areal trends, loss of vowel length and acquisition of tones, so that words with an originally

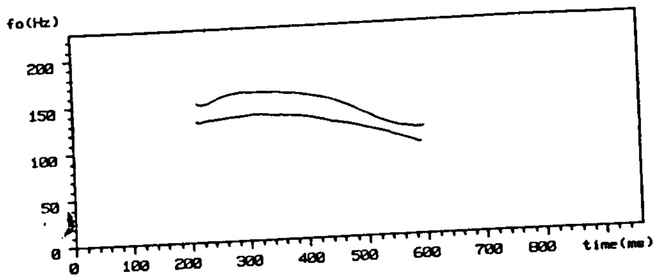


FIGURE 1. F_0 contours for Kammu tones: tɨ́ɨ "to beat", tɨ̀ɨ "place". The words were said in isolation by a male speaker.

short vowel have developed high tone, and words with an originally long vowel have low tone. Examples of this are given below. Cognates from Kammu and Lamet are given, since these languages preserve vowel length. Lamet is a Palaungic language which has developed tense (´) and lax (`) registers under conditions similar to those which have given rise to high and low tones in Kammu.

	Hu	Kammu	Lamet	
*short:	yám		yàm	"to die"
	phɨ́ɲ	píɲ	pɨ́ɲ	"to shoot"
	θúk		khúk	"hair"
*long:	yàm	yàam	yàam	"to cry"
	thàɲ	táaɲ	táaɲ	"to weave"
	nasòk		yóok	"ear"

One possible phonetic explanation why short vowels are associated with high tone and long vowels with low is that a short vowel would get a higher average F_0 than a low one, provided that the usual intonation pattern is falling, that the long and the short vowels start at the same F_0, and that the F_0 slope is constant. Whether or not this is true for Hu is not known, but these assumptions do not seem unreasonable. It may also be the case that the originally long and short vowels have slightly different quality in Hu. According to Hartmut Traunmüller (pers. comm.) long low vowels have slightly lower intrinsic pitch than short low vowels, but the difference is too small to be audible.

U

U has a four-tone system, with a high level (´), a low level (`), a rising (˘) and a falling (^) tone (Svantesson forthc.b.). The falling tone does not occur on syllables with a final stop, and the rising tone is rare on open syllables and syllables ending in a sonorant.

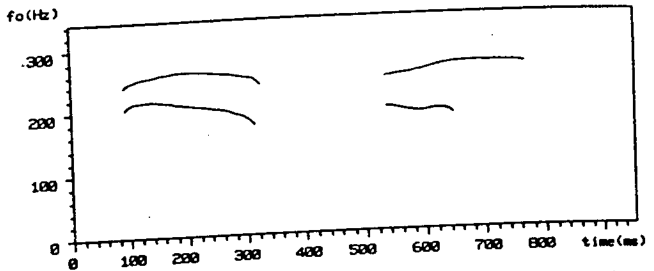


FIGURE 2. F_0 contours for Hu tones: yám "to die", yàm "to cry" (left); khát "sick", khàŋ "jaw" (right). The words were said in isolation by a female speaker.

The tones are illustrated in Figure 3.

Proto-Angkuic lacked tones, and tones have developed independently in Hu and U as is proved by the following facts: final nasals changed into the corresponding stops after originally short vowels in U but not in Hu. This process, which thus took place after U and Hu separated, must have preceded loss of vowel length in U. Since syllables with originally short vowel followed by a *stop and a *nasal have different tones, tone development must have started before vowel length disappeared in U. In Hu, however, tones developed in connection with loss of vowel length.

The Angkuic consonant shift, which had taken place already in Proto-Angkuic (thus before tonogenesis) made all obstruents in the language voiceless, so that the oppositions voiceless/voiced and obstruent/sonorant are equivalent in the Angkuic consonant system. For this reason, a Kammu type of tonogenesis is impossible in Angkuic.

Based on these observations, the following scenario for U tonogenesis can be given. This is somewhat speculative, but each step can be motivated phonetically, and is accompanied by segmental changes which transfer functional load from segments to tones.

(1) A final sonorant (or open syllable) lowers F_0 and a final voiceless obstruent raises F_0 in the final part of the preceding vowel. It is well-known that voiceless consonants raise, and voiced consonants lower F_0 in the following vowel, and in some of the investigations cited by Hombert 1978:92 a similar but smaller effect on the vowel preceding a consonant was found.

The tones created by this rule became phonemic when final nasals were denasalized after short vowels, resulting in minimal or near-minimal pairs as:

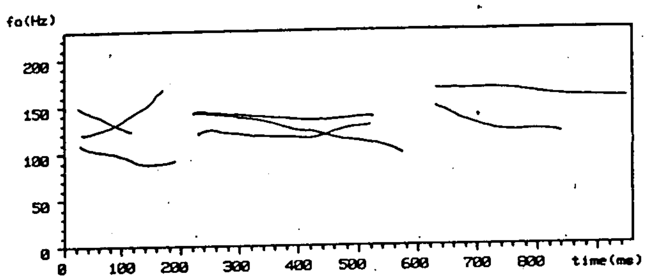


FIGURE 3. F_0 contours for U tones: khát "cold", sàt "five", lăt "fear" (left); phón "four", mphùn "seven", yâm "to cry" (mid); sí "rope", sì "tree" (right). The words were said in isolation by a male speaker.

U	Hu	
sáʕ	θák	"rice"
sàʕ	sáŋ	"bitter"
mpét	pét	"to spit"
phèt	phɨ́ɲ	"to shoot"

The Hu high tone shows that these words had short vowels. The first member of each pair had a final stop and the second a final nasal. When the nasal changed into a stop (k later became ʕ in U), merger was prevented by this tone development.

(2) Because of the raising and lowering of fundamental frequency in the final part of the vowel, short vowels get level (high or low) tone, and long vowels get contour tones (rising or falling). In tone languages with a vowel length opposition, it is not unusual that contour tones occur only on long syllables (in Thai, for instance, short vowels followed by a stop do not normally carry contour tones). After long and short vowels had merged, these tone patterns remained and became phonemic:

		FINAL:	
		*voiceless	*voiced
VOWEL:	*long	rising	falling
	*short	high	low

In this way, the vowel length opposition was replaced by a tone opposition in U, but in a different way than in Hu. The stage reached after this rule has applied is identical to the present state of the language, except that the falling tone is retained only when the initial cluster was completely voiced, i.e. consisted of sonorants only. Otherwise it has been modified by rules (3) and (4) below. Examples of the different tones created by rules (1)-(2):

U	Hu	Lamet	
*short vowel, *voiceless ending:			
khát	khát	kát	"cold" (Hu: "sick")
káʕ	kák	kàk	"to bite"
súʕ	θúk	khúk	"hair"
ʔáʕ	phaʔát	ʔɛ́s	"to swell"
*short vowel, *voiced ending:			
yàp	yám	yàm	"to die"
sàt	paθán	phán	"five"
mphà	phɨ́ʁ	mpɨ́r	"to fly"
ŋàw	ŋál	ŋàl	"fire"
*long vowel, *voiceless ending:			
ntshăt	nθàc	máac	"sand"
qhăʕ	tʁàk	tráak	"buffalo"
sŭʕ	nasòk	yóok	"ear"
ʔăʕ	ʔàk		"crossbow"
*long vowel, *voiced ending:			
mâ	mà	màar	"field"
yâm	yàm	yàam	"to cry"
ŋâ	ŋáʔ	ŋàaʔ	"to itch"
mî	méʔ	mìi̯ʔ	"you"

The falling tone was further changed in the following ways:

(3) When a vowel with a falling tone was preceded by a voiceless consonant (obstruent) or a cluster containing a voiceless consonant, it became a high level tone. There was probably an allophonic variation between a high falling and a low falling tone, conditioned by voiceless and voiced initial cluster. (Some words from another U language given by Zhōu and Yán 1983 seem to confirm this.) Reduction of initial clusters led to phonemization of the tone allophones, and the high falling tone then became high level:

U	Hu	Lamet	
thám	thàm	ktáam	"crab"
pán	pàɲ	pàaɲ	"white"
kíɑ̃	càŋ	cèeŋ	"foot"
wáy	kaʔɔ̀y	ʔlɔ́ɔy	"three"

A minimal pair is xáɑ̃ "thorn" vs. xâɑ̃ "flower". The word for "thorn" is ráaŋ in Lamet, where the tense register shows that there was a h cluster, whereas Kammu ràaŋ "flower" with low tone points to a voiced initial.

(4) In open syllables (corresponding to final glottal stop in Hu and Lamet), the high level tone split into high and low tones, depending on vowel height, so that high vowels got high tone and non-high vowels got

low tone. This rule is phonetically well motivated, since high vowels have higher intrinsic pitch than non-high vowels, but this mechanism is seldom used to generate tones in actual languages (see Hombert 1978:96). Examples from U:

	U	Hu	Lamet	
*high vowel:	qí	pʁí?	prìi?	"nature"
	sí	pasí?	plsí?	"rope"
	ŋkú		ŋkùu?	"skin"
	nthú		ntú?	"hole"
*non-high vowel:	khà		káa?	"fish"
	salè	salé?	slɛ̀ɛ?	"rain"
	sì	θé?	khé?	"tree"
	sò	sɔ́?	sɔ́?	"dog"
	ŋkhù	ŋkhó?		"rice"

Under certain conditions, mid vowels have become high, so that the oppositions i/e and u/o have been partially replaced by tone oppositions í/ì and ú/ù (cf. the pairs sí/sì and ŋkú/ŋkhù).

CONCLUSION

Tones have developed independently in the three closely related languages Kammu, Hu and U, showing that presence or absence of tones in a language cannot be taken as an indicator of genetic relationship. Each of these languages has acquired tones in its own way, which shows that at least in areas where there is a strong areal pressure on languages to acquire tones, this can be done by other mechanisms than those generally used to explain the origin of tones.

REFERENCES

Gårding, Eva and Kristina Lindell. 1977. "Tones in northern Kammu: a phonetic investigation". **Acta Orientalia** (Copenhagen) 38, 321-32. Also published in **Working Papers** (Dept. of Linguistics, Lund University) 16 (1978), 19-29.

Hombert, Jean-Marie. 1978. "Consonant types, vowel quality, and tone" Chapter 3 in Victoria Fromkin, ed. **Tone. A linguistic survey**, 77-111. New York: Academic Press.

Svantesson, Jan-Olof. 1983. **Kammu phonology and morphology**. Lund: Gleerup, 1983.

Svantesson, Jan-Olof. forthc.a. "Hu - a language with unorthodox tonogenesis". In Jeremy Davidson, ed. **Contributions to Mon-Khmer studies: essays in honour of Professor H.L. Shorto**. SOAS, London.

Svantesson, Jan-Olof. forthc.b. "U". Unpublished manuscript..

Zhōu Zhízhì and Yán Qíxiāng. 1983. "Bùlǎngyǔ gàikuàng". **Mínzú yǔwén** 1983:2, 71-81.

PHONETIC CONDITIONING FOR THE DEVELOPMENT OF NASALIZATION IN TEKE

JEAN-MARIE HOMBERT

C.R.L.S. - Université Lyon 2 et LACITO - C.N.R.S.
69500 BRON - FRANCE

ABSTRACT

Diachronic generalizations concerning vowel nasalization have been made on the basis of very restricted data. The development of nasalized vowels is a current on-going process in Teke languages. Comparative data from three different languages of this group allow better understanding of the interaction of the various phonetic factors at work (e.g. vowel quality and length, place of articulation of nasal consonants, etc.)

UNIVERSAL TENDENCIES OF VOWEL NASALIZATION

A number of fairly recent studies [1,2,3,4] propose generalizations concerning nasalized vowels both from a synchronic and a diachronic point of view. Synchronically, the number of nasal vowels (V̰) in a given language never exceeds the number of oral vowels (V). Among the languages which have V̰, approximately half of them have a number of V̰ equal to the number of V. Furthermore the quality of V̰ is generally more centralized than its oral counterpart. Diachronically it is claimed that vowel nasalization:

- originates from the loss of a nasal consonant (N) in postvocalic position more often than in prevocalic position, i.e. VN > V̰N > V̰ˑ is more common than NV > NVˑ > V̰.
- affects low vowels first and high vowels last.
- affects front vowels before back vowels of similar height [5] (i.e. [e] earlier than [o] and [i] earlier than [u]).
- occurs first before labial nasal consonants second before dental nasals and last before velar nasals.

It should be emphasized that if synchronic generalizations seem to be well founded because of the size of the language samples taken into consideration, the situation is quite different for diachronic generalizations. They have been made almost exclusively on the basis of two language groups: Chinese and Romance. In order to distinguish between language specific and language universal conditioning factors it is crucial to increase our diachronic data base by examining the development of nasalization in other language groups, either directly - through the use of written documents - or indirectly, through the study of closely related languages currently acquiring vowel nasalization. This second case is the one that we will now consider.

TEKE LANGUAGES

There are about 15 Teke languages spoken in Congo, Gabon and Zaïre. They belong to the Bantu branch (B70 [6]) of the Niger-Congo family. Some Teke languages have nasalized vowels (e.g. Ibali, Ndzindziu) while others lack them (e.g. Atege).

In Ibali, V̰ in noun and verb forms[1] are found only with long or double vowels: e.g. –tá̰á̰ (sole of foot), –gìɔ̰́ (bat), –kúɔ̰̌ (to sweep). In Ndzindziu, V̰ can be short: –lô̰ (husband), preceded by a glide: –kʲô̰ (monkey), or double: –táɔ̰́ (sole of foot), –gìɔ̰́ (bat), –kúɔ̰̌ (to sweep).

Most Bantu nouns have a C_1 V_1 C_2 V_2 structure; verb forms have a similar structure with a V_2 = a in the infinitive. Table 1 summarizes the correspondences between nasalized vowels (in Ibali and Nzindziu) and non-nasalized forms (in Atege). Examples can be found in Annex A. It is clear from Table 1 that nasalization results from the loss of a labial nasal consonant in C_2 (intervocalic) position. This process is at a more advanced stage in Ndzindziu than in Ibali. Thus, the comparison between these languages allows us to propose a relative chronology of nasalization development.

[1] for a more complete presentation see [7]

Table 1. C_1 V_1 m V_2 in Atege and their correspondences in Ibali and Ndzindziu.

ATEGE	IBALI	NDZINDZIU
C_1 i m V_2	C_1 i m V_2 [2]	C_1 ʲɔ̰
C_1 e m V_2	C_1 e m V_2	
C_1 a m V_2	C_1 a m V_2	C_1 ɔ̰
C_1 o m V_2	C_1 o m V_2	
C_1 u m V_2	C_1 u m V_2	C_1 oɔ̰ , C_1 o̰
C_1 ii m V_2	C_1 ii m V_2	C_1 iɔ̰
C_1 ie m V_2	C_1 iɔ̰	
C_1 aa m V_2	C_1 a̰a̰	C_1 aɔ̰
C_1 uo m V_2	C_1 uɔ̰	C_1 uɔ̰
C_1 uu m V_2	C_1 uu m V_2	C_1 uɔ̰ , C_1 uo̰

The following parameters appear to play a significant role:

- vowel length: long vowels[3] (bottom half of Table 1) are more prone to nasalization than short ones; in Ibali short vowels have not been nasalized yet.
- vowel quality: close vowels nasalize last[4]. In Ibali [ii] and [uu] have not undergone nasalization.
- place of articulation of nasal consonant: only the labial nasal consonant [m] triggered nasalization, dental nasals are preserved in all three languages and do not nasalize adjacent vowels as illustrated in Annex B. As for velar nasals, they have disappeared without leaving traces of nasalization.

Two final remarks can be made concerning the quality of the resulting nasalized vowels:

- there are no instances in Ibali and in Ndzindziu of front unrounded nasalized vowels (i.e. ḭ, ḛ or ɛ̰).
- when V_1 is [u] or [uu] in Atege and Ibali, we have two sets of correspondences in Ndzindziu ([oɔ̰] and [o̰] correspond to [u] and [uɔ̰] and [uo̰] to [uu]).

A closer look at corresponding forms in Atege show that the less open nasalized vowels in Ndzindziu ([o̰] (as opposed to [oɔ̰]) and [uo̰] as opposed to [uɔ̰]) are found when V_2 is a high vowel ([i] or [u]) in the non-nasalized Atege forms.

Table 2. Effect of V_2 on the quality of Ndzindziu nasalized vowels

	ATEGE	NDZINDZIU
chief	pfúmú	pfó̰
powder (against rhumatisms)	bùmì	bɔ̰̀
name	kúúmí	kfúó̰
middle part of body	lùùmù	lùɔ̰̀
to rumble	džùmà	dzɔ̀ɔ̰̀
to send	túmà	tóɔ̰̀
to buy	súúmà	fúɔ̰̀
to rest	wúúmà	wúɔ̰̀

It is only when V_1 is [u] (or [uu]) that V_2 seems to play a role in the determination of the quality of the resulting nasalized vowel.

PHONETIC CONDITIONING

Numerous fiberscopic and X-ray studies of the velum position as a function of vowel quality have shown that low oral vowels are produced with a relatively low velum position. Consequently it is not surprising that a nasal leakage could occur with these low oral vowels resulting in nasalized vowel quality. Our data showing that [i] and [u] are the last vowels to undergo nasalization are in agreement with this phonetic observation.

The fact that vowel length can play a determinant role in the development of nasalization has not been emphasized in the phonetic literature. It does, though, make perfect phonetic sense that, on a long vowel, the soft palate has more time to anticipate its lowering movement in preparation for the following nasal consonant. Moreover, the positive correlation between length and nasalization can also be linked to the role of vowel quality mentioned above: it could also be that it is because they are phonetically longer that low vowels are nasalized first.

The order proposed by Chen [5] with respect to the role of the place of articulation of the nasal consonant is only partially followed in Teke; [m] is the first nasal to trigger vowel nasalization and [n] seems to follow as suggested by data recently collected by Paulian [8] but [ŋ] has disappeared in Teke languages without nasalizing adjacent vowels.

It seems clear that in our data the quality of ṽ is strongly influenced by the triggering labial nasal consonant[5]: there is no unrounded ṽ even when the original vowel was (front) unrounded. The labiality of the consonant has been transferred onto the adjacent vowel.

It is extremely difficult at this point to decide whether the effect of the nasal consonant was perseveratory or anticipatory (i.e. whether intermediate stages between Atege and Ibali for instance, were C_1 V_1 m V_2 > C_1 V_1 m ṽ > C_1 V_1 ṽ or rather C_1 V_1 m V_2 > C_1 ṽ m V_2 > C_1 ṽ m > C_1 ṽ.

The fact that V_2 played a role in the determination of the quality of ṽ (at least when V_1 was [u] or [uu]) pleads in favor of the first solution. On the other hand, though, the fact that V_1 quality and length also played a role in triggering nasalization favors the anticipatory solution[6]. It is also possible that perseveratory and anticipatory assimilations played a role simultaneously. Notice that if perseveratory assimilation is proposed one has to explain why it did not affect V_1 when C_1 was [m]. It seems that it is because in these languages[7] (as in most Bantu languages) the first syllable is accented and consequently is not subjected to a number of phonetic changes commonly found in pre or post-accented syllables (e.g. vowel reduction or loss). Only data from other Teke languages illustrating intermediate stages will allow us to get a better understanding of the respective role of perseveratory vs anticipatory assimilation.

CONCLUSION

Data presented in this paper:

- confirm the role of vowel height in the development of vowel nasalization;
- are in partial agreement with respect to the role of place of articulation of the nasal consonant (m before n);
- are unclear with respect to perseveratory vs anticipatory assimilation;
- do not make any claim concerning the role of front/back vowel quality;
- illustrate the role of a phonetic factor generally not mentioned: vowel length.

It is obvious that we need similar studies from other language groups in order to be able to sort out language specific conditioning factors from more universal phonetic constraints. It is from these universal phonetic constraints that sound changes can originate; whether or not they are actually activated depends on non-phonetic factors (e.g. phonological or sociolinguistic).

[2] There is an apparently irregular correspondence between Ibali and Ndzindziu for three forms with [i] in V_1 position:

to dig	Ib.	−tsímà	Nd	−tšúɔ̰̀
to think	Ib.	−tsìmà	Nd.	−tšùɔ̰̀
to sing	Ib.	−yímà	Nd.	−yúɔ̰̀

Table 1 would lead us to expect Ndzindziu forms with ʲɔ̰. This discrepancy is partially explained by the fact that, for these words, V_1 first changed to [u] before the loss of the nasal consonant as attested by the corresponding Atege forms tšúmà, tšùmà and yúúmà.

[3] Historically these long vowels come from a lengthening of short vowels before prenasalized stops: C_1 V_1 mb V_2 > C_1 V_1: mb V_2 > C_1 V_1: m V_2 These prenasalized stops are still preserved in neighboring closely related (but non-Teke) languages such as Duma, Nzebi, Tsengui, Wandzi.

[4] Atege i.e. [ie] et [uo] correspond to earlier [ee] et [oo] respectively.

[5] In Ngungwel, when the triggering nasal consonant was n, the resulting quality of ṽ is [ɛ̰]: dental consonants are well known to push the vowels towards the front of the vowel space.

[6] to get a more complete picture, nasalization of prefixes should also be taken into account (see [7]).

[7] see for instance [9].

REFERENCES

[1] C.A. FERGUSON, L.M. HYMAN and J.J. OHALA (eds.), "Nasalfest: Papers from a symposium on nasals and nasalization", Standford University: Language Universals Project, 1975.

[2] M. RUHLEN, "Nasal vowels", in: J.H. GREENBERG (ed.), "Universals of Human Language", Standford University Press, 203-241, 1978.

[3] J. CROTHERS, "Typology and Universals of Vowel Systems", in: J.H. GREENBERG (ed.), "Universals of Human Language", Standford University Press, 93-152, 1978.

[4] I. MADDIESON, "Patterns of Sounds", Cambridge University Press, 1984.

[5] M. CHEN, "An areal study of nasalization in Chinese", in: C.A. FERGUSON, L.M. HYMAN and J.J. OHALA (eds.), "Nasalfest", 81-124, 1975.

[6] M. GUTHRIE, "Comparative Bantu", vol. 2, Gregg Int. Publishers Ltd., 1971.

[7] J.M. HOMBERT, "The development of nasalized vowels in the Teke language group (Bantu)", in: K. BOGERS, H. van der HULST and M. MOUS (eds.) "The Phonological Representation of Suprasegmentals", Foris Publications, 359-373, 1986.

[8] C. PAULIAN, "Nasalization in Ngungwel", (in preparation).

[9] C. PAULIAN, "Le Kukuya: langue Teke du Congo", Paris: SELAF, 1975.

ANNEX A. Examples illustrating Table 1.

	ATEGE	IBALI	NDZINDZIU
monkey	−kímà	−kímà	−kʲɔ̰̂
hoe	−tɛ́mì	−tɛ́mù	−tʲɔ̰̂
slowness	−lɛ̀mɛ̀	−lɛ̀mɛ̀	−lʲɔ̰̀
conversation	−sàmí		−sɔ̰̌
to shout	−yámà		−yɔ̰̂
python	−bòmò	−bòmò	−bɔ̰̀
to enter	−sɔ̀mɔ̀	−sɔ̀mɔ̀	−sɔ̰̀
husband	−lʊ́mì	−lʊ́mì	−lô̰
to climb	−kʊ́mà	−kʊ́mà	−kóɔ̰̀
to swell	−bíímà	−bíímà	−bíɔ̰̀
debt	−bíímì		−bíɔ̰̀
to touch	−bìɛ̀mɛ̀	−bìɔ̰̀	−bìɔ̰̀
finger	−líɛ́mì	−líɔ̰̀	−líɔ̰̀
to patch	−bààmà	−bà̰à̰	−bà̰ɔ̰̀
lizard (k.o)	−báámì	−bá̰á̰	−bá̰ɔ̰̀
to sweep	−kʊ́ómò	−kʊ́ɔ̰̀	−kʊ́ɔ̰̀
musical bow	−gʊ̀ɔ̀mí		−gʊ̰̀ɔ̰́
buyer	−sʊ́ʊ́mì		−fʊ́ɔ̰̀
to buy	−sʊ́ʊ́mà	−fʊ́ʊ́mà	−fʊ́ɔ̰̀

ANNEX B. Dental nasals in C_2 position in Atege, Ibali and Ndzindziu.

	ATEGE	IBALI	NDZINDZIU
to dance	−kínà	−kínà	−kínà
to finish	−mànà	−mànà	−mànà
to plant	−kʊ́nà	−kʊ́nà	−kʊ́nà
to be black	−pììnà	−pììnà	−pììnà
rat (k.o.)	−bíɛ́nɛ́	−bíɛ́nɛ́	−bíɛ́nɛ́
to begin	−báánà	−báánà	−báánà

TONOGENESIS IN NORTHERN TEPEHUAN

BURT BASCOM

SUMMER INSTITUTE OF LINGUISTICS
INTERNATIONAL LINGUISTIC CENTER
7500 CAMP WISDOM ROAD
DALLAS, TX 75236

ABSTRACT

Northern Tepehuan, a Uto-Aztecan language of Mexico, displays contrastive pitch on clusters of two vowels. These pitch contrasts have been described as phonemic tone[1]. Northern Tepehuan, Southern Tepehuan, Upper Piman and Lower Piman form the Tepiman Branch of Uto-Aztecan. The loss of Proto-Tepiman *ˀ and *h has resulted in vowel clusters in Northern Tepehuan, thus providing some of the environments for the contrasting tones. The other three Tepiman languages display stress in corresponding environments. This incipient tone system presents an ideal situation in which to examine once more the ways in which tone develops in a language.

INTRODUCTION

In his article "Tonogenesis in Southeast Asia", James A. Matisoff (1973) says: "...it appears that to become truly tonal a language must have a basic monosyllabic structure. Polysyllabic languages may develop 'pitch accent systems...'" Matisoff refers to the latter as "marginally or incipiently tonal"[2].

It is precisely this "incipient" or "marginal" nature of the contrastive pitch phenomena in Northern Tepehuan which provides the motivation for this paper. The precise definition of a tone languages has yet to be agreed upon. This paper will reflect the view that tone is present where contrastive pitch is found on the lexical level.

Northern Tepehuan is spoken by approximately eight thousand people living in the mountains of Chihuahua in Northern Mexico. Northern Tepehuan (NT), Southern Tepehuan (ST), Papago (UP), and Pima (LP) form the TEPIMAN sub-group of the Sonoran Branch of Uto-Aztecan. This paper reflects the field work done by the author in these languages.

Because NT has a relatively simple tone system (only two tones), and because the tone contrasts are restricted to vowel clusters, tone has a very low functional load. Since the other three Tepiman languages do not have tone, it seems very likely that tone is just developing in Northern Tepehuan and that it is therefore an ideal situation in which to inquire about how tone originates in a language.

SYLLABLE STRUCTURE

The syllable in NT must always have a V or a VV as its nucleus. It may have a C onset and/or coda. It may be short, i.e., contain a single vowel; or long, i.e., contain a vowel cluster. The VV of the long syllable may be geminate or diverse. This unit displays the contrasting pitch patterns of NT. All four possible tone sequences of high and low tone on a sequence of two vowels occur in NT. I accept this fact as a part of the evidence for tone in NT. The following examples show the structure of the syllable as described:

V	á.ki	'stream'
VV	áá.ki	'popcorn'
VC	áš.t^{y}a.ñi	'throw it out'
CV	bá.vi	'beans'
CVC	táš.ka.li	'tortilla'
CVV	daá.ka	'nóse'

NT words may consist of as many as eight syllables (or more if one includes clitics) as seen in: ga.ma.máá.tɨ.tu.li.t^{y}a.dai 'he was teaching someone'.

While long syllables play an important role in NT phonology, not every word is required to have a long syllable. In one limited environment long consonants appear to "take the place of" long vowels. Following a short high-toned initial syllable, a consonant is lengthened as in: /bávi/ [ˈbábּ·iI̥] or [ˈbábּ·ƀiI̥] 'beans'. But bavígadɨ 'his beans' has no long syllable.

TONE IN NORTHERN TEPEHUAN

Northern Tepehuan has two contrastive pitches (tones), high (´) and low (\` in phonetic representations and unmarked in phonemic representations). There is at most one high-toned syllable in a stem. Any of the following qualifies as a high-toned syllable: V́, V́V́, V́V, or VV́. Pitch contrasts occur only on VV (sequences of two vowels). A VV is the nucleus of a long syllable. There is at most one long syllable in a stem, which is considered to be the nucleus of the stem. The long syllable is not always the high-toned syllable. Stems may be composed of from one to three syllables. For example:

/móo/	['móòO]	'head'
/móódɨ/	['móó.dìɨ]	'his head'
/maákaɨ/	['màá.kàiI]	'he gives'.
/kɨlívi/	[kɨ.'lí.ƀiI]	'he shells corn'

The claim that stems have no more than one long syllable in **NT** is not contradicted by maákai since the final -i is not a part of the stem. It is the "present" or better the "atemporal" verb suffix. The final -i of nouns is the "absolutive" suffix.

The evidence for tone is found on the following (C)VV sequences:

/áàši/	['áàšiI]	'catfish'
/áaši/	['ááššiI]	'are they others ?'
/aáši/	['āášiI]	'he laughed'
/aašíši/	['āāšíšiI]	'is it a catfish ?'
/tuudákɨi/	['tūūdákɨìI]	'he dances'
/tóódakɨi/	['tóódākɨìI]	'it beats' (the heart)

For further evidence of contrastive pitch see Pike, Barrett and Bascom[1].

STRESS

Stress in **NT** is predictable. Phonetically, stress is loudness. Most frequently there is only one stress per word. It occurs with high tone, as in:

/nóvi/	['nóƀ·iI]	'hand'
/gɨmóo/	[gɨ'móòO]	'your head'.

In words which begin with CVVCV́ or CVV́ stress fluctuates. In this special environment the stress fluctuates between the low-toned V(V) and the high-toned vowel which follows, or both may be stressed, as in:

/naadámi/	['nāādámiI]	
or	[nāā'dámiI]	
or	['nāā'dámiI]	'six'
/kášiš koí/	['kášiš 'kōí]	
or	['kášiš kō'í]	
or	['kášiš 'kō'í]	'Did he already go to sleep ?'

Compound words may occur with two high-toned syllables accompanied by stress.

THE DEVELOPMENT OF TONE IN NORTHERN TEPEHUAN

Some of the most significant sound changes in the development of the Tepiman languages from Proto-Tepiman (**PT**) involve the loss of *ʔ and *h (h is a glottal fricative in **ST** and **UP**, and a velar fricative in **NT** and **LP**). These changes are related to the development of tone in **NT**, since they have in some instances resulted in the vowel clusters where tone contrasts occur.

All **PT** *ʔ are lost in **NT**. Word initial *ʔ has been retained in **ST**, **UP**, and **LP**. Non-initial **PT** *ʔ is retained in **ST** and is split to ʔ/0 in **UP** and **LP** (*ʔ is the onset of a syllable, not part of the syllable nucleus). The following examples show some of the correspondences displaying these sound changes:

PT		**NT**	**ST**	**UP**	**LP**	**Gloss**
*ˈʔoobai	>	oóbai	ˈʔoob	ˈʔoobɨ	ˈʔoob	'foreigner'
*baˈʔagai	>	báágai	baˈʔaaʔ	ˈbaʔagɨ	ˈbaʔag	'eagle'
*ˈmoʔo	>	móo	ˈmoʔ	ˈmoʔo	ˈmoʔo	'head'
*ˈʔaapiʔi	>	aápi	ˈʔaapiʔ	ˈʔaapi	ˈʔaapi	'you'

Initial **PT** *h has almost completely disappeared in **NT** (some speakers use the archaic forms with initial h); has split to h/0 in **ST** and has been retained in **UP** and **LP**. Non-initial **PT** *h has split in all four languages to h/0. The following examples show some of the correspondences displaying these sound changes (see Bascom 1965[3] for a complete set of correspondences).

PT		**NT**	**ST**	**UP**	**LP**	**Gloss**
*ˈhaaki		ááki	ˈhaak	ˈhaaki	ˈhaahak	'popcorn'
*hioˈsigai	>	yoošígai	ˈyoošiʔ	ˈhɨosigɨ	ˈhioškam	'flower'
*ʔiahaˈtagi	>	yaatági	ˈʔiatgi-	ˈʔiatogi-	ˈʔiahtg-	'to lie'

The purpose of this list of examples is to present evidence for the contrasting development of **NT** phonology as compared with **ST**, **UP**, and **LP**, especially to show that, while tone has developed in **NT**, stress has remained contrastive in the other three languages. The examples have been chosen to illustrate the four tone patterns on vowel clusters in **NT** (high-high, high-low, low-high and low-low).

Source of high-high tone sequence.

The high-high tones occurring on a (C)VV sequence have three sources in **PT**.

Loss of *ʔ. Both *ˈCVʔV... and *CVˈʔV... > **NT** CV́V́... The three dots (here and subsequently) indicate that the sequence is followed by one or more syllables in the same word, as in:

PT		**NT**	**Gloss**
*ˈgɨʔɨri	>	gɨ́ɨ́li	'boy'
*ˈkoʔokori	>	kóókoli	'chile'
*ˈmuʔidu	>	múíd^{y}u	'there are many'
*ˈvaʔigɨi	>	váígɨi	'she fetched water'
*baˈʔagai	>	báágai	'eagle'
*koˈʔoko	>	kóóko	'it hurts'
*vaˈʔaki	>	vááki	'house'
*ʔiˈohogɨi	>	yóógɨi	'he coughs'

Loss of *h. (either initial or medial) *ˈhVV... > **NT** V́V́ and *ˈCVhV... > CV́V́. If *h represented an unequivocal example of the loss of a laryngeal resulting in a sequence of high tones, this would be worthy of note. However, only two of the daughter languages actually have the glottal fricative for the reflex of *h, while the other two have a velar fricative reflex.

*ˈhaaki	>	ááki	'parched corn'
*ˈhaahaga	>	áága	'leaves'
*ˈhoonita	>	óóñt^{y}ai	'he takes a wife'
*ˈhuuhutu	>	úútu	'fingernails'
*ˈbihugimu	>	bíúgimu	'he is hungry'
*ˈhahaduñi	>	ááduñi	'relatives'
*ˈhihina	>	ííña	'he shouts'
*ˈmɨhida	>	mɨ́ɨ́d^{y}a	'he burns it'

Two forms do not follow this rule:

*ˈtahapai	>	taápai	'he split it'
*ˈtɨhanai	>	tɨánai	'he orders'

They seem to follow the rule that **PT** *ˈCVV > **NT** CV́ which might indicate that they lost the *h before the tone rule was applied.

Non-final *Vi. *ˈCVi... > CV́í when followed by other syllables in the same word.

*ˈdaikaroi	>	dáíkaroi	'chair'
*ˈkaidɨ	>	káíd^{y}ɨ	'its seed'
*ˈkɨisa	>	kɨ́íša	'he stepped on it'
*ˈsoiga	>	sóíga	'domesticated animal'
*ˈvoisikai	>	vóíšikai	'he sweeps'
*ˈkoi-	>	kóí	'he killed them'

Source of low-high tone sequence.

The low-high sequence of tones on a CVV syllable in **NT** comes from *ˈCVV under the following conditions: #*ˈCVV(...) > **NT** CVV́(...), # means that *ˈCVV is initial in the word; (...) means that #*CVV is optionally followed by one or more syllables in the same word; neither *CVi (non-geminate) nor *hVV is followed by another syllable; if *C is *h or *ʔ then **NT** is 0.

*ˈbaaba	>	baába	'mother's mother'
*ˈbiitai	>	biít^{y}ai	'excrement'
*ˈdaada	>	daáda	'mama'
*ˈmɨɨ	>	mɨɨ́	'he ran'
*ˈtuu	>	tuú	'he put out the fire'
*ˈʔii	>	ií	'he went'
*ˈhuu	>	uú	'he ate'
*ˈdai	>	daí	'he set it down'
*ˈkoi	>	koí	'he slept'

Note that in **NT** the contrast between kóí 'he killed them' and koí 'he slept' the former comes from a **PT** bisyllabic form while the latter comes from a **PT** monosyllabic form.

Source of high-low tone sequence.

The high-low sequence of tones on a CVV syllable in **NT** comes from **PT** *ˈCVʔV when this sequence represents the whole stem. It is followed in the same word only by 0 or -i.

*ˈmoʔo	>	móo	'head'
*ˈmuʔi	>	múi	'many'
*ˈtuʔi	>	túi	'flour'
*ˈkoʔoi	>	kói/kóóyi	'snake'
*ˈnɨʔɨi	>	nɨ́i/nɨ́ɨ́yi	'he sings'
*ˈʔiʔɨi	>	yɨ́i/yɨ́ɨ́yi	'he drinks'

Source of low-low sequence of tones.

The low-low sequence of tones on a CVV syllable in **NT** comes from an unstressed long syllable in **PT**, as in:

*baaˈbahi	>	baabáhi	'tails'
*baaˈbanai	>	baabánai	'coyotes'
*daaˈdaka	>	daadáka	'noses'
*daaˈkadɨ	>	daakádɨ	'his nose'
*duuˈkami	>	duukámi	'official'
*gɨɨˈsimi	>	gɨɨšími	'he is falling'

An alternate approach.

This study is not an exhaustive one. That would require much more comparative work. Nevertheless, the primary thrust of the paper, namely, that the pitch contrasts observed in **NT** have developed in part from the loss of **PT** *ʔ and *h, is well established by the evidence presented. How phonetic facts are interpreted in a phonological analysis depends in part on the basic assumptions of the analyst. While this paper has analyzed the pitch contrasts as tone, the fact that **NT** tone contrasts occur in such a limited environment makes it desirable to entertain the possibility of an alternate solution. Nancy Woo has done this in her paper on "Tone in Northern Tepehuan"[4]. Her analysis claims to account for all of the pitch phenomena in **NT** on the basis of a set of rules involving the shapes of stems, historical and comparative information, and the introduction of a rule involving a special kind of "syllabic" feature. Since she was not working directly with native speakers of Northern Tepehuan she did not have all of the information about the language in hand, thus her analysis does not handle all the forms. For example: marádɨ 'his child' or 'its branch' has two plural forms whose tones contrast, i.e., maamáradɨ 'his children' vs. máámaradɨ 'its branches'. If I have understood Woo's rules correctly, they predict the former but not the latter.

Some of the assertions Woo makes which reveal her underlying assumptions about the nature of phonological analysis and which seem to be essential to her thesis are at best debatable. The claim, at one point in her argument, that speakers of a language "remember" lost laryngeals (historically lost, not synchronically) is not a concept accepted by all linguists.

Even if one grants that Woo's rules do handle the material she had access to and might well be expanded to cover all of the phonological facts, it remains that some prefer to stay closer to the observable phonetic phenomena and live with fewer generalizations; and also to treat historical and comparative facts as information to be considered after the synchronic picture has been developed.

Northern Tepehuan phonology has moved away from a strictly "stress-accent" system in Proto-Tepiman to a "pitch accent system" synchronically. As long as one keeps the total word in focus and does not introduce morphology, specifically stem boundries, into the phonological analysis, contrastive pitch is phonemic. I prefer to call this contrastive pitch tone, but am willing to settle for the term "pitch accent" as a viable alternative.

References

[1] K.L. Pike, R.P. Barrett and B. Bascom, "Instrumental Collaboration on a Tepehuan (Uto-Aztecan) Pitch Problem", Phonetica 3, 1959, pp. 1–22.

[2] James A. Matisoff, "Tonogenesis in Southeast Asia", Southern California Occasional Papers No. 1, July, 1973, pp. 71–95. Edited by Larry M Hyman.

[3] Bascom, Burton W., "Proto-Tepiman (Tepehuan-Piman)" Unpublished dissertation, 1965, University of Washington.

[4] Woo, Nancy, "Tone in Northern Tepehuan", International Journal of American Linguistics Vol. 36, 1970, pp. 19–30.

[5] B. Bascom, "Tonomechanics of Northern Tepehuan", Phonetica 4, 1959, pp. 71–88.

[6] B. Bascom, "Northern Tepehuan Grammatical Sketch" In R. Langacker (Ed), Uto-Aztecan Grammatical Sketches, Part III, Studies in Uto-Aztecan Grammar, 1982 pp. 269–93.

ÜBER DIE NATUR DER VOKALISCHEN ALTERNATIONEN IN DER KETISCHEN SPRACHE

H. Werner, N. Schablo

Pädagogische Hochschule Taganrog (UdSSR)

En utilisant le matériel de la langue des kets - celle d'une petite peuplade de Sibérie qui se rapporte au groupe des langues iénissiénnes on justifie dans l'exposé le point de vue d'après lequel les alternations vocaliques ont pris naissance historiquement sous l'influence des tons syllabiques. Les nuances qualitatives et quantitatives des phonèmes de voyelles acquièrent une signification phonologique et deviennent les phonèmes indépendants, une fois les tons sillabiques disparus. De cette manière les alternations vocaliques ont pris naissance, par ex. des "œil", pl. dɛs etc. La liaison des tons syllabiques à telles alternations vocaliques dans la langue des kets se conserve jusqu'à nos jours.

In der Fachliteratur über die Tonsprachen wird behauptet, daß die Silbentöne keinen Einfluß auf die Phonembasis ausüben. Vom phonologischen Standpunkt ausgehend könnte man diese Ansicht annehmen, denn der Phonembestand der entsprechenden Silben bleibt bei verschiedenen Silbentönen immer ein und derselbe, z.B. chinesisch ma^1 "Mutter", ma^2 "Hanf" ma^3 "Pferd", ma^4 "schimpfen". Auf der phonetischen Ebene liegen aber die Dinge anders. Die Einwirkung von Tönen verschiedenen Charakters verursacht die Erscheinung von verschiedenen qualitativen und quantitativen Schattierungen der entsprechenden Phoneme, welche infolge der historischen Konvergenz der Silbentöne unvermeidlich zu unterschiedlichen selbständigen Phonemen werden, da sie nun differenzierende Funktionen bekommen, die früher den Silbentönen zukamen. Diese differenzierenden Funktionen bekommen dabei vor allem die Vokale, und die in solchen Fällen entstandenen vokalischen Alternationen bleiben das einzige Reflex der gewesenen Tonalitätsunterschiede. Auf Grund der germanischen Sprachen wurde diese Voraussetzung von E.Sievers ausgesagt /1, S. 148-198/ und später von S.D.Kaznelson bestätigt /2, S. 308-312/. Im vorliegenden Bericht wird diese Ansicht auf Grund der ketischen Sprache entwickelt.

Das Ketische ist die Sprache der gegenwärtigen Jenissejer (der Keten), die am Mittel- und Unterlauf des Jenissej, sowie an einigen seiner Nebenflüsse wohnhaft sind. Bekanntlich gehört das Ketische zu den jenissejschen Sprachen, die eine isolierte Sprachinsel in Mittelsibirien darstellen; die meisten davon sind schon im XVIII. und XIX. Jahrhundert von osttürkischen, samojedischen und anderen Sprachen aufgesogen worden. Die Silbentonalität wurde im Ketischen vor mehr als zwei Jahrzehnten entdeckt und ist seitdem gründlich untersucht und beschrieben worden /3; 4/.

An der vokalischen Phonembasis der 4 ketischen Silbentöne läßt sich folgende bemerkenswerte Besonderheit verfolgen: die Vokale der mittleren Zungenhebung sind immer beim 1. Silbenton geschlossen, bei den anderen Tönen aber stets offen, z.B. des^1 "Auge" - dɛs^4 "Augen", dɛ?^{2} "See" - den^1 "Seen", qɔj^4 "Bär" - qon^1 "Bären" usw. Selbstverständlich gibt es diesen Unterschied auch bei den Vokalen der hohen und der tiefen Zungenhebung, obwohl er in solchen Fällen schwerer zu bemerken ist, z.B. s'ul'^{1} "Blut" - s'u?l'^{2} "sibirischer Weißlachs" - s'u:l'^{3} "Polarschlitten" - s'ul'^{4} "Haken für die Kinderwiege". Deutlich genug ist auch der Unterschied zwischen /ä/ (ein geschlossenes "a", das einem æ-Laut nahe steht) und /a/, z.B. täp^1 "Reifen" - pl. ta:ŋ^3; tät^1 "Stoßzahn" - ta:l'^{3} "Otter", s'ä^1 "Hausrat" - sa:l^3 "übernachten". Es sei aber darauf hingewiesen, daß der ä - Laut nicht gern vor solchen Konsonanten wie ŋ,q,k erscheint, vgl.: tak^1 "Kranich", taŋ^{1}a "Zugriemen (am Hundeschlitten) takt^1 russ. dial. "Tschir" (eine sibirische Fischart", qaq^1 russ. dial. "Jelez" (eine Fischart).

Man könnte vermuten, daß einige Erscheinungen der ketischen Sprache der oben formulierten Voraussetzung über die Einwirkung der Silbentöne auf die Phonembasis widersprechen. Es wären in dieser Hinsicht folgende Fälle zu erwähnen /5, S. 205/:

(a) in Beispielen wie di?^{2} "Baumstamm" - pl. da?n^2, i?^{2} "Bewahrungszelt" - pl. ɛ?ŋ^2, wo in grammatischen Formen ein und desselben Wortes bei verschiedener Phonembasis ein und derselbe Silbenton repräsentiert ist;

(b) in Beispielen wie ki^1 "Fangfalle" - pl. ki?ŋ^2, s'ul'^{1} "Blut" - s'u?l'^{2} "sibirischer Weißlachs", wo bei ein und derselben Phonembasis verschiedene Silbentöne vorkommen;

(c) der ä-Laut kommt nicht nur als Phonembasis des 1.Silbentones vor, sondern erscheint auch mit dem 2. und 4. Silbentönen, z.B. s'a?l'^{2}/s'ä?l'^{2} "Karausche", tar'^{4}/tär'^{4} "schlagen".

Hier läßt sich also eine Einwirkung von anderen Entwicklungstendenzen verfolgen, die späteren Charakters sind; der letztere Fall kommt besonders oft im südketischen Dialekt vor. Diese Erscheinungen sind meistenteils mit den Veränderungen in der lexisch-grammatischen Distribution der Silbentöne verbunden, die im Laufe der Zeit kontinuierlich vor sich gehen. Das läßt sich vor allem dadurch beweisen, daß ein und dasselbe Wort in verschiedenen jenissejschen Sprachen verschiedene Silbentöne haben kann, z.B. ketisch ta:l'^{3}, jugisch ta:r^4 "Otter"; ketisch täp^1, pl. ta:ŋ^3, jugisch tap^1, pl. ta:p^4 "Reifen"; ketisch ty?s'^{2}, pl. tʌ?ŋ^2, jugisch čy?s^2, pl. čʌ?ŋ^2, kottisch ši:š^1, pl. šeŋ < šɛ?ŋ^2 "Stein" usw.

Wie es dem auch sei, kann man behaupten, besonders auf Grund des ketischen Vokalismus der mittleren Zungenhebung, daß die qualitative und quantitative Charakteristik der Vokale in den gegenwärtigen ketischen Wörtern vor allem durch die Silbenakzentuation bestimmt ist und daß also bei den vokalischen Alternationen die Akzentuationserscheinungen dominieren. Man dürfte vermuten, daß durch diese Besonderheit der ketischen Sprache auch die Entstehung des Ablautes in Fällen wie s'es'^{1} "Fluß" - pl. s'as'^{4}, tip^1 "Hund" - pl. ta?p^2 verursacht ist. Jedenfalls ist eine andere akzeptable Deutung des Ablauts im Ketischen vorläufig unmöglich. Dabei lassen sich selbstverständlich auch einige Nebenerscheinungen verfolgen, die den Prozeß der historischen Entwicklung von Silbentönen beglitten haben. Hier wäre vor allem auf die Silbenkontraktion hinzuweisen, die zu qualitativen und quantitativen Veränderungen, sowie Veränderungen in der Akzentuation der entsprechenden Wörter führte. Dabei wurden labiale, velare und uvulare Konsonanten synkopiert, z.B. jugisch čafyr "Hundefutter" < ča?p^2 "Hunde" + ur^1 "Wasser", aber ketisch ta:l'^{3} dasselbe < ta?p^2 "Hunde" + ul'^{1} "Wasser"; jugisch afyŋ,ketisch a:ŋ "heiß"; jugisch xo:x^4 - pl. xoxyn, ketisch qoR^4 - pl. qo:n "Stern"; jugisch kʌxyn - pl. kʌxynyŋ, ketisch kʌ:n^3- pl. kɔn^1 "Fuchs". In anderen Fällen wurden bei der Kontraktion Vokale im Auslaut apokopiert, z.B. kʌ?n^2 "hell", aber kɔn^1 "Morgenröte" < kʌ?n^2 "hell" + i?^{2} "Tag", kys'^{1} "Russe" < kʌ?n^2 "hell" + si prädikatives Suffix. Ähnliche Erscheinungen findet man historisch in den Wörtern mit dem 4. Silbenton vor, der sich in bestimmten Fällen aus dem 2. Silbenton entwickelt hat, z.B. jugisch xa:m^4"Großmutter" < xɛ?^{2} "groß" + am^1 (< ama) "Mutter"; jugisch xɛ:s^4, ketisch qɛs'^{4} "Vorgesetzter", "Befehlshaber" < jugisch xɛ?^{2}, ketisch qɛ?^{2} "groß" + si prädikatives Suffix.

Im Wort daR^4 "Adlerin" (die Benennung des mythischen Vogels aus dem Mythus über den Zerstörer der Adlerhorste/6, S. 134/) könnte man eine uralte Zusammensetzung vermuten, deren erster Teil sich sehr leicht auf di?^{2} "Adler" zurückführen läßt.Der Vokalwechsel i : a, der sich dabei verfolgen läßt, ist in solchen Fällen in den lenissejschen Sprachen üblich. Die kritischen Bemerkungen von J.A. Krelnowitsch zur Übersetzung des Wortes daR als "adlerin" /7, S. 109-110/ sind völlig insolvent.

In mehrsilbigen Wortformen wird der 2. Silbenton in der Regel versimpelt, und dadurch entsteht auch ein Vokalwechsel wie z.B.in: i?n^2 "Nadel" -

pl. ɛnaŋ, i?s'^{2} "Spindel" - pl. ɛs'aŋ, hu?n^{2}, pl. hunaŋ (vgl. jugisch fu?n^{2}, pl. fɔnyŋ) "Tochter".

Wie an den angeführten Beispielen zu sehen ist, kann der vokalische Lautwechsel im Ketischen formbildenden oder wortbildenden Charakters sein. Der formbildende Vokalwechsel tritt entweder als das einzige oder als ein begleitendes (zusätzliches) Differenzierungsmittel auf dem Gebiet der Morphologie auf. Somit wären die vokalischen Alternationen im Ketischen als morphonologischer Lautwechsel zu bezeichnen; da sie aber meistenteils mit der Akzentuation verbunden sind, so ist auch der entsprechende Wechsel von Akzenten als Wechsel morphonologischen Charakters zu betrachten. Eben diese Frage wäre weiterhin ausführlicher zu erörtern.

Was die Paradigmatik des Nomens anbetrifft, so sind die Akzente (resp. Silbentöne) und die vokalischen Alternationen vor allem am Ausdruck der Kategorie der Zahl beteiligt. Dabei lassen sich folgende Fälle des Wechsels feststellen /8, S.22/:

1) der 1. Silbenton wechselt mit dem zweiten: ɛ?j^{2} - pl. eŋ^{1} "Ei", ki^{1} - pl. ki?ŋ^{2} "Fangfalle", ku?t^{2} - pl. ku^{1}reŋ "Gürtel", dɛ?^{2} -pl. deŋ^{1} "See";

2) der 1. Silbenton wechselt mit dem dritten: ba^{1} - pl. ba:n^{3} "Schlammläufer", boq^{1} - pl. bɔ:n^{3} "Handschuh", dap^{1} - pl. da:ŋ^{3} "Schulter", qop^{1} - pl. qɔ:^{3} "Wipfel";

3) der 1. Silbenton wechselt mit dem vierten: s'es'^{1} - pl. s'as'^{4} "Fluß", ej^{1} - pl. ɛj^{4} "Zunge", ɛr'^{4} - e^{1}r'eŋ "Schornstein";

4) der 2. Silbenton wechselt mit dem dritten: ɛ?p^{2} - pl. ɛ:ŋ^{3} "Schneespaten", kɔ?p^{2} - pl. kɔ:n^{3} "Erdeichhörnchen", sa?q^{2} - sa:n^{3} "Eichhorn", qɔ?^{2} - pl. qɔ:n^{3} "Waldzwiebel";

5) der 2. Silbenton wechselt mit dem vierten: sɛl'^{4}- pl, sɛ?n^{2} "Renntier";

6) der 3. Silbenton wechselt mit dem vierten: qaj^{4} - qi:n^{3} "Elentier".

Außer den 4 Silbentönen gibt es im Ketischen noch 2 kurzsilbige Akzentuationstypen, die in mehrsilbigen Wörtern vorkommen /9, S. 380-384/. Auch diese Akzentuationstypen*) (sie sind hier mit * und ** vor den entsprechenden Wörtern bezeichnet) wechseln in den Formen des Singulars und des Plurals:

1) die Singularform hat den 1. Silbenton, die Pluralform hat einen von den zwei kurzsilbigen Akzentuationstypen, z.B. baY^{1} - pl. *bakŋ "Klotz", dit^{1} - pl. *dɛkn "Auerhahn", am^{1} - pl. **amaŋ "Mutter", des'^{1} pl. **des'taŋ "Auge";

2) die Singularform hat den 2. Silbenton, die Pluralform hat den ersten kurzsilbigen Akzentuationstyp: dɔ?n'^{2} - pl. *dɔn'aŋ "Messer", bɛ?s'^{2} - pl: *bɛs'n "Hase", ki?s'^{2} - pl. *kis'en "Bein";

3) die Singularform hat den 3. Silbenton, die Pluralform hat den zweiten kurzsilbigen Akzentuationstyp: dɔ:l'^{3} - pl.**dol'eŋ "Unterfutter", ba:m^{3} - pl.**bamaŋ "Großmutter", ba:t^{3} - pl. **batan "Großvater";

4) der 4. Silbenton in der Singularform wechselt mit einem der kurzsilbigen Akzentuationstypen in der Pluralform: aj^{4} - pl. **ajeŋ "Sack", as'^{4} - pl. *as'eŋ "Feder", ɛr'^{4} - pl.*ɛtn "Zobel";

5) in der Singular- und Pluralform wechseln mit einander die kurzsilbigen Akzentuationstypen, z.B. *qoqpun' - pl. **qoqpun' "Kuckuck",**qol'et - pl. *qol'eraŋ "Wange"

Außerdem könnte man Fälle angeben, wo in der Singular- und Pluralform ein und derselbe Akzentuationstyp vorkommt: *bis'l' - pl. *bis'l'aŋ "Flosse", bit^{1} - pl bi^{1}kn "Taucher", qɔ?^{2} - pl. qɔ?ŋ^{2} "Horn".

Es ist leicht zu bemerken, daß in den angeführten Beispielen die vokalischen Alternationen im Vergleich zu der Akzentuation höchstens eine Nebenerscheinung sind. Unabhängig von der Akzentuation sind nur wenige Fälle der vokalischen Alternation(historischer Lautwechsel): i?^{2} - pl. ɛ?ŋ^{2} "Aufbewahrungszelt", di?^{2} - pl. da?n^{2} "Baumstamm", ty?s'^{2} - tʌ?ŋ^{2} "Stein".

Der Vokal- und Akzentwechsel tritt im Ketischen auch beim Ausdruck der Kategorien des Geschlechts und des Kasus auf. In den Kasusformanten drückt der a-Laut im Singular das männliche Geschlecht und der i-Laut das weibliche Gesclecht und in der Regel auch das sächliche Geschlecht aus, vgl. qɔjdaŋ "zum Bären", qɔjdiŋ "zur Bärin"; qytdanal "vom Wolf", qytdinal "von der Wölfin" usw.

Im Deklinationsparadigma verändern sich nur der

[*] Bei dem ersten kurzsilbigen Akzentuationstyp ist die Tonhöhe der zweiten Silbe tiefer, als die der ersten; bei dem zweiten Akzentuationstyp ist es umgekehrt, z.B. *as'el' "Schneeschuh", **as'el' "bedecktes großes Boot". Hier muß also immer die Tonbewegung in 2 Silben berücksichtigt werden.

zweite, dritte und vierte Silbentöne in den indirekten Kasusformen, vgl. di?^{2} "Adler", *diRan' "ohne Adler",*diRas' "mit dem Adler", *didata "für den Adler", *didaŋal' "vom Adler",*didaŋ "zum Adler".

Bei den ketischen Verbalformen findet man auch sehr häufig die vokalischen Alternationen und den Akzentwechsel vor. In den Nominalformen des Verbs lassen sich genau dieselben Fälle verfolgen,wie beim Nomen, z.B. bɛr'^{4} -*bɛtn "schlagen", dʌ?k - *dʌ?kŋ "fallen". Und genauso wie sich im Deklinationsparadigma die Akzentuation der Wortformen verändert, so verändert sie sich auch im verbalen Paradigma, vgl. i?l'^{2} "singen", aber *il'di "ich singe/ich kann singen"; sa:l^{3} "übernachten",*baYissal "ich übernachte"; dɔ?k^{2} "fliegen", *dirɔk "ich fliege" usw. Diese Veränderungen in der Akzentuation können vom Vokalwechsel i/a,e/i,y/a,u/o, ʌ/a u.a. beglitten werden /10, S,43/: kʌ?j "gehen", *tajga< tajkʌ?j "ich gehe"; bej^{1} "Wind", *il'vij "es blies der Wind"; qin^{1} "Strömung", *biŋsaRan biŋsaqin "es fließt" usw.

Das wären nur die allgemeinen Bemerkungen zum Vokal- und Akzentwechsel in der Konjugation, aber jedes Paradigma hat noch seine spezifischen Besonderheiten, wie es die Beschreibung des jugischen Verbalsystems gezeigt hat /11/.

Unabhängig von der heutigen Akzentuation sind im Ketischen Verbalsystem folgende Fälle der Alternation: (a) Vokalwechsel in den Formanten der Zeit, z.B. kavaRat "es läßt sich abschaben" - kovl'aRat "es ließ sich abschaben", kavatij "es hört auf" - komnatij "es hörte auf"; (B) Vokalwechsel in den Personalaffixen. Im letzteren Fall lassen sich mehrere Reihen von Affixen verfolgen, die von K.Bouda /12, S.98/ als Affixe der Klasse D und der Klasse B bezeichnet wurden, z.B. dilokŋ "ich zittere",kulokŋ "du zitterst"; boks'ivij "es treibt mich der Wind", kuks'ivij "es treibt dich der Wind" usw.

Auf Grund deranalysierten Beispiele läßt sich also behaupten, daß auch der Wechsel von verschiedenen Akzenten(Silbentönen) morphonologischen Charakters sein kann, und daß den vokalischen Alternationen im Ketischen die Unterschiede in der Akzentuation zugrunde liegen. Nur in manchen Fällen ist heute der Vokalwechsel von der Akzentuation unabhängig.

LITERATUR

1. SIEVERS,E. Steigton und Fallton im Ahd.- Aufsätze zur Sprach- und Literaturgeschichte (Festschrift Braune), Dortmund,1920.
2. KACNEL'SON S.D. Sravnitel'naja akcentologija germanskich jasykov. Moskva-Leningrad, 1966.
3. IVANOV V.V. O proischoždenii laringalisacii/faringalisacii v jenissejskich jasykach. - Fonetika. Fonologija. Grammatika. Moskva, 1971.
4. VERNER G.K. Ketskaja akcentologija.Avtoreferat dokt. diss. Leningrad, 1974.
5. VERNER G.K. Vsaimodejstvije tonal'noj i fonemnoj sistem v sovremennych jenissejskich jasykach. - Issledovanija v oblasti sravnitel'noj akcentologii indoevropejskich jasykov.Leningrad, 1979.
6. IVANOV V.V. Ketsko-amerindejskije svjasi v oblasti mifologii. - Ketskij sbornik. Antropologija etnografija, mifologija, lingvistika. Leningrad, 1982.
7. KREJNOVIČ J.A. Analis odnoj ketskoj legendy o ptice daR. - Formal'nyj analis jasykovych jedinic. Moskva, 1983.
8. VERNER G.K. Akcentuacionnaja charakteristika grammatičeskich paradigm imeni i mestoimenij v sovremennych jenissejskich dialektach.- Voprosy tjurkskoj filolcgii.Kemerovo, 1973.
9. WERNER,H.K. Die Akzentuation der mehrsilbigen Wörter in den gegenwärtigen Jenissej-Dialekten. - ZPSK, Bd. 27, Hf. 5, 1974.
10. KREJNOVIČ J.A. Glagol ketskogo jasyka. Leningrad, 1968.
11. VERNER G.K. Akcentuacija glagol'nych paradigm v jasyke symskich ketov (jugov).- Jasyki i toponimija, vyp. 7, Tomsk, 1980.
12. BOUDA,K. Die Sprache der Jenissejer. Genealogische und morphologische Untersuchungen. - Anthropos, vol. LII, 1957, N 1-2.

PHARYNGEAL SPLIT IN SYRIAC

SOLOMON SARA

Georgetown University
School of Languages & Linguistics
Washington, D.C. 20057 U.S.A.

ABSTRACT

The fact that Modern Chaldean is related to Classical Syriac is a chance to examine the changes that have occurred in the pharyngeals in the process of change. The analysis shows that they do not behave as a class; rather, the voiceless [H] changes to velar [xy], while the voiced [9] changes to a laryngeal stop [?]. The changes that occur in all the positions of the triliteral roots are not dependent on vocalic, but on root consonantal contexts.

Linguists assume that living languages change with time, and that daughter language inherits a portion of the lexicon from the mother language. Changes will be uneven; some will be more marked in certain areas of the lexicon than in others, and that changes will be systematic in going from the old to the new phonological system. The fact that Chaldean is related to classical Syriac is an opportunity to examine some of the phonological changes that have taken place in the process of going from one to the other. This paper will concentrate on the changes in the two pharyngeals [H,9].

The term "classical Syriac" refers to the language that has been in use since the third century A.D. Syriac is related to ancient Aramaic and is considered a later form of it. Syriac is still in use today in liturgical functions in many Christian communities of the Middle East. Chaldean refers to a modern dialect of Syriac that is currently spoken in parts of Iraq. Sara (1974).

The procedure followed in the study was to take the dictionary of Jacob Manna (1975) as the source of lexical items. All the lexical items that contained pharyngeals in this lexicon were isolated, then a list was drawn up of all the lexemes that have come into Chaldean, and that share the same semantic references with Syriac. A transcription was made of both lists. The transcription of Syriac depends primarily on the orthography given in the lexicon, while that for Chaldean is based on native speaker pronunciation. The study concentrated mainly on the triliteral roots in both languages, and only marginally includes their derivational or non-triliteral forms.

THE PHARYNGEAL [H]

[H] is a voiceless pharyngeal fricative. It occurs in initial, medial and final positions in words,i.e. as first, second and third radical, and in clusters. Since one is aware of the differences in the pharyngeal occurrences in both languages, the point of interest is: Where does Chaldean differ from Syriac in the shared lexical items, and are the differences haphazard or rule governed? The proper method is to isolate all the contexts in which pharyngeals occur in the shared items of the two languages, and to determine whether the changes that occur are contextually determined or not. A look at a list of shared items where [H] occurs initially in words is given below in parallel columns. The two columns highlight the differences between the two languages:

SYRIAC	CHALDEAN	
[Hpr]	[xpr]	'dig
[Hwr]	[xwr]	'white`
[HwH]	[xwx]	'peach`
[H y]	[x y]	'life`
[Hyt̲]	[xyt̲]	'sew`
[Hyp]	[xyp]	'bathe`
[Hyr]	[xyr]	'look`
[Hyl]	[xyl]	'power`
[Hmr]	[xmr]	'ass`
[Hmm]	[xmm]	'hot`
[Hmt]	[xmt]	'anger`
[Hmŝ]	[xmŝ]	'five`
[Hmr]	[xmr]	'bead`
[Hmr]	[xmr]	'yeast`
[Hms̲]	[xms̲]	'ferment`
[Hm9]	[xm?]	'leaven`
[H t]	[x t]	'sister`
[Htn]	[xtn]	'groom`
[Htm]	[xtm]	'conclude`
[Ht̲r]	[xt̲r]	'strike`
[Ht̲p]	[xt̲f]	'snatch`
[Ht̲y]	[xt̲y]	'sin`
[Hd̄]	[xd̄]	'one`
[Hdt]	[xt]	'new`
[Hdr]	[xdr]	'turn`
[Hdy]	[xdy]	'rejoice`
[Hsr]	[xsr]	'lose`
[Hs]	[xs.]	'lettuce`
[Hsl]	[xsl]	'wean`
[Hs̲d]	[xs̲d]	'harvest`
[Hs̲y]	[xs̲y]	'spay`
[Hl̄]	[xl̄]	'vinegar`
[Hlp]	[xlp]	'exchange`
[Hlm]	[xlm]	'thick`
[Hlm]	[xlm]	'dream`
[Hly]	[xly]	'sweet`
[Hlt̲]	[xlt̲]	'mix`
[HrH̄r]	[xrx̄r]	'snore`
[Hrp]	[xrp]	'sharp`
[Hry]	[xry]	'defecate`
[Hrb]	[xrw]	'spoil`
[Hrz]	[xrz]	'string`
[Hŝl]	[xŝl]	'strike`
[Hŝk]	[xŝk]	'darken`
[HŝH]	[xŝx]	'suitable`
[Hŝb]	[xŝw]	'think`
[HH]	[xx]	'plum`

****

[Hbb]	[Hbb]	'love`
[Hnq]	[Hnq]	'choke`
[Hnn]	[Hnn]	'kind`
[Hnp]	[Hnp]	'pagan`
[Hzq]	[Hzq]	'tighten`
[Hzm]	[Hzm]	'tie`
[Hzrn]	[Hzrn]	'june`
[Hkm]	[Hkm]	'govern`
[Hql]	[Hql]	'field`
[Hqq]	[Hqq]	'true`
[Hŝŝ]	[Hŝŝ]	`passion`

In the items listed above, and separated by *** one notices that [H] changes to [x] in some contexts but not in others. Listing the contexts gives an interesting pattern of change and stability in initial position:

[H] = [H]	[H] > [x]
[-b]	[-p]
[-n]	[-m]
[-z]	[-s,ŝ,t,d,t̲,s̲,l,r]
[-k,q]	[-w,y]

The above pattern of occurrence is consistent. Only exception to the above distributional occurrences of [H] in Chaldean was noticed. i.e: [Hrŝ] 'magic`.

The occurrences of [H] in medial position fare in a similar manner. The following list of shared lexical items illustrates this point:

[lHm]	[lxm]	'bread`
[tHm]	[txm]	'boundry`
[nHt]	[nxt]	'descend`
[ŝHn]	[ŝxn]	'warm`
[t̲Hn]	[t̲xn]	'grind`
[?Hn]	[?xn]	'brother`
[lHd]	[lxd]	'alone`
[ŝHt]	[ŝxt]	'dirt`
[kHl]	[kxl]	'mascara`
[ŝHlp]	[ŝxlp]	'change`
[rHŝ]	[rxŝ]	'walk`
[gHk]	[kxk]	'laugh`
[mlHb]	[mlxw]	'pitch fork`
[s̲Hy]	[s̲xy]	'swim`

****

[rHq]	[rHq]	'far`
[t̲Hl]	[t̲Hl]	'spleen`

Though the items in the above list are less numerous than the ones where [H] occurs initially, they are, nontheless, informative on the relevance of the environments in which [H] is retained in the Chaldean lexemes. One notices that the number of environments in which the original [H] is retained has shrunk and, the [-n,-k] environements are no longer effective in retaining the [H].

[H] = [H]	[H] > [x]
[-b]	[-p]
[-z]	[-m,n]
[-q]	[-s,ŝ,t,d,s̲,l;r]
	[-k]
	[-w,y]

It appears then that there is a gradation in the strength of the environments, i.e: the initial position in the root is the strongest and gives the maximum number of contexts in which [H] is retained, while the other environments will lose some of their conditioning power.

If the tendency to reduce the potency of the environment in retaining the Syriac [H] as we move towards the final radical of the word is valid, then there should be no restriction on the change of all the final [H]s to [x]s. Since by definition only the root consonantal patterns are operative in the change, there are no other root consonants occuring after the final radical. The following representative list of lexical items with the final [H] illustrates the change:

[?H]	[?xn]	'brother`
[pH]	[fx]	'trap`
[npH]	[npx]	'blow`
[rtH]	[rtx]	'boil`
[ptH]	[ptx]	'open`
[ps̲H]	[ps̲x]	'happy`
[ŝt̲H]	[ŝt̲x]	'spread`
[tl̄H]	[tl̄x]	'demolish`
[ŝlH]	[ŝlx]	'take-off`
[plH]	[plx]	'work`
[mlH]	[mlx]	'salt`
[ryH]	[ryx]	'smell`
[s̲rH]	[s̲rx]	'shout`
[b̄rH]	[b̄rx]	'lamb`
[?rH]	[?rx]	'road`
[mŝH]	[mŝz]	'smear`
[pŝH]	[pŝx]	'warp`
[lkH]	[lkx]	'lick`
[pqH]	[pqx]	'blossom`
[mwH]	[mwx]	'brain`
[ŝmH]	[ŝmx]	'trampled`
[nnH]	[nnx]	'mint`
[pwH]	[pwx]	'wind`
[t̲lwH]	[t̲lwx]	'pea`
[pyH]	[pyx]	'cool`
[nyH]	[nyx]	'rest`
[myH]	[myx]	'smell`

****

[gnH]	[gnH]	'blasphemy`
[ŝlh]	[ŝlH]	'apostle`
[mŝH]	[mŝH]	'messiah`
[zyH]	[zyH]	'psalmody`
[mdnH]	[mdnH]	'East`
[mdbH]	[mdbH]	'altar`

The above list of items indicates that there is no segmental that restricts the change of Syriac [H] to Chaldean [x] in final position. The items below ***** retain the final [H], but they are all liturgical terms that have been kept in their original place for special purposes.

From the above study one can see the tendency of syriac [H] to change to [x] in Chaldean. There are some environments in which it is kept. These environments are more operative when [H] is the first radical of the stem than when it is medial. In the final position there are no phonological restrictions but special uses of certain words have kept the original word in tact.

THE PHARYNGEAL [9]

[9] is a voiced pharyngeal fricative. The voiceless pharyngeal fricative [H] changed to another fricative, i.e: [x]. They differ only in their points of articulation. [H] changes from the pharyngeal to the velar [x] articulation. The articulation of [9], however, does not change its point of articulation to a corresponding velar fricative, but moves in a different direction. It changes more than one feature of its articulation as the following sets of data indicate.

[9] in initial position,i.e. as the first radical:

[9pr]	[?pr]	'dust, dirt`
[9ps̲]	[?ps̲]	'gall nut`
[9bb̄]	[?bb̄]	'lap`
[9bd]	[?wd]	'do,make`
[9m]	[?m]	'with`
[9mq]	[?mq]	'depth`
[9nz]	[?zz]	'goat`
[9nb]	[?nw]	'grape`
[9zqt]	[?sqt]	'finger ring`
[9zl]	[?zl]	'weave`
[9sr]	[?sr]	'ten`
[9s̲r]	[9s̲r]	'squeeze`
[9s̲y]	[9s̲y]	'rebel,mutiny`
[9d̄]	[?d̄]	'feast`
[9t̲m]	[?t̲m]	'thigh`
[9rbl]	[?rbl]	'sieve`
[9rq]	[?rq]	'run`
[9qrb]	[?qrw]	'scorpion`
[9qbr]	[?qbr]	'mouse`
[9qt]	[?qt]	'narrow,tight`

****

[9wl]	[9wl]	'moral evil`
[9md]	[9md]	'baptize`
[9lm]	[9lm]	'world`

The above list indicates that the Syriac [9] changes consistently into a glottal stop [?] in the corresponding Chaldean items in initial position. The list of exceptions in which [9] occurs initially is very limited and the items tend to be associated with matters liturgical, e.g. items below the *** in the above list.

[9] in medial position, i.e. as the second radical.

[t̲9n]	[t̲?n]	'to carry`
[t̲9m]	[t̲m?]	'taste`
[b̄9t]	[b̄?t]	'egg`
[ŝ9t]	[ŝ?t]	'yellow`
[l9s]	[l?s]	'chew`
[t9l]	[t?l]	`fox`
[ŝ9l]	[ŝ?l]	'cough`
[r9l]	[r?l]	'shiver`
[z9r]	[z?r]	'small`
[q9r]	[q?r]	'unlodge`
[s9r]	[s?r]	'sexton`
[d9k]	[d?k]	'knead`
[d9k]	[d?x]	'extinguish`
[b9y]	[b?y]	'wish,want`
[r9y]	[r?y]	'graze`

****

[ŝ9nn]	[ŝ9nn]	'hossana`
[z9prn]	[z9prn]	'saffron`

[9] in mid position in Syriac changes to a glottal stop in the corresponding Chaldean items. The few exceptions fit into the liturgical context or are abvious borrowings.

[9] in final position i.e. as the last radical.

[dm9]	[dm?]	'tear`
[nb9]	[nb?]	'spring`
[ŝb9]	[sw?]	'satisfy`
[ŝm9]	[ŝm?]	'hear`
[zd9]	[zd?]	'fear`
[?d9]	[?d?]	'know`
[qt̂9]	[qt̂?]	'sever`
[ŝl9]	[ŝl?]	'uproot`
[bl9]	[bl?]	'swallow`
[qr9]	[qr?]	'squash`
[mr9]	[mr?]	'pain`
[zr9]	[zr?]	'plant`
[dr9]	[dr?]	'arm`
[ŝy9]	[ŝy?]	'paint,smear`
[my9]	[my?]	'melt`
[tŝ9]	[tŝ?]	'nine`
[rq9]	[rq?]	'patch`
[?ŝ9]	[?ŝ]	'jesus`

****

[?ŝ9]	[?ŝ9]	'Jesus`
[rŝ9]	[rŝ9]	'blasphemy`

In the above list of items, the finally occuring [9] in Syriac almost invariably changes to a glottal stop in Chaldean. The rare forms in which [9] is retained are liturgically oriented lexemes that have been kept in their original forms, and are used in these restricted contexts.

The tendency of the two pharyngeals to change over time is evident from the above data. What is of interest is that their change is not parallel. They do not change to their corresponding fricatives in another class. Rather each changes into a separate segment, in a separate subclass of sounds. If there are contexts that have retained some of the Syriac [H] in certain positions,e.g. as the first or second radicals of the roots, none seem to be evident in the case of the pharyngeal [9], as far as the assembled data indicate. The change in [9] has been more radical than the change in [H].

This paper has concentrated on the consonantal environment of the pharyngeals as a basis for their retention or deletion. The question arises as to whether the vowels are operative in this process of change? From the data studied so far, there is no evidence that the vowels have been a factor of change in these cases. The changes are maintained in the derived forms irrespective of affixation or vowel variations in the derived forms of the same root. e.g.

[Hsr]	[xsr]	'lose'

[xasrin]	'I lose`
[xaaasɨr]	'He loses'`
[xsɨrri]	'I lost`
[maxsoorɨ]	'make lose`

[9rq]	[?rq]	'run`

[?arqin]	'I run`
[?aarɨq]	'He runs`
[?ɨrɨqli]	'I ran`
[ma?rooqɨ]	'make run`

The above sketch of the pharyngeal occurrences in Syriac and Chaldean shows the tendecy of these sounds to change to other sounds. The conclusions is that this class of sounds does not change to the corresponding sounds of another class, but the class members change into sounds of different classes.

The break down on the number of items that were borrowed from Syriac into Chaldean is shown below.

HCC	79/230	34%	9CC	30/155	19%
CHC	19/117	16%	C9C	21/115	18%
CCH	30/114	26%	CC9	20/103	19%

This indicates that the number of lexical items inherited from Syriac into Chaldean is roughly speaking about 22% of the items that contain one of the pharyngeals.

SOURCES

Jacob Manna. 1900/1975. CHALDEAN-ARABIC DICTIONARY. reprinted with a new appendix by Raphael J. Bidawid. (Beirut: Babel Center Publications) pp. 28 & 986.

Solomon I. Sara. 1974. A DESCRIPTION OF MODERN CHALDEAN. (The Hague: Mouton Publishers).

[Legend: H= , 9= , ŝ= ,s̲= , t̲=]

ARGUMENTS EN FAVEUR D'UN TRAITEMENT NON UNITAIRE DE LA TYPOLOGIE SYLLABIQUE

JAAP J.SPA

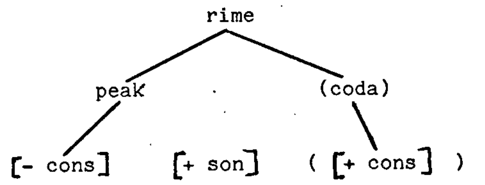
DÉPARTEMENT FRANCO-ROUMAIN
UNIVERSITÉ D'AMSTERDAM

ABSTRACT
Des arguments seront fournis pour étayer l'hypothèse que les schèmes syllabiques ne sont pas identiques pour tous les noyaux. Notammant les syllabes à schwa ou à consonne syllabique ont des structures qui dévient de celles qu'il faut admettre pour les syllabes à voyelle pleine.

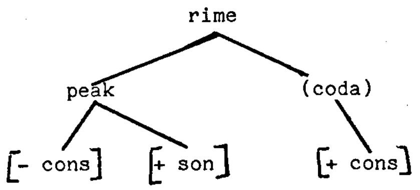
INTRODUCTION

En néerlandais il existe trois catégories principales de voyelles:
I les voyelles longues et les diphtongues
II les voyelles brèves
III le schwa.

Ce nonobstant, Trommelen a proposé de ramener les rimes de toutes les syllabes néerlandaises possibles à un seul type (1):

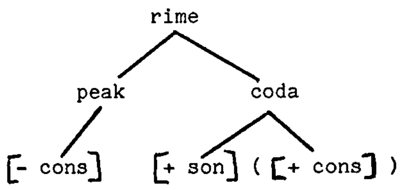


La coda aussi bien que le segment [+ cons] sont facultatifs. Le segment [+ son] est flottant dans la mesure où il peut se rattacher tantôt à la coda, tantôt au noyau (= peak). Soit: (pour les voyelles citées sous I)



z	e	e	m	[ze:m] (chamois)
d	u	u	w	[dy:w] (pousse)
p	a	u	s	[pɔws] (Pape)
z	e	e	Ø	[ze:] (mer.)

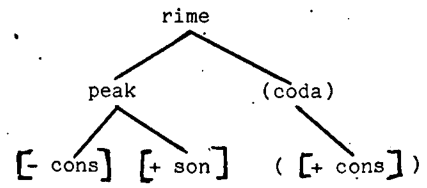
(pour les voyelles citées sous II)



k	e	r	k	[kɛrk] (église)
p	a	l	m	[pɑlm] (palmier)
r	a	m	Ø	[rɑm] (bélier)
k	i	Ø	p	[kIp] (poule)
sch	u	Ø	b	[sxʌp] (écaille)

On constate que dans les deux derniers cas le segment [+ son] obligatoire domine un élément Ø. Cet état de choses est justifié dans (1).

(pour le schwa)



lep	Ø	ə	l	[le:pəl] (cuillère)
bez	Ø	ə	m	[be:zəm] (balai)
monn	Ø	ə	k	[mɔnək] (moine)
tub	Ø	ə	Ø	[tybə] (tube)
vet	Ø	ə	Ø	[ve:tə] (zizanie)

Tout comme les éléments présentés sous I le schwa est considéré comme une unité bimorique par Trommelen (1). Cela se justifie par les considérations suivantes:

A. les voyelles longues, les diphtongues et le schwa peuvent se trouver en syllabe ouverte, sans coda (voir ci-dessus)
B. les voyelles longues, les diphtongues et le schwa ne peuvent être suivis du groupe: consonne sonante + consonne non coronale.

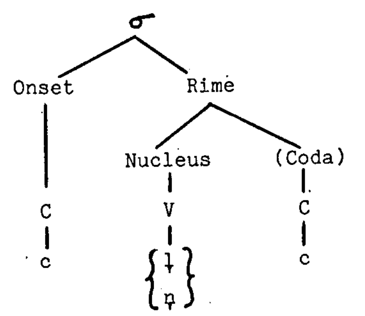
OBJECTION

Il est cependant possible de trouver un certain nombre d'éventualités où le schwa ne présente pas, en néerlandais, la même distribution que les voyelles longues et les diphtongues. C'est le cas notammant pour le groupe obstruante + liquide qui ne peut être l'attaque d'une syllabe dont le noyau est un schwa, alors que ce groupe peut parfaitement bien être l'attaque d'une syllabe à voyelle pleine.

HYPOTHÈSES

Je voudrais par conséquent mettre en question l'énoncé "There are no phonotactic restrictions at all for the language which must involve onset and peak"(2). L'environnement prévocalique peut jouer un très grand rôle dans la distinction des types syllabiques. L'hypothèse à soutenir est donc que dans certaines langues certains types de syllabe sont structurellement différents d'autres types de syllabe. Quelque choses de semblable a déjà été signalé sous une forme légèrement différente: "If a language contains syllables with vocal nuclei of the form C^mVC^n and syllables with consonantal nuclei of the form $C^{m'}$Ç$C^{n'}$, then $m' \leqslant m$ and $n' \leqslant n$". (3)

Cet énoncé est vérifié par les faits suivants tirés de l'allemand: dans cette langue les seules consonnes syllabiques sont [l̩] et [n̩]. La structure de la syllabe où elles se trouvent est relativement simple:



Toutes les consonnes ne sont d'ailleurs pas admises dans une telle structure: si le noyau est [n̩], l'attaque ne comporte que des consonnes non sonantes, alors que si le noyau est [l̩], l'attaque peut en outre contenir des consonnes nasales. Si l'on considère par la suite les syllabes allemandes ayant des noyaux composés de voyelles, sauf schwa, on y trouve une distribution consonantique beaucoup plus riche. cf. [ʃtraʊs, pflɪçt, (gə)braxt, ʃtro:, ʃvarts, fli:(ən)] etc. Pour de tels cas on voudrait proposer une structure syllabique ayant au moins deux consonnes dans l'attaque et deux dans la coda.

Les syllabes allemandes comportant des consonnes syllabiques sont par ailleurs dérivées et doivent leur existence à l'effacement d'un schwa. Les syllabes allemandes présenteraient dès lors des caractéristiques qui les font se distinguer elles aussi des autres syllabes à noyau vocalique. Cf.:"Da keine dieser Kombinationen -- il s'agit de /bl,dl,gl,zl,ml, ŋl/ suivis de schwa; JJS -- im Lexem realisiert wird, kann man den Kombinationen mit unbetontem /ə/ im System einen besonderen Status geben" (4).

En français les syllabes à schwa sont à générer par les règles syntagmatiques suivantes

$$\sigma \rightarrow O\ R \qquad R \rightarrow N$$
$$O \rightarrow C_1^2 \qquad N \rightarrow V$$

Ces dernières sont à compléter par un filtre qui spécifie l'ordre des consonnes dans l'attaque et qui adopte la structure que voici:

$$*[\ldots X_a X_b \ldots]_{\text{Onset}}$$
$$(a \geqslant b;\ a+b = 2,3;\ a+b \geqslant 4)$$

Ce filtre stipule que deux consonnes consécutives X_a et X_b ne peuvent faire partie de l'attaque si leurs valeurs respectives sont telles qu'elles sont exprimées dans les formules se trouvant entre les parenthèses. Les valeurs sont attribuées aux consonnes par l'échelle de force suivante:

obstruantes	force 0
liquides	force 1
nasales	force 2
glides	force 4

Exemples: *pjə , *vwə , *rjə , *rdə , *stə ,
*tnə , *lnə .

Les syllabes françaises à voyelle pleine sont par contre spécifiables par les règles -différentes-:

$$\sigma \rightarrow (O)\ R \qquad R \rightarrow N\ (Cd)$$
$$O \rightarrow C_1^3 \qquad N \rightarrow V$$
$$Cd \rightarrow C$$

et par le filtre -différent-:

$$*\ [\ldots.\ X_a\ X_b \ldots]_{\text{Onset}}$$
$$(a \geqslant b;\ a+b = 2,3)$$

Les syllabes françaises à voyelle pleine se caractérisent en outre par la possibilité d'adjoindre à leur attaque des consonnes supplémentaires, soumises à des conditions extrêmement strictes. En comparant la syllabe possible contenant un schwa et celle contenant une voyelle pleine, on constate que l'hypothèse précitée se vérifie également pour le français. On peut signaler pour conclure que le modèle présenté s'inscrit dans le cadre de la phonologie lexicale: les syllabes seront assignées aux structures segmentales ayant déjà subi les règles lexicales. C'est ce qui explique par ex. qu'en français le schwa peut se trouver sans consonnes précédentes dans une structure sousjacente comme /luəre/ "louerai" .

RÉFÉRENCES

(1) M. Trommelen "The Syllable in Dutch", Foris Publication, Dordrecht etc. 1984.

(2) E. Selkirk, The Syllable, in: "The Structure of Phonological Representations" (H.v.d.Hulst & N. Smith, eds.),Foris Publications, Dordrecht etc., 1982, pp. 337-383.

(3) A.Bell, Syllabic Consonants, in: "Universals of Human Language, vol. 2, Phonology" (J.H. Greenberg, ed.), Stanford University Press, Stanford, 1978, pp. 153-201.

(4) M. Philipp "Phonologie des Deutschen", Verlag W. Kohlhammer, Stuttgart etc. 1974.

SYLLABLES AND CONSONANTS

DAVID MICHAELS

Department of Linguistics
University of Connecticut
341 Mansfield Rd., Rm. 230, U-145
Storrs, CT 06268, USA

ABSTRACT

Syllables are projections of vowels. The first projection licenses a consonant position (coda) which is optional. The second projection licenses a consonant position (onset) which is obligatory. Consonantal phenomena such as assimilation and deletion are the consequence of 'fitting' consonants into syllable structure positions licensed by vowel projections.

Korean

In Korean there is widespread deletion of consonants at morpheme boundaries where two consonants compete for a single syllable structure position. For example, consider the alternation of [s] and Ø in [kap] 'price', [kapto] (foc), and [kapsi] (subj). The focus ending is [to], the subject ending is [i], and the basic noun is /kaps/. Where the two final consonants (/p/ and /s/) compete for a single coda position word finally or before an onset consonant, then /s/ invariably deletes. Where a vowel follows the consonant cluster, both are retained. Similarly, there is an alternation between [l] and Ø in [samki] 'to steam', [salmini] 'as one steams'. Here in lexical /salm/, /l/ and /m/ compete for a single coda position before an onset consonant and /l/ deletes. Again, where the cluster is followed by a vowel both consonants are retained. In the case of /kaps/ it is the second consonant of the cluster that deletes, in the case of /salm/ it is the first. Both the fact that a consonant deletes in the particular context, and the particular consonant that deletes in a given cluster follow from syllable structure principles.

Let us assume an account of syllable structure where syllables are projections of vowels: that is, a vowel is the head of a syllable. The first vowel projection (V' or rhyme) can optionally license a consonant position (coda). The second vowel projection (V" or syllable) obligatorily licenses a consonant position (onset). Onset positions, which are obligatory, are filled first. Furthermore, a syllable has a single sonority peak. In this framework, the consonant cluster reduction phenomena in Korean follows from the fact that the onset and coda positions can be filled by no more than one consonant each. The consonant that wins the position where two compete for it is the consonant with the lowest sonority value. If we assume that the syllable structure principles outlined above are universal, then all that a Korean child must learn is that the consonant positions licensed by syllable structure in Korean are limited to single consonants. Which consonants delete, and under what circumstances, will follow from universal principles of syllabification and general considerations of sonority. The evidence that the child will need to arrive at the language particular property of Korean that only a single consonant can fill a syllable structure position is readily available from the fact that words end in no more than one consonant. The fact that it is the lowest sonority member of the pair is the unmarked case if syllables tend to maximize the single sonority curve. In this way a fairly restricted, but general theory of syllable structure and sonority, not only gives an account of cluster reduction in Korean, but suggests a way that the learnability question (how complex grammatical facts can be mastered so rapidly with the limited data available to the child) can be addressed in phonology.

There is one additional aspect of Korean consonant cluster reduction that I will consider here. The combination of /ilk+ta/ gives [ikta] 'to read' as expected. That is where /l/ and /k/ compete for the single coda position of the first syllable, the lower sonority /k/ wins. However, the combination of /ilk+ki/ gives [ilki]. Here it appears that /l/ has won the coda position and /k/ has been deleted in violation of the sonority principle. It seems that when two of the three consonants in the sequence are a geminate cluster, that degemination takes precedence in cluster reduction. This appears to be a strictly linear fact that has nothing to do with syllable structure positions. However, if consider that consonant clusters can also be analyzed as hierarchical structures and that deletion is a last resort to "fit" a sequence of consonants into positions licensed by syllable structure, then there is a syllable structure account of degemination. The analysis is as follows: consonant clusters are C-projections, where the first projection C dominates the head C and a complement C. The single consonant for each C-position in Korean syllable structure can then be seen as a requirement that C′ have a single interpretation. Where two different consonants are dominated by C′, then C′ takes the lowest sonority interpretation in the unmarked case. Where the two C′s are identical (i.e. geminate clusters), then C′ still gets a single interpretation since the two C s it dominates send up nondistinct feature representations. Thus, degemination is automatic where geminate clusters must be interpreted in a single syllable structure position, and the facts in question here do not contradict the syllable structure and sonority approach to cluster reduction in Korean [1].

Southern Paiute

Southern Paiute also has degemination, but under different circumstances from those in Korean. In Southern Paiute, geminate clusters are reduced before a stressed vowel. The interaction of stress and syllable structure in this case is quite deep. First, it is important to note that Southern Paiute has an alternating stress pattern where every even numbered vowel counting from the left is stressed (except for the last syllable of the word, or in bisyllabic words where the first syllable is stressed). The interaction of stress and syllable structure in Southern Paiute seems to be the following. Vowels project a syllabic category (V′ or rhyme) which optionally licenses a consonant position (coda). In Southern Paiute, whether or not a coda position is licensed depends on stress. Specifically, a stressed syllable can have a coda, an unstressed syllable cannot. Thus, the syllable before a stressed syllable is always unstressed under the alternating stress pattern and cannot license a consonant as coda. If two consonants intervene between an unstressed syllable and a stressed syllable then the both must be interpreted under the onset C′ of the stressed syllable. If they are nondistinct (i.e. a geminate cluster) and if C′ in Southern Paiute (as in Korean) must have a single interpretation, then degemination once again is the single interpretation of nondistinct feature representations at C′. In Korean, degemination is forced in onset position when the coda of the preceding syllable is filled by a consonant. In Southern Paiute it is forced when the preceding syllable is unstressed and, therefore, cannot license a coda position. In both cases cluster reduction follows automatically from the requirements of syllable structure. In Southern Paiute the learnability of degemination requires only that evidence of the alternating stress pattern in open and closed syllables be available. No particular rule of degemination need be learned.

Let us consider one other closely related aspect of Southern Paiute here: the other side of degemination, gemination. In Southern Paiute, as pointed out by Sapir [2], certain morphemes have the effect of geminating the initial consonant of following morphemes. The initial segment of the geminate cluster may be an obstruent or a nasal. Following Chomsky and Halle [3], let us assume that the geminating effect is the result of adjacency of the final consonant of the first morpheme and the initial consonant of the second morpheme. In the syllable structure framework under consideration, both consonants will be given an interpretation under C′ in the onset of the second syllable. If the first consonant is unspecified, at C′ it will be interpreted as nondistinct from the other, specified consonant. If the first consonant is specified only for nasality, at C′ it will be interpreted an nondistinct for all other features from the other, fully specified consonant. Finally, if the preceding syllable is stressed and, therefore, licenses a coda position, this geminated consonant will receive an interpretation there. This will be the case of gemination. If the preceding syllable is unstressed and, therefore, cannot license a coda position, then the geminated consonants will receive a single interpretation in the onset position of the second syllable. This is the case of degemination. Thus, assimilation (in this case complete assimilation or gemination) can be seen to follow from the requirement of giving adjacent consonants an interpretation at the immediately dominating projection licensed by syllable structure. See Chomsky and Halle′s [3] analysis of Sapir′s [2] observations on Southern Paiute for further discussion.

Zoque

There is a similar gemination and degemination process at morpheme boundaries in Zoque (see Wonderly [4] for the data, Dell [5] for a generative analysis of the data). In morpheme combinations such as /kihp+u/ ([kihpu]) ′he fought′, the /h/ appears to be realized as coda of the first syllable, the /p/ as onset of the second. When lexical /kihp/ comes together with /pa/ (past tense), the resulting geminate /p/-cluster degeminates giving [kihpa]. Thus, where the coda position of the preceding syllable is filled as in Korean, the geminate cluster must be interpreted as a nondistinct single consonant in the onset C′ position of the following syllable. However, where an /h/ is followed by two nonidentical consonants, then the /h/ deletes. For example, /kihp+yahu/ ′they fought′ is realized as [kipyahu]. This case, too, can be analyzed in a similar way to cluster reduction in Korean. Thus, /p/ and /y/ being nonidentical cannot both be interpreted in the onset C′ position of the second syllable. /p/, then, must compete with /h/ for the single coda position of the first syllable, where sonority determines which segment wins.

Zoque also shows evidence of assimilation in onset position similar to that of Southern Paiute. Here, a morpheme apparently specified only for nasality essentially geminates with the initial consonant of the stem it is prefixed to. Thus we get the following pair: [puhtu] ′he went out′, [mbuhtu] ′I am going out . If the stop consonants are unspecified for voicing and the nasal prefix is unspecified for place of articulation, the at C′ in onset position, the nasal is interpreted with the place of articulation features of the stop and the stop is interpreted with the voicing feature of the nasal. Once again, if syllabification requires the two consonants to be interpreted in a single onset position, the assimilation follows from the requirement that each C be interpreted at C′. See Michaels and Tiedeman [6] for further discussion of consonantal alternations in Zoque within a syllable structure framework.

Japanese

The last set of examples is from Japanese. A verb such as [tabe] ′eat′ is [taberu] in the present, [tabeta] in the past, which enables us to identify the present tense suffix as /ru/, the past tense suffix as /ta/. When these suffixes are added to consonant final verb stems, we get instances of degemination and gemination creating an interesting array of alternations. Thus, /wakar+ru/ ′understand′ gives [wakaru] (degemination), and /wakar+ta/ gives [wakatta] (gemination). In the case of /wakar+ru/, if we assume that the final syllable of the stem does not license a coda, then both /r/s must be interpreted in the C′ onset position of the suffix syllable giving degemination in circumstances similar to Southern Paiute. In the case of /wakar+ta/, the nonavailability of a coda position will force the interpretation of /r/ and /t/ in the C′ onset position of the suffix syllable. Here, however, because the two consonants are distinct, we would expect either deletion of one (presumably /r/ since it is more sonorous than /t/), or that /r/ is interpreted in the coda position of the preceding syllable. However, there are two things wrong with this analysis. First, if we′ve argued that degemination occurs in /wakar+ru/ because there is no coda position in the final syllable of the stem, then why does such a position show up to license the first member of the geminate cluster in [wakatta]. Second, if gemination is nondistinctness at C′ how is it that /r/ and /t/ can have nondistinct feature representations.

Assuming that the general syllable structure account of deletion, assimilation, gemination and degemination that we have been outlining is correct, then we must make several assumptions about language particular representations in Japanese to get the analysis to come out right there. First, we must assume that the final stem syllable in the verbs in question license a coda position that is partially specified as either a nasal or a voiceless stop (i.e. [nasal, voice]). That is, any segment that is interpretable in that position will be interpreted as a voiced nasal or a voiceless nonnasal. Second, we must assume that /r/ is a maximally unmarked liquid. Then in the case of /wakar+ta/, /r/ is interpreted as nondistinct from /t/ at C′ in the onset position of the suffix syllable. Once it has the interpretation of a voiceless stop it can be realized in the marked coda position of the final syllable of the stem. The fact that all stem final consonants are interpreted as either /n/ or /t/ in this position before the /ta/ suffix is consistent with the assumption that the position itself is partially marked for the features [nasal, voice] (e.g. /yom+ta/ [yonda], /yob+ta/ [yonda], /iw+ta/ [itta], etc.) See Kawai [7] for a detailed discussion of syllable structure and assimilation phenomena in Japanese.

In conclusion, it seems unlikely that similar phenomena in such disparate languages as those illustrated above are the result of accidental language particular rules. The syllable structure approach suggests why such phenomena might be expected to show up in again and again across languages. It is in the process of vowels gathering consonants up into syllables that adjustments to the consonants have to be made. The different adjustments made in different languages are the result of the particular configuration that syllables take in those languages. Once those configurations are identified, the range of adjustments available to consonants can be predicted to a large extent.

REFERENCES

[1] Hong, S. and D. Michaels, 'Syllable Structure and Sonority in Korean,' ms. University of Connecticut, 1985.
[2) Sapir, E., 'The Psychological Reality of Phonemes,' in D. G. Mandelbaum (ed.), Selected_Writings_of_Edward_Sapir, U. of Calif. Press, 1949.
[3] Chomsky, N. and M. Halle. The_Sound Pattern_of_English, Harper & Row, 1968.
[4] Wonderly, W. 'Zoque II: Phonemes and Morphemes,' IJAL 17.105-23.
[5] Dell, F., Generative_Phonology, Cambridge Univ. Press, 1980.
[6] Michaels, D. and R. Tiedeman, 'Rules and Syllable Structure in Zoque,' ms. University of Connecticut, 1986.
[7] Kawai, M., 'Syllable Structure and Assimilation in Japanese,' ms. University of Connecticut, 1986.

DIRECTIONAL, LEXICAL AND POSTLEXICAL SYLLABIFICATION AND VOWEL DELETION*

ROLAND NOSKE

Institute for General Linguistics
University of Amsterdam

ABSTRACT

In this paper it will be shown that the concept of syllabification, i.e. the assignment of syllable structure, can account for the at first sight disparate vowel deletion phenomena in a much discussed Amerindian language, viz. Tonkawa. More specifically, it will be shown that the specification of the **direction**, the **domain of application** and the **elements triggering the syllabification** can account for the data in question. The Tonkawa case thus provides a good illustration for the view that certain phonological processes involving syllable structure, like vowel deletion, epenthesis and semivocalization, are typically the result of the assignment of syllabic structure, and need not be stated as independent rules.

INTRODUCTION

Consider the following set of Tonkawa forms:

(1)

a. picno?	<	picena+o?	'he cuts it'
b. wepceno?	<	we+picena+o?	'he cuts them'
c. picnano?	<	picena+n+o?	'he is cutting it'
d. wepcenano?	<	we+picena+n+o?	'he is cutting them'
e. picen	<	picena	'steer castrated one'

The following affixes can be identified:

(2)

a. we-	3rd person plural, pronominal object
b. -o?	3rd person singular, declarative, present tense
c. -n-	progressive (continuative)
d. (unmarked)	3rd person singular, pronominal object

The following phonetic variants are exhibited by the stems:

(3) c. picn-, -pcen-, pecna-, -pcena-, picen,
/picena/ 'cut'

In order to account for these data, Kisseberth [4] posits the following rules:

(4) a. Word-Final Vowel Deletion
$V \rightarrow \emptyset \; / \; \underline{\quad} \, \#$

b. Vowel Elision
$V \rightarrow \emptyset \; / \; \#CVC\underline{\quad}C\begin{bmatrix} V \\ +stem \end{bmatrix}$

c. Vowel Truncation
$V \rightarrow \emptyset \; / \; \underline{\quad} \, V$

The derivations are given in (5):

(5)

	(a)	(b)	(c)
UR	picena+o?	we+picena+o?	picena+n+o?
Delete			
Elide	picna+o?	we+pcena+o?	picna+n+o?
Truncate	picn+o?	we+pcen+o?	
SR	picno?	wepceno?	picnano?

	(d)	(e)
UR	we+picena+n+o?	picena
Delete		picen
Elide	we+pcena+n+o?	
Truncate		
SR	wepcenano?	picen

The specification is [+stem] for the final vowel in the SD of Vowel Elision (4b) is needed in order to prevent Elision to take place in (6c):

(6)

a. pilo?	<	pile+o?	'he rolls it'
b. weplo?	<	we+pile+o?	'he rolls them'
c. pileno?	<	pile+n+o?	'he is rolling it'
d. wepleno?	<	we+pile+n+o?	'he is rolling them'

In (6c) the second vowel of the word does not elide, although it is in the environment CVC__CV. The final vowel in these forms does not belong to the stem, but to the suffix **-o?** (see (2b)).

Kisseberth adduces additional paradigms in order to show that the vowel that is to be deleted by Vowel Elision must belong to the stem:

(7)

a. yakpo?	<	yakapa+o?	'he hits it'
b. weykapo?	<	we+yakapa+o?	'he hits them'
c. wexaykapo?	<	we+xa+yakapa+o?	'he hits them with force'

In (7c) it is not the second vowel of the form that elides, (which is what rule (4b) would predict), but its third vowel, which is the first stem vowel. Therefore, Kisseberth restricts Vowel Elision further so that only a vowel that is specified as [+stem] is affected by the rule. He observes ([4]:117) that if there is a CV prefix, the first stem vowel deletes and that if there is no prefix, the second vowel of the stem deletes.

Kisseberth reformulates Vowel Elision as:

(8) Kisseberth's reformulation of vowel-elision

$$\begin{bmatrix} V \\ +stem \end{bmatrix} \dashrightarrow \emptyset \; / \; \left\{ \begin{array}{c} V{+}C \\ \left\{ \begin{array}{c} \# \\ C{+} \end{array} \right\} CVC \end{array} \right\} \_\_ \; C \begin{bmatrix} V \\ +stem \end{bmatrix} \quad \begin{array}{c} (a) \\ (b) \\ (c) \end{array}$$

Subrule (a) accounts for stems preceded by a CV prefix; subrule (b), for stems without a prefix and subrule (c) for stems preceded by a CVC prefix. The three subrules restrict elision to the context VC__CV.

The complexity of rule (8) does not satisfy Kisseberth and he therefore mentions the need for a simpler rule, combined with a derivational contraint.

AN ALTERNATIVE ANALYSIS

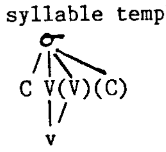
Tonkawa allows the following syllable types:

(9) possible Tonkawa syllables:

a. CV b. CVC c. CV: (CV_iV_i) d. CV:C

The syllable template in (10) expresses the possible forms a syllable can take. Note that the first C and V are the obligatory elements of the syllable (CV often being referred to as the core syllable).

(10) syllable template:



(σ =syllable)

We propose that the syllabification parameters are set as follows:

(11) Tonkawa syllabification:

a. first cyclically (lexically), exclusively in derived environments; then postcyclically (postlexically) (**cyclicity parameter**);

b. iteratively leftward (**directionality parameter**);

c. syllabification is triggered by unsyllabified C's (**obligatory incorporation parameter**);

d. measures taken when syllabification fails:
   i. the direction is reversed (left to right);
   ii. when this also fails: necessary elements on the left side are extracted from other syllables and incorporated into the newly formed syllable.

The **cyclicity parameter** given iŋ (11a) is a well-known parameter in phonology, especially with regard to the assigment of prosodic structure. This parameter also exists for the assignment of the lowest prosodic level: the syllable. A consequence of the particular setting of this parameter in Tonkawa is that existing syllable nodes are **not** erased by a subsequent cycle (an example of strict cyclicity; if syllable nodes could be erased by a syllabification applying at a later cycle, the results of the former syllabification would not be detectable). However, as we will see later, existing links between individual elements **can** be broken, if this is necessary for obtaining a permissible syllable structure.

The **directionality parameter** in (11b) and directionality in syllabification in general have been argued for repeatedly (see Kaye and Lowenstamm [3], ter Mors [5], Noske [6,7]). Here, we will see that the setting of this parameter (from right to left) allows us to explain which vowels are deleted.

The **obligatory incorporation parameter** in (11c) is crucial in our theory of syllabification. According to this theory, syllabification is triggered by the elements that **must** be incorporated into syllabic structure. Three situations are possible:

i. only C's are triggers
ii. only V's are triggers
iii. both C's and V's are triggers

The theory entails that the fourth logical possibility, i.e., neither C's nor V's are triggers of syllabification, does not occur. This is because in that situation, syllabification would not be triggered at all, neither C's nor V's would be linked to syllable nodes, and there would be no phonetic outcome.

In the case where only C's are triggers of syllabification, i.e., C's are the elements that must be syllabified, V's will be skipped at a stage of the syllabification process where this process can only incorporate a C. The rightmost V will be skipped if the syllabification is applying from left to right, the leftmost if syllabification applies in the opposite direction. (This latter case can be found in (14a,b) (below)). If in this type of language two contiguous **C's** are encountered by the syllabification mechanism, the mechanism will project a V in between them (e.g., in the environment CVC__CV). This V will be filled with the neutral vowel value (often a schwa).

The situation is symmetrically opposite if V's are the triggering elements. This may be the case in languages with consonant truncation phenomena. In this type of language, again assuming CV as the only permissible syllable type, in a CVCCV environment, one of the two contiguous C's will be ignored by the syllabification mechanism, the rightmost if the syllabification is applying from left to right, the leftmost if syllabification applies in the opposite direction. In the case of a CVVC environment, a C will be projected between the two contiguous V's, and will be filled with the neutral consonant value (often a glottal stop).

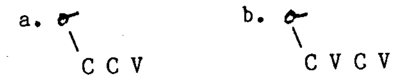
In case iii, in which both C's and V's are triggers, The syllabification mechanism will resort to projection of both C's and V's if it encounters disallowable sequences of elements. The mechanism will project a C in the environment CV__VCV and a V in the environment CVC__CV.

An interesting consequence of our theory is that it predicts that the reverse situation will not occur: there will be no language where both C's and V's can be skipped during syllabification, hence no C's as well V's will be deleted as a result of the syllabification process. This is precisely because of the fact that there should be at least one type of element, either C or V, that triggers syllabification.

The particular parameter setting for Tonkawa in (llc) has the consequence that V's may be skipped, but not C's.

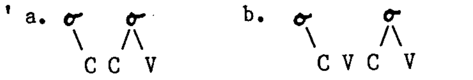
The **measures** taken if syllabification of an obligatory element (in the Tonkawa case: a consonant) fails boils down to two basic situations:

(12) case i



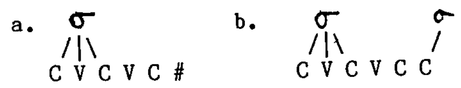
In this case, the rightmost C is not yet syllabified and therefore at a given cycle (or in the case of a C in a nonderived environment: postcyclically) triggers syllabification. Because of the direction parameter setting (right to left), syllabification will proceed leftward. With the material on the left side, however, no legitimate syllable can be formed: there is no free consonant that can function as the beginning of the syllable (note that a Tonkawa syllable always starts with a C). Therefore, the direction will be reversed and the V to the right will be incorporated into the syllable. The thus formed CV syllable is legitimate, and syllabification has succeeded:

(12)'



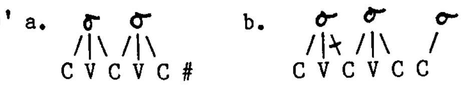
In (13), this latter strategy, reversal of the direction of syllabification, will also fail:

(13) case ii



In this situation, there is no V to the right of the unsyllabified C that can be incorporated. Now strategy ii takes effect: the C to the left is detached from the preceding syllable (note that the syllable to the left remains licit (CVC --> CV)), and this C is incorporated into the newly formed syllable:

(13)'

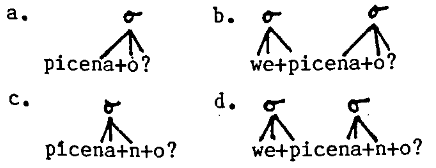


Having outlined our analysis of Tonkawa syllabification, we will provide illustrations for each of the type of cases mentioned in section 1. We will show that in these cases, the correct deletion is forcast by our syllabification algorithm. Let us now look at the cases (1a-d & 5a-d), repeated here as (14):

(14) a. picena+o? b. we+picena+o?
c. picena+n+o? d. we+picena+n+o?

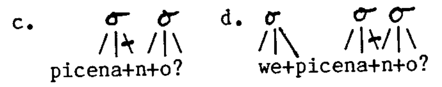
In these forms, during the first cycle the morphemes which are adjacent to the stem will be syllabified:

(14)'



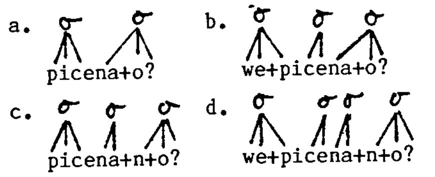
It is assumed that in these forms, the prefixes and suffixes adjacent to the stem are attached on the same cycle. However, this is not crucial. In (14'a,b), it is the glottal stop that triggers syllabification. syllabification, in accordance with the directionality setting (11b), now proceeds leftward, and the **o** is incorporated into the syllable. Next, **a** is ignored by the syllabification mechanism, because it cannot incorporate this in its syllabic structure (cf. template (10)). It is thus that the data for which Kisseberth has formulated his rule of vowel truncation (6b) are borne out. In (14b,d), the **w** of the prefix also triggers syllabification. Because there are no elements to the left, the direction of syllabification is reversed by virtue of (11d.i), cf. (12). Note also that on this first cycle, the morphemes of the second cycle are invisible. Let us now look at the second cycle:

(14)'' a. (vacuous) b. (vacuous)



Here, we see that in (14''c,d) the syllabification mechanism has delinked the **n** from the preceding syllable (by virtue of (11d.ii), cf. (18)). We now come to the postcyclic (postlexical) syllabification, in which unassociated consonants belonging to the stem trigger syllabification:

(14)'''

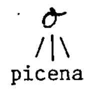


When postcyclic syllabification takes place in (14'''a,c), both **p** and **c** are still unsyllabified. Going from right to left (by virtue of (11a)), the syllabification algorithm creates a syllable incorporating the **c**, **i** and **p**. Now, all consonants are incorporated into the syllabic structure. Hence there is no need to incorporate the **e** and therefore the phonetic outcomes are **picno?** and **picnano?** respectively.

When postcyclic syllabification takes place in (14'''b,d), one stem consonant is still unsyllabified: **c**. Leftward syllabification fails, because there is no C to its left. Therefore, by virtue of (11d.i), the direction of syllabification is reversed, (cf. (12)). The syllabification mechanism will incorporate the **e** to the right of the **c** into the syllabic structure. The subsequent **n** is already syllabified and therefore will not be incorporated into the syllable just formed. The sylllable thus created (**ce**) has the form CV which, according to template (10), is a permissible structure. The **i** has been left unsyllabified and is hence not realised. Our model thus correctly predicts that **picnano?** and **wepcenano?** are the correct surface forms.

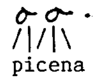
We still have to explain the vowel deletion in example (3e, 7e), for which Kisseberth has formulated his rule of Word Final Vowel Deletion (6a). Our model accounts for this deletion in a straightforward manner. Since there are no morphemes attached to the stems, there is only one, postcyclic, application of syllabification. Syllabification starts form the right in accordance with (11a):

(15)



Thus, the syllable **cen** is formed. Arriving at the **p**, no material can be found to the left of this element. Therefore, the direction is reversed by virtue of (11d.i) (cf. (12)):

(15)'

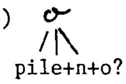


Now, all consonants, which are the syllabification triggering elements (cf. 11c), have been syllabified. The final **a** has been left unsyllabified, and hence is not realised, which is the correct prediction.

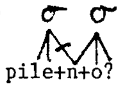
Let us now turn to the example given in (6c) where the second stem vowel is not elided.

first cycle: second cycle: postcyclic syll.:

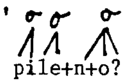
(16)



(16)'



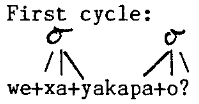
(16)''



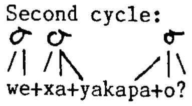
We thus see here that the nonerasure of the **e** (the vowel which finds itself in the environment VC__CV) is the consequence of the setting of the cyclicity parameter (11a): the **e** had to be incorporated into syllable structure at the first cycle, at which syllabification was triggered by **n** (the progressive morpheme). This example illustrates also the working of (11d.ii): during the postcyclic syllabification, the **n** was extracted from the previous syllable and was incorporated into the final syllable (an instantiation of (13a)).

Finally, let us look at the form in (7c). Here it was the third V and not the second one that was deleted.

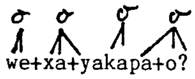
(17) First cycle:



(17)' Second cycle:



(17)'' Postcyclic syllabification:



Here, the **k** was the only consonant that had not yet been syllabified during the cycle. Leftward syllabification will fail, because the preceding consonant **y** is already incorporated into a syllable. Therefore, by virtue of (11d.i) (cf. (12), the direction of the application of syllabification is reversed and the following **a** is incorporated into the syllabic structure. The first stem vowel is left unsyllabified, because all consonants are already syllabified, and there is no need for further syllabification. It is thus correctly predicted that **wexaykapo?** is the phonetic outcome.

We have thus seen that the phenomena for which Kisseberth's word-final vowel deletion rule ((4a), the truncation rule (4c) as well as his vowel elision rule (8) were formulated are all correctly predicted in our syllabification model. Hence, there is no need for formulating separate rules.

CONCLUSION

In this paper, we have provided an explanation for the different phenomena of vowel deletion taking place in Tonkawa. We have shown that it is possible to account for them by analysing them as a result of the assignment of syllable structure. It has also been shown that, for Tonkawa at least, a lexical and a postlexical stage of syllabification must be assumed. Furthermore, the theoretical relevance of a newly proposed parameter the **obligatory incorporation** parameter has been outlined. Its importance for the account of the Tonkawa fact has subsequently been shown. Finally, it was demonstrated that the concept of directional syllabification can not only explain the correct epenthesis sites in certain languages (as shown in Noske [6,7]) for Yawelmani and Tigrinya), but also the correct sites for vowel deletion in certain other languages, like Tonkawa.

NOTE

* I wish to thank Martha Wright for a valuable discussion on the Tonkawa data and especially for pointing out that the assumption that the order in which syllabification applies in the different morphological domains is of crucial importance. This paper is part of work in progress, where more data of Tonkawa will be treated. The main source on Tonkawa is Hoijer [2].

REFERENCES

[1] van der Hulst H. & N.S.H. Smith (eds.), **Advances in Nonlinear Phonology**, Dordrecht: Foris, 1985.

[2] Hoijer, H., "Tonkawa - An Indian Language of Texas", in Boas, F. (ed.), **Handbook of American Indian Languages**, Part 3., 1933.

[3] Kaye, J.D. & J. Lowenstamm, "Syllable Structure and Markedness Theory", in Belletti, A. et al. (eds.), **Theory of Markedness in Generative Grammar**, Pisa: Scuola Normale Superiore di Pisa, 1982, p. 287-315.

[4] Kisseberth, C.W., "Vowel Elision in Tonkawa and Derivational Constraints", in Sadock, J.M. & A.L. Vanek, **Studies presented to Robert B. Lees by his students**, Edmonton, AB: Linguistic Research, 1970, p. 109-137.

[5] ter Mors, C., "Empty V-nodes in Klamath and their Role in Klamath Vowel Alternations", in [1], p. 313-334.

[6] Noske, R., "Syllabification and Syllable Changing Processes in Yawelmani", in [1], p. 335-361.

[7] Noske, R., "A Parameter of Syllabification", in **Papers and Studies in Contrastive Linguistics**, to appear.

ON THE RELATIONSHIP BETWEEN COARTICULATORY EFFECT OF LIP ROUNDING AND SYLLABIC BOUNDARY IN FRENCH

ANTONELLA GIANNINI

Ist. Universitario Orientale
Fonetica SPerimentale
Napoli, Italy

ABSTRACT

This study investigates the relationship between the extension of the coarticulatory effect of lip rounding and the syllabic boundary. The choice of lip rounding is due to the fact that it is particularly evident on the spectrogram, the formants of a labialized articulation being noticeably lower that those of its non-labialized counterpart. We have chosen French because it is a language having a vocalic system strongly affected by lip protrusion. The aim of this research is, on one hand, to verify whether among the rounded vowels different levels of labialization and, consequently, different extensions of coarticulation exist; on the other hand, to propose a model of syllabic organization for French and to compare it with those already obtained for English and Italian following the same method of analysis.

INTRODUCTION

Traditionally the term "coarticulation" has been used to indicate the phenomenon by which a speech sound is modified from its basic form because of the influence of a neighbouring sound.

The need to study this phenomenon ensued from the observation that all the speech elements always showed different characteristics according to the context. However such a way of facing the problem didn't give later on satisfactory results because to consider speech as a chain of discrete units that only marginally could be modified by the context was inadequate to the dynamic reality of speech. The point is that these supposed "discrete units", even if had gained an apparent objectiveness thanks to the larger and larger number of firstly physiological and then acoustical studies, did nothing but to propose in a new dress the classical distinction between vowels and consonants.

In the 60' a new way of dealing with the problem was suggested, leading consequently to a redefinition of the coarticulation.

In fact coarticulation, once seen as the effect of the mechano-inertial constraints of the speech apparatus, is now considered as a reflex of the speech organization in programming units. The starting point is given now by the "articulatory syllable" that is considered as a string of elements coproduced at the level of motor command. At the beginning of each programming unit the articulators receive all the information necessary to realize the whole articulatory syllable. In the light of these considerations, as both coarticulation and syllable are the result of a single motor command, the limits of extent of the former should not go beyond the limits of extent of the latter.

The problem, this time, is to specify the limits of extension of these units and in order to do this many researches have been carried out using different analysis techniques, that is electropalatography /1/, cinefluorography /2/ and electromyography /3/ /4/.

Once more the results are not univocal and different models have been proposed that individualize the articulatory syllable now in CV now in CVC now in VC..CnV. The disagreement persists also among studies employing the same technique. In the case of the electromyography, for instance, the different results can be due to the fact that there isn't a close relationship between muscular intensity and the position assumed by the articulators. This has been clearly pointed out by Lubker and Gay /5/ who, as regards lip rounding, say that the same muscular contraction can give rise to different labial movements.

In order to overcome this impasse, in this research we'll examine the coarticulatory phenomenon from an acoustic point of view. This kind of approach is justified by the fact that each articulatory mechanism must correspond to a different acoustic signal. The risk of this kind of approach is to ascribe erroneously a given acoustic feature to an articulatory parameter. So it's necessary in this method of analysis, to choose an acoustic feature that can be ascribed without any doubt to a specific

articulatory act. For our purposes, we have chosen the coarticulatory effect of lip rounding, because, if we consider the different occurrences of nodes and antinodes of velocity of the resonances along the vocal tract, lip rounding is the only articulatory feature to cause the simultaneous drop in all formants.
The studies concerned with the extension of the lip rounding, based on electromyographic analyses, have proposed two different hypotheses: on one hand; Lubker /6/ says that the anticipation of lip rounding can start at a maximum time of about 600 ms before the rounded vowel according to the length of the consonantal string; on the other hand Bell-Berti and Harris /4/ and Gay /7/ say that the anticipation of lip rounding is of about 250 ms independently from the number of consonants preceding the rounded vowel.
Following a different kind of approach based on the analysis of the acoustic signal, Pettorino and Giannini /8/ /9/ have studied the phenomenon of the lip rounding in Italian and Pettorino /10/ in English. They point out that the time of anticipation of lip rounding is not fixed and that there is a relationship between the extension of the anticipation of the lip rounding and the system of language. In fact in Italian the anticipation is related to the position of the stop inside the consonantal string and to the number of consonants: it is of about 150 ms when only one consonant is labialized whereas it is of about 190 ms when two consonants are labialized. In English these values are longer, being of about 190 ms and 220 ms respectively, regardless of the position of the stop consonant.
In this research we have chosen French because it is a language with a vocalic system strongly affected by lip protrusion, with six oral and two nasal rounded vowels. The aim of this experimental research is, on one hand, to verify whether among the rounded vowels different levels of labialization and consequently different extensions of coarticulation exist; on the other hand, to propose a model of syllabic organization for French and to compare it with those already obtained for English and Italian following the same method of analysis.

PROCEDURE

For the purposes of this experimental study we have prepared a list of about 200 French meaningful words containing the sequences VCV and VC..CnV. In these sequences V1 is /a/ /e/ or /i/ and V2 is /a/ /e/ /ε/ /i/ /u/ /o/ /ɔ/ /y/ /ø/ or /oe/. Our analysis has been limited to the oral vowels in order to avoid any possible interference caused by the added nasal cavities.
As regards VCV sequences the consonant is either a velar stop or a dental fricative. The choice of the velar stop was determined by the fact that, on the basis of the locus theory formulated by Delattre /11/, velar articulations have an F2 locus at about 3000 Hz when they are followed by an unrounded vowel, and at a lower value when they are followed by a rounded vowel. This allows us to verify spectrographically whether the velar stop is labialized or not. The choice of the fricative was determined by the fact that it shows a different distribution of the signal along the frequency scale according to whether it is labialized or not. In VCCV sequences one of the consonants is always /k/ or /g/ and the other one is selected among /s/ /l/ and /r/.
The randomized list of words has been read by two French male native speakers in an anechoic room and then recorded and analyzed using a Series 700 Sound Spectrograph by Voice Identification Inc. For each word a broad band spectrogram was obtained. Fo and intensity were measured using an FFM 650 and an IM 360 by F-J Electronics ApS.
We have compared the formant pattern of V1 in unrounded context with that of the same V1 in rounded context. As during their steady state the formant patterns don't show any noticeable difference, we have considered as onset of the lip rounding the point in which the two formant patterns diverge.

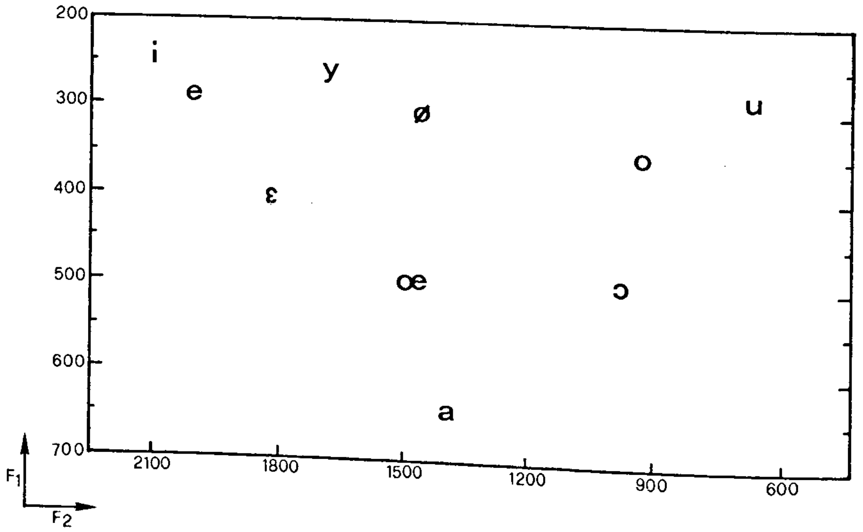


FIG. 1. Average values of F1 and F2 of French vowels.

RESULTS

Figure 1 shows the average values of F1 and F2 of all vowels. We have to notice that F1 of the front rounded vowels is slightly higher than F1 of the corresponding front unrounded vowels. From an articulatory point of view this means that the difference between the two series is not only due to the presence or absence of lip rounding but also to a different opening degree and place of articulation.
Table I shows the average durations of vowels and consonants in the different sequences.
As regards VssV sequence we have to notice that, even if the traditional grammars say that in French there isn't the functional opposition between short and long /s/ (see for instance Fouché /12/), our data show that when /s/ is graphically represented by double symbol, its average duration is of about 160 ms, that is twice as long as the dental fricative graphically represented by only one symbol. In the light of this, the sequence VssV, that from a phonological point of view is considered as VCV, is here considered as VCCV.

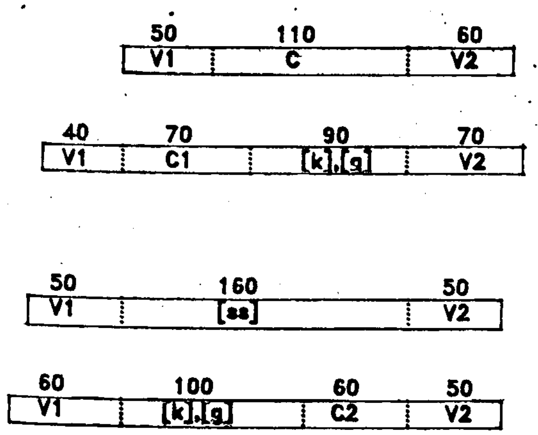


TABLE I. Average durations of vowels and consonants in ms.

Spectrograms show that in VCV sequences, where V2 is always /u/ /o/ or /ɔ/ the consonant is always labialized and the lip rounding begins inside V1 at about 30 ms before the onset of the consonant.
When V2 is /y/ /ø/ or /oe/, no anticipatory effect of lip rounding is noticed.
As regards VCCV sequence the onset of lip rounding occurs inside the offset of C1 except when it is a stop. In this case the anticipation of lip rounding starts from the offset of V1. So in VCCV sequence we have two possible kinds of syllabic organization that is VC-CV and V-CCV.
Table II shows the anticipation of the lip rounding in the different sequences.

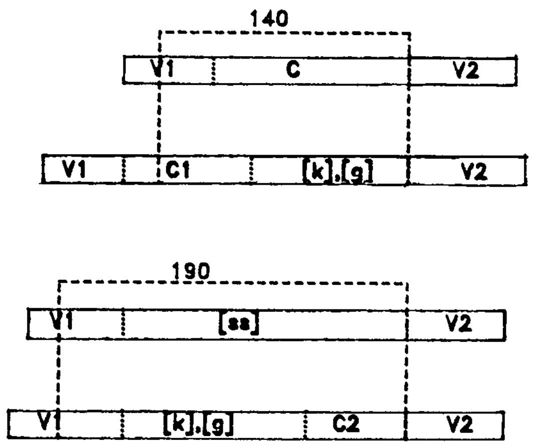


TABLE II. Anticipatory coarticulation of lip rounding in ms.

CONCLUSIONS

The data gathered in this experimental research show that the time of anticipation of the lip rounding is not fixed but it varies according to the number of consonants preceding the rounded vowel.
The times of anticipation for French are of about 140 ms when only one consonant is labialized and of about 190 ms when two consonants are labialized.
These values correspond to those found for Italian by Pettorino and Giannini /8/ /9/ but are considerably shorter than those found for English by Pettorino /10/. As we have said above, the limits of extension of the coarticulatory effect must coincide with those of the articulatory syllable. So, on the basis of our data, two different models of syllabic organization, one for Italian and French and one for English can be proposed. In addition to the difference in time of lip rounding anticipation, the two models differ also by the fact that in English in VCCV sequences the syllabic boundary always occurs inside V1

whereas in French and Italian it varies according to the position of the stop inside the consonantal string. It occurs inside V1 when C1 is a stop, inside C1 when C2 is a stop.

In order to clarify whether coarticulation has to be considered independent from the language system being "supplied by universal rules" /13/ or, on the contrary, "language specific" /14/, we must say that our data don't contribute to solve the problem. In fact on one hand it seems that a relationship exists between a specific language system and its syllabic organization, English being different from Italian and French; on the other hand, as regards the anticipation of lip rounding, there isn't any difference between French and Italian even though in these languages this articulatory feature plays a different role from a phonological point of view. In fact if in Italian lip rounding can be considered redundant, in that it is always co-occurrent with backness, in French it plays a distinctive role in that it occurs independently from the place of articulation.

In order to clarify whether a relationship exists between a specific language system and its syllabic organization it would be useful, using the same experimental method, to examine other Romance and Germanic languages.

As regards the vowels /y/ /ø/ /oe/, our data show that in no sequence examined there is anticipation of lip rounding. If we consider that /y/ /ø/ and /oe/ are front rounded vowels and /u/ /o/ and /ɔ/ are back rounded vowels, there must be a relationship between place of articulation and coarticulatory effect.

In order to verify this hypothesis it would be useful to examine systems of language with a large variety of front rounded and back unrounded vowels.

REFERENCES

/1/ V.A. Kozhevnikov, L.A. Chistovich, "Speech: Articulation and Perception" (translated from Russian), Joint Publications Research Service, Rep. 30, 543, Washington D.C., 1965.

/2/ R.G. Daniloff, K.L. Moll, "Coarticulation of Lip Rounding", J. Speech Hear. Res. 11, 707-721, 1968.

/3/ P.F. MacNeilage, J.L. DeClerck, "On the Motor Control of Coarticulation in CVC Monosyllables", J. Acoust. Soc. Amer. 65, 1268-1270, 1979.

/4/ F. Bell-Berti, K.S. Harris, "Anticipatory Coarticulation: Some Implications from a Study of Lip Rounding", J. Acoust. Soc. Am.65, 1268-1270, 1979.

/5/ J. Lubker, T. Gay, "Anticipatory Labial Coarticulation: Experimental, Biological and Linguistic Variables", J. Acoust. Soc. Amer. 71, 437-448, 1982.

/6/ J.F. Lubker, "Temporal Aspects of Speech Production: Anticipatory Labial Coarticulation", Phonetica 38, 51-65, 1981.

/7/ T. Gay, "Coarticulation in some Consonant-Vowel and Consonant Cluster-Vowel Syllables", Frontiers of Speech Communication Research Ed. by B. Lindblom and S. Ohman, 1979.

/8/ M. Pettorino, A. Giannini, "Some Aspects of Coarticulation in Italian: a Spectrographic Analysis of VV, VCV and VCCV Utterances", Wiener Linguistische Gazette, Suppl. 3, Wien, 1984.

/9/ M. Pettorino, A. Giannini, "Le rapport entre syllabe et coarticulation en italien", GALF, Paris, 1985.

/10/ M. pettorino, "A Model of Syllabic Organization: a Spectrografic Study of Coarticulation in English", IEE, London, 1986.

/11/ P. Delattre, "Coarticulation and Locus Theory", Studia Linguistica XXIII, 1-26, 1969.

/12/ P. Fouché, "Traité de prononciation française", Paris, 1959.

/13/ N. Chomsky, M. Halle, "The Sound Pattern of English", New York, 1968.

/14/ R. Hammamberg, "The Metaphysics of Coarticulation", Journal of Phonetics, 4, 353-363, 1976.

Grateful acknowledgement is made to Umberto Cinque for his precious technical collaboration.

PHONOLOGICAL CONCEPTION OF A SYLLABLE (APPLIED TO THE LANGUAGES WITH PHONIC SYSTEM)

VLADIMIR RUDELYOV

Russian Language Chair
Tambov State Pedagogical Institute
Tambov, Russia, USSR 392000

ABSTRACT

So far there is no phonological theory of a syllable. There are only some approaches to such theory in some linguists' papers /1/ or unexpected guessing at the genuine laws /2;3/.The syllabic theory has not enough explanatory power and no theoretico-informational basis in any of the cases,while these very points distinguish present-day phonology from empirical phonetics /4/.

INTRODUCTION

It turned out that farthest from phonology is the part of syllable theory studying syllable-boundary, i.e. the boundaries between separate syllables. Just here can we find the greatest bulk of the material for phonological interpretations. The material can become the basis of syllables classification and the basis for constructing the overall syllabic theory as a part of word phonology.
The essence of the above mentioned interpretations is the statement confirmed long ago intuitively and experimentally. The statement is that in the languages with phonic system a non-syllabic phoneme chooses the strongest position - either explosive ore implosive position. The position is stronger if it provides more correlations and relevance of the most important distinctive features /5/.
The Russian language is a brightest realization of the phonic system, just Russian is selected by us define the phonological rules of syllable-boundary; some of the rules hold good for other languages, some of them are uniqe and have different bases.

RULES OF SYLLABLE-BOUNDARY

<u>Rule-1</u>: A consonant or a sonorant preceding a vowel is explosive (symbol ".").
The examples of syllabication are:
/ḳa-ṛó-ṿa/ 'cow', /ẓ̂ə-ḷ'é-ẓə/ 'iron',

/pə-ra-xó-da/ 'of a ship',

/∂̣u-ṃá/ 'of the mind', /∂̣'é-x̣∂/ 'echo'.

The phonological background of this rule is that the pre-vowel position is a strong position for all the Russian phonemes.Vowels seldom occur in this position - such are the cases when writing fixes the vocalic beginning which does not exist in reality. According to rule-1 we can speak about two syllable types:

ṬA and ṚA;

the former begins with a consonant phoneme, the latter begins with a non-consonant phoneme,i.e.a sonorant or a vowel phoneme (cf.: /6, 173/ and /7, 100/.
<u>Rule-2</u>: A consonant preceding a sonorant is explosive.
This rule concerns not only Russian, but also the languages in which there is a consonantal correlation in resonance - voicelessness. In Russian it has the following background: consonants adjoin the explosive or some other sonorant because it provides for them a strong position in the feature of resonance - voicelessness, the most important and the most unstable feature of the phonological system /9, 6/.
The examples of syllabication are:

/ṃa-ṭṛó-ṣy/ 'sailors', /ə̣a-ḅḷa-ḳá/

'clouds', /ə̣a-ḅṃá-ṇa/ 'of the deception'.
Syllable type-2:

ṬṚA,

where Ṭ is a consonant, Ṛ is a sonorant, A is
a vowel. Final combinations of the kind contain the same explosive consonant and implosive sonorant (syllable typ-3: ṬR̭)

The examples of syllabication are:
/ṣṃó-ṭr̭/ 'review', /ṛu-ḅḽ'/ 'rouble',

/ṿó-ṗḽ'/ 'howl', /ḳá-ẓṋ'/ 'execution'.

<u>Rule-3</u>: A consonant or a sonorant in the absolute ultima

of a word is implosive.

This rule is as trite as the first one. But it is deduced, i.e. proven phonologically, in rather a complicated way.Not any consonant at the end of the word can be cosidered implosive. If in a language an implosive position does not differ from an explosive position in the number of correlations we cannot speak about implosion. Implosion is the loss of some relevant features,it is the position of neutralization.

In Russian (in most of its dialects) the position of a consonant in the word final is really an implosive position, while here the most important feature of the system - the feature of resonance - is dephonologized. Cf.: /ḳóṱ/ 'code' and 'cat', /lúk/ 'meadow' and 'onion'. The sonorants indifferent to the feature of resonance also have every ground to be in the implosive position as in it they lose their sonority getting mixed with vowels. However it happens rarely, when sonorants have vocalic pairs. In the Russian literary dialect only one sonorant /j'/ has a vocalic pair - the vocalic zero /ə/ (cf.: /8/, /9, 6/, /10/. Cf.: /ṃóə̭'/ 'my' (m.) and /ṃaj̣'á/ 'my'(f.)

In subdialects vocalic pairs have sonorants /l/ and /w/. Cf.: / ̣u-p̣á-ḷa/ 'fell down' (f.) and /əu-p̣áṷ/ 'fell down' (m.); /ṭṛa-ẉá/ 'grass' - /ṭṛáṷ/ 'of the grasses'.

The phonological feature of resonance for consonants is functionally identical with the feature of sonority for non-consonants (vowels and sonorants). Dephonologization of the both is the index of an implosive position for the Russian language.

Rule-3 gives syllable types: ṬAṰ, ṬAR̭, ṚAṰ, ṚAR̭, ṬṚAṰ, ṬṚAR̭. Rule-4: A sonorant preceding a sonorant is implosive.

Phonologically this rule is confirmed by the fact that this position is weak in the feature of "sonority - vocality", i.e. sonorants are vocalized in it if they have vocalic pairs. Cf.: /ṿaə̭'-ṇá/ 'war', dial. /ẉoṷ-ṇá/ 'wave'. If a sonorant has no vocalic pair it is not vocalized befor another sonorant, but it gains implosion: /ṿaḽ-ṇá/ 'wave', /ṭóṋ-ṇa/ 'ton', /ḳar̭-ṃá/ 'stern', /ṭ'ur̭'-ṃá/ 'prison', /ə̣ar̭-ḷá/ 'of the eagle'. A sonorant can be implosive at the beginning of a word too (syllable type R̭). Cf.: /m̭-ṛák̭/ 'gloom', /ḽ'-ṇá/ 'of the flax', /m-nú/ 'am crumpling', /m-n'ý/ 'crumple!'.

At the end of a word two sonorants present an example of double implosion, but the implosive segments belong to different syllables:

/ṿóə̭'-ṋ/ 'of the wars', /g̣ór̭-ṋ/ 'bugle'.

Rule-5: A sonorant preceding a consonant is implosive.

Phonological background for this rule is similar to that for rule-4. The examples of syllabication are:

/ṭṛóə̭'-ḳa/ 'troika', /p̣áḽ-ḳa/ 'stick', /ḳ'ýr̭-ḳa/ 'pick', /ṭaḽ-p̣á/ 'growd', /ə̣áḽ'-f̣a/ 'alfa', /ḳar̭-ṭó-f̣'əḽ'/ 'potatoes'. Cf. also: /ḽ'-ḍá/ 'of the ice', /ḽ'-ḍ'ý-ṇa/ 'block of ice', /r̭-ẓ̂ý/ 'of the rye', /ḽ-g̣ú/ 'am telling a lie'.

At the end of the word an implosive sonorant together with an implosive consonant present the same syllable demonstrating an extraordinary mirror reflection in the arrangement of structural elements. Cf.: /s̤ḳáḽ'ṱ/ 'skald', /ṃór̭s̭/ 'fruit juice', /púl't/ 'control panel' - the implosive elements of a syllable come the reverse way as comared with the explosive elements.

Rule-6: A consonant preceding a consonant is neutral in the way of explosion-implosion,

i.e. non-explosive and non-implosive or both explosive and implosive.

The neutral consonant T̤ is a certain intersyllabic element, however it is not an independent syllable and according to the situation it belong to one of the adjacent syllables.

In fact the neutral explosive-implosive position mixes consonants in the feature "resonance - voicelessness" and consequently it follows the law of implosion. But it reveals assimilative dependence on the next syllabic.

Cf.: /ṿaḽ-ẓ̂b̤á/ 'magic', /ṿa-g̤zál/ 'railway station' and/ṇóŝ̤-ḳa/ 'leg'. Cf. also: /ə̣a-c̤cá/ 'of the father', /ə̣óĉ̤'-ĉ̣'ym/ 'stepfather'.

The existence of the intersyllabic elements which are not independent syllables and which adjoin either the preceding or the following syllable complicates the procedure of word syllabication.

Not independent syllables, neutral segments have to adjoin the preceding or the following syllable becoming positionally implosive or explosive. The essence of positional explosion or implosion is that it depends on the informational force of syllables. As a rule the most informational syllable of a word is stressed and this very syllable subordinates the segment neutral in the way of explosion - implosion and makes this segment a part of its own.

The neutral segment T̤ complicates the syllable without changing its structure much. Cf.: syllables TA and TTA. But in some cases adjoining the neutral segment makes the syllable of one type resemble the syllable of another type.

Cf. syllables: ṬAR̭, ṬAR̭T̤ and ṬAR̭Ṱ,

ṬṚAR̭, ṬṚAR̭T̤ and ṬṚAR̭Ṱ

Quite evident is the actual neutralization of syllables wich results from adjoining the neutral segment T̤. While such neutralization brings about the increase in the number of close syllables we can admit that just open syllables have the greatest informational value in present-day Russian and in the languages of the same system.

At the beginning of a word the neutral segment T̤ is always positionally explosive:

/s̤ṭṛáx̭/ 'fear', /p̤ṭ'ý-c̣a/ 'bird' etc.

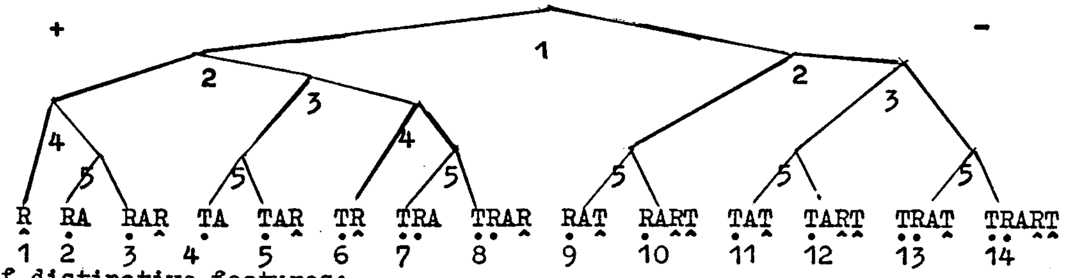
At the end of a word the position of a neutral segment is unusual: the syllables having an implosive consonantal element are complicated by the segment T̤ so that the latter takes the position after the vowel and precedes the implosive segment proper: /p̣ós̤ṱ/, /ṿós̤k̭/ 'wax',

cf. also: /Ṣ'ýṋk̤x̭/ 'Singh',

/ḅór̭ŝ̤'ŝ̭'/ 'bortsch',

/ṿóə̭'s̤k̭/ 'of the troops'.

Classification scheme of Russian syllabemes:



Names of distinctive features:

DF-1 "absence of consonantal implosion",
DF-2 "absence of consonantal explosion",
DF-3 "absence of sonorant explosion",
DF-4 "absence of the vocalic syllable-bearer,
DF-5 "absence of sonorant implosion".

CONCLUSION

In conclusion we can state for sure that in Russian there are strict rules of arranging phonemes about syllables. These rules obey phonological laws which are 6 in number. As to the syllables proper there stand out among them the syllabic structures which do without a vowel syllable-bearer. A consonant cannot form an independent syllable.

LITERATURE

/1/ L.Hjelmslev, "The syllable as structural unit", "Proceedings of 2nd CPhS", London, 1938;
A.Rosetti, "Sur la théorie de la syllabe", "Bulletin linguistique",III,Bucureşti,1935;
E.Курилович,"Очерки по лингвистике", Москва, I962.

/2/ Р.И.Аванесов, "Фонетика современного русского литературного языка", Москва, 1956.

/3/ Л.В.Щерба, "Фонетика французского языка", Москва, I939, с. 75-76.

/4/ H.Karlgren, "Speech rate and information theory", "Proceedings of the 4th ICPhS", The Hague, 1962, p. 670.

/5/ В.Г.Руделев,"Слогоотделение в русском языке", "Материалы научн. филолог. конф. вузов Уральской зоны", Свердловск, I969.
/6/ Е.Курилович, "Очерки", с. I73.
/7/ В.Г.Руделев, "Фонология слова", Тамбов, I975.
/8/ Т.-Р.Вийтсо, "Об одной возможности описания фонологии русского языка","Труды по русской и славянской филологии",Тарту, I963, с. 405-409.
/9/ В.Г.Руделев /ред/, "Теория нейтрализации", Тамбов, I980.

/I0/ В.Г.Руделев,"Вокалический нуль и его место в системе вокальных фонем", "Материалы и тезисы докл. ХУ итог. научн. конф.", Оренбург, I967.

VOT PRODUCTION IN APHASIA: CONTEXTUAL AND LEXICAL INFLUENCES.

WOLFRAM ZIEGLER DETLEV VON CRAMON

Neuropsychological Department
Max-Planck-Institute for Psychiatry
D-8000 München 40, FRG

ABSTRACT

Two experiments revealed that the voice onset time (VOT) of initial plosives in aphasic speech may be conditioned by the voicing of medial consonants and by the lexical status of the involved stimulus.

INTRODUCTION

The voice onset time of plosives is considered to play an important role in the classification of aphasic speech errors and in the differentiation of aphasic syndromes (e.g. /1,2/).
In normal speech the VOT of (initial) stop consonants is known to be sensitive to context influences such as the quality, tenseness, and duration of the subsequent vowel or the voicing of the post-vocalic consonant (see /3/ for references). To the extent that certain aphasic patients are particularly prone to disturbances of sequential processing the role of such effects in aphasic speech deserves particular consideration.
On the other hand, linguistic variables which are not known to play a role in normal speech production may nonetheless influence voice onset time in the condition of aphasia. Among the factors considered to influence the error rates of many aphasics are for instance the grammatical class, the frequency, and the lexical status of a word (e.g. /4, 5/).
When uncontrolled, such effects may cause a systematic increase in the variability of VOT data in aphasic speech. On the other hand, the VOT may provide us with a quantitative measure to study contextual interdependences and lexical effects and thus to promote our understanding of the processes underlying the phonetic and/or phonological impairments in aphasia.
The present study focussed on two of the aforementioned effects. In a first experiment, the anticipatory influence exerted by voiced vs. voiceless medial plosives upon the VOT of initial stop consonants was investigated in aphasic patients with and without apraxia of speech.
A second experiment was designed to assess VOT differences in voiced and voiceless initial stops of word-nonword minimal pairs in speech apraxics.

CONTEXT INFLUENCES

Methods

Subjects: Context influences were examined in six aphasic patients and in a normal and a dysarthric control. All aphasic patients had suffered from occlusions of the left middle cerebral artery. Aphasia testing revealed two cases each of Broca's and Wernicke's aphasia. The two remaining patients had unclassifiable aphasic disturbances. Together with the two Broca's aphasics they presented with the clinical symptoms of apraxia of speech: Their speech was characterized by numerous substitutions and distortions of speech sounds and an inconsistency in the articulatory pattern. All of the speech apraxics had prosodic impairments.
The dysarthric patient had suffered from a midbrain hemorrhage and presented with the symptoms of ataxic dysarthria.

Materials and procedure: A pseudorandom test list was prepared, containing four bisyllabic nonwords of the form /'daCo/, /'taCo/ with C = /b,p/ interspersed with a number of dummy words. The subjects were required to repeat the test utterances upon aural presentation by an examiner. Each word was produced at least 20 times within one session. Patient examinations were performed in a sound-treated room and recorded using high-quality equipment. Preceding each session the discriminatory abilities of the patient with respect to the voicing contrast were examined in a same-different task with taped presentations of test stimuli. Each patient reported here performed well on the discrimination task, yielding almost 100% correct discriminations, yet often with delayed responses. The recordings were digitized on a LSI 11/73. Words containing an error unrelated to the voicing feature in either the initial or the medial plosive were excluded from further analysis. For the initial plosive of each target word two examiners measured the voice onset time as defined by the interval between the plosive burst and the onset of periodicity.

Results and discussion

Figure 1 presents the cumulative VOT distributions resulting from the productions of the normal subject (left) and the dysarthric patient (right). In both cases the data reveal a slight tendency towards increased VOT values in utterances with voiced medial plosives as compared to the voiceless context, an effect which proved significant in the case of the dysarthric patient's /t/-tokens (Mann-Whitney, two-tailed; p<0.02). The difference contours depicted at the bottom of fig.1 describe both the direction and the extent of the measured effects.

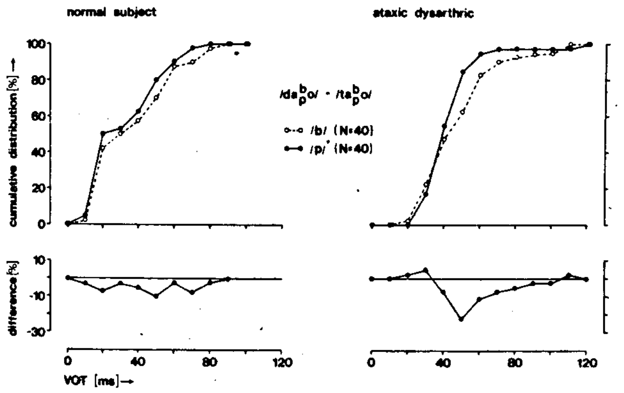


Fig.1: Sensitivity of voice onset time of initial plosives to medial stop voicing: Cumulative distributions (top) and voiceless-minus-voiced difference contours (bottom) in a normal and a dysarthric control.

A possible explanation of this effect which has been described earlier /3/ might be that the two controls, in the sense of an "elaborate pronunciation", attempted to increase the syntagmatic contrast between the two plosives in the test words.
The difference contours given in figure 2 demonstrate consistent contextual effects in two patients with apraxia of speech and two Wernicke's aphasics. Unlike the two controls of fig.1, the aphasics' samples with voiced medial plosives tended to have shorter VOT than when the medial plosive was voiceless. In three of the six cases the measured effects were rather marked and proved significant (Mann-Whitney, two-tailed; p<0.001). In the remaining patients the same effect was present, yet not statistically significant.

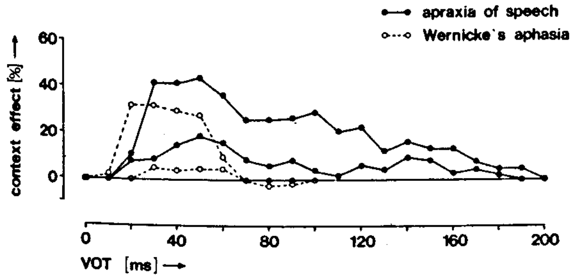


Fig.2: Difference contours as in fig.1 for two Wernicke's aphasics and two patients with apraxia of speech.

Following this pattern, the aphasic patients were, unlike the two controls, influenced in the sense of an assimilation of the initial plosive to the voicing feature of the subsequent medial plosive. Notably, this assimilation occurred across place differences between the two neighboring stop consonants.

A major difference between apraxic and non-apraxic aphasics was that the former covered a broader overall VOT range in their productions. Moreover, the two groups differed in the pattern of context sensitivity (fig.3).
The diagrams of fig.3 describe VOT distributions for the examined word pairs as approximated by two-component least-squares fit models /2/. In none of the cases was the approximation error greater than 5%. In the normal speaker (left), the distributions obtained for /tabo/ and /tapo/ were similar in their shape with a slight shifting of the former towards higher VOT values. The patient with apraxia of speech (middle) presented /dabo/-samples with increased VOT values, concentrated at almost equal proportions in two modes around 50 ms and 80 ms. In the condition of voiceless context the 80 ms peak became overproportionate and a number of outliers above 100 ms occurred at the expense of VOT values below 40 ms. It should be pointed out that this patient's /t/ productions occupied a range above 100 ms, meaning that neither of the two peaks in the /d/-distribution of /dabo/ or /dapo/ can be considered as representing literal paraphasias (for a broader discussion of this issue see /2/). The measured context effect therefore is presumably not a result of phonemic changes in the voicing category.
In the Wernicke's aphasic (right), on the other hand, the effect exerted by context variation produced a dual pattern in the VOT distributions of /dabo/ and /dapo/: the majority of /d/-productions in the voiced context assumed values between 5 and 25 ms, i.e. within a normal range, and a smaller proportion of voice onset times was distributed around 50 ms. In the voiceless context, these relations were reversed. It is of importance to know that the peaks near 50 ms were characteristic of this patient's realizations of /t/. In contrast to the speech apraxic the context influence observed in the Wernicke's aphasic may therefore be described as a triggering of literal paraphasias in the voicing djmension.

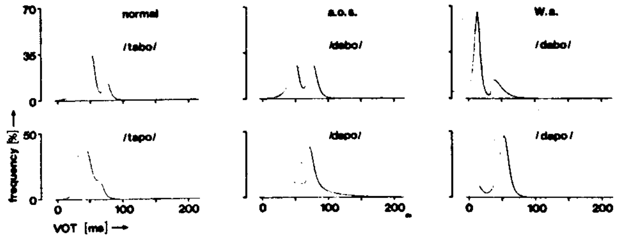


Fig.3: VOT distributions (least squares fit models /2/) for /d/ or /t/ with varying medial plosives for a normal subject and two aphasic patients.

LEXICAL INFLUENCE

Methods

Subjects: In a second experiment, three normal subjects and six aphasic patients (two Broca's, one conduction and three patients with very mild, unclassifiable aphasic symptoms) were involved. Again, all patients had suffered from occlusions of the middle cerebral artery. All patients were diagnosed as presenting the symptoms of apraxia of speech.

Materials and procedure: The test words used in this experiment consisted of the two minimal pairs /dynə/ ("dune") - /tynə/ (nonword) and /dyrə/ (nonword) - /tyrə/ ("door"). The first of these two pairs will in the following be referred to as the 'word-nonword pair', the second as the 'nonword-word pair'. The test words were arranged in a pseudo-randomized order with interspersed dummy words and repeated upon taped presentation for at least 20 times each. Testing was performed as described above, again with an examination of auditory discrimination preceding each session. The VOT of the initial plosive of each target word was measured according to the criteria mentioned earlier.

Results and discussion:

The curves plotted in figure 4 were obtained by subtracting the cumulative percentage of word-nonword samples from that of nonword-word samples within 10 ms VOT bins for three of the speech apraxics and a normal subject. In the latter, the difference contour oscillated around 0, indicating that there was no systematic difference between the VOT distributions of the word-nonword and the nonword-word pairs. In contrast, the three patients presented considerable positive deviations from 0 in their difference contours, meaning that in their productions the nonword-word pairs tended to have larger VOT values than the word-nonword pairs.

This was corroborated by testing the differences between words and nonwords for each of /d/ and /t/ separately (Mann-Whitney, two-tailed; N=80). Among the patients a significant (p<0.005) bias of nonwords towards words was obtained in four out of six cases, yet always in only one of the two stop cognates. The respective median values in these cases differed by approximately 10 ms (8 ms to 13 ms). In a further patient a less marked effect was found (p<0.025), which, however, remained stable when this subject was re-examined (N=160). Only the patient with the mildest aphasic and apraxic symptoms demonstrated an opposite tendency.

Notably, a significant bias in the nonword-word direction occurred also in one of the three normals. This effect, although small, was preserved upon increasing the data base (N=160).

These results reveal that, obviously, the meaningful words in the two pairs played the role of "attractors", causing in their meaningless counterparts a measurable change in plosive VOT towards the "word" - target. These deviations consisted of gradual shiftings of the entire distribution rather than categorical jumps of single tokens and could therefore not be explained by a literal paraphasia model. Hence the lexical effect must arise at a stage of processing where the motor patterns pertaining to the target word are specified in their phonetic detail. In terms of the two-route model hypothesized by McCarthy & Warrington /4/ one could speculate that the auditory-phonological and the semantic-phonological transcoding process are simultaneously active in the repetition of test words. When the input is a nonword, its phonetically "neighboring" lexical entries, including presumably the voicing counterpart of the test stimulus, are activated. Yet, at the level of VOT specification the articulatory programs implemented along the lexical route and the direct route interfere with each other, yielding the described bias. However, this interference becomes effective only if the desired stimulus, i.e. the "word"-member of the respective word-nonword pair, actually receives the highest activation load among several potential candidates in the mental lexicon. It was probably due to this weakness in the paradigm that the effect could be measured in only either /d/ or /t/. If an appropriate paradigm were available that controls lexical access effectively a more significant influence would possibly result.

The observations made regarding one of the normal subjects suggest that the hypothesized interaction may in principle be effective in normals, too, which would provide an analogy to the lexical effects observed in word perception experiments /6/. Nevertheless, in the examined speech apraxics the bias in favor of words against nonwords was considerably greater, meaning that these patients are more vulnerable to lexical influences in speech production. One might speculate that a similar effect is present in words of differing frequency.

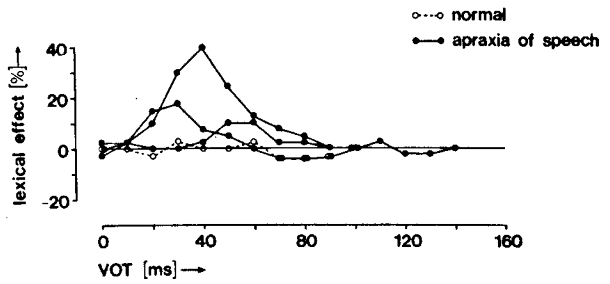


Fig.4: Difference contours resulting from cumulative VOT distribution functions (10 ms intervals) as explained in fig.1 (nonword-word minus word-nonword). Open circles: normal subject; closed circles: patients with apraxia of speech.

CONCLUSIONS

The outcome of both experiments reported here strongly suggests that plosive VOT in aphasic speech is sensitive to linguistic variables. With regard to the influence exerted by neighboring voiced vs. voiceless consonants on a target plosive the measured effects were opposite to expectations from normal speech. Apraxic and non-apraxic aph~~asics, in~~ their own way, tended to reduce the syntagmatic voicing contrast. The lexical effect observed here included a VOT bias of meaningless stimuli towards meaningful counterparts. This effect was interpreted along the lines of a two-route model of word repetition.

Our results stress the requirement of controlling the speech materials used in VOT studies of aphasic speech with respect to linguistic variables.

REFERENCES

/1/ S.E.Blumstein, W.E.Cooper, H.Goodglass, S.Statlender, J.Gottlieb: Production deficits in aphasia: A voice-onset time analysis. Brain & Language 9: 153-170 (1980)

/2/ W.Ziegler: On the phonetic realization of phonological contrast in aphasic patients. In: J.Ryalls (ed.), Phonetic Approaches to Speech Production in Aphasia and Related Disorders. College Hill Press (in press)

/3/ G.Weismer: Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. J. of Phonetics 7: 197-204 (1979)

/4/ R.McCarthy, E.K.Warrington: A two-route model of speech production. Evidence from aphasia. Brain 107: 463-485 (1984)

/5/ J.M.Dunlop, T.P.Marquardt: Linguistic and articulatory aspects of single word production in apraxia of speech. Cortex 13: 17-29 (1977)

/6/ W.F.Ganong III: Phonetic categorization in auditory word perception. J. Exp.Psychol.: Hum.Percept.Perform. 6: 110-125 (1980)

NATURALLY OCCURRING VOT CONTINUA IN APHASIC SPEECH: PERCEPTUAL CORRELATES

WOLFRAM ZIEGLER PHILIP HOOLE

Neuropsychological Department
Max-Planck-Institute for Psychiatry
D-8000 München 40

ABSTRACT

Using plosive-vowel stimuli spliced out of aphasic speech VOT-continua were prepared and presented to experienced listeners for identification of the voicing of the initial consonant. Possible range effects on listeners' judgments were studied. Stimuli with unexpected response rates are discussed.

INTRODUCTION

Errors of consonant voicing are considered to play an important role in the categorization of aphasic production deficits. They have often been reported in transcription studies (e.g. /1/) and have, at least with respect to certain aphasic syndromes, been interpreted as reflecting a deficit in the temporal coordination of articulatory and laryngeal gestures.
A major shortcoming of the perceptual assessment of consonant voicing is that it leads to a description based on discrete entities where continuous measures would probably be more adequate. Therefore, disturbances of plosive voicing have been adressed in voice-onset time studies of various aphasic syndromes, revealing that aphasic speakers can vary considerably in their VOT /2/.
Yet, to our knowledge there has not yet been any attempt to systematically examine a particular corpus of aphasic speech material both from the point of view of judgments of discrete categories and measurements of a continuous variable such as VOT. Such an inquiry, however, would be of great interest since it cross-relates the results obtained on different methodological bases. One would predict that the perception of stop consonants produced by aphasic patients is largely conditioned by the measured VOT and that only productions with voice onset times near the category boundary are ambivalent in perception. Yet, other cues to stop voicing may be present in aphasic speech which override the perceptual significance of voice onset time.
The fact that VOT in some aphasic patients covers a broad range of considerable density was used in this study to construct sets of plosive-vowel stimuli along VOT continua. On the basis of such continua, category boundaries were estimated for the productions of three aphasic patients in identification experiments, asking whether these boundaries are invariant over the different subjects. The possible influence of different VOT ranges covered by a stimulus set /3/ was also investigated. Stimuli which did not attract the responses expected from their VOT are discussed with regard to possible acoustic features that override the VOT cue.

EXPERIMENT 1

Method, subjects and material

The material for the perception experiments was taken from speech utterances of 3 female patients (aged 46, 55 and 64) suffering from apraxia of speech following occlusion of the left middle cerebral artery. In a production experiment reported in detail elsewhere the patients had been required to produce 20 repetitions of the nonsense words /dabo/, /dapo/, /tabo/ and /tapo/. For these words the length of the segment from release burst of the initial consonant to onset of periodicity (VOT) was determined by obtaining a consensus of opinion from three examiners (/d/ generally has no voicing lead in South German). The initial consonant and 20 periods of the vowel /a/ were then digitally spliced out, after eliminating those words showing gross deviations in place of articulation for the consonant, or in loudness or fundamental frequency for the vowel. The vowel length of 20 periods was chosen so that the vowel would be as long as possible, but without including the transition to the following plosive. These spliced segments represent the actual stimuli used in the experiments; a total of 68 stimuli spanning the VOT continuum were obtained for patient A1, 56 for A2 and 70 for A3. In contrast to classical synthetic continua the stimuli were not absolutely equidistant along the VOT axis.
A panel of listeners with considerable experience in perception experiments took part in the listening tests; six listeners for Patient A1 and five listeners for A2 and A3. Each listener was tested separately. In each session the randomized stimuli of one patient were presented to the listeners, with 4 repetitions of each stimulus. Each patient was presented in 3 sessions with a different randomization each time; thus each listener heard each stimulus of each patient 12 times. The listeners were required to identify the stimuli as /da/ or /ta/ (forced-choice). Presentation of stimuli was performed by a laboratory computer using a 20 kHz sample rate for D/A conversion with a 9kHz low-pass filter prior to output over high-quality loudspeakers.

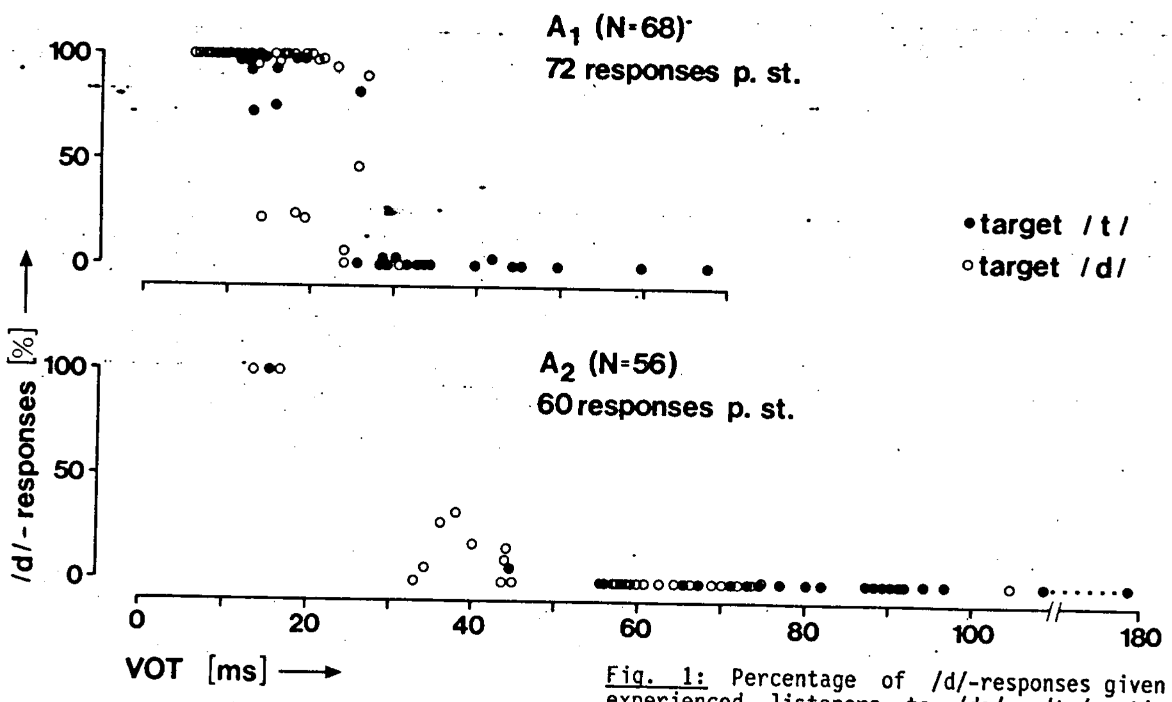


Fig. 1: Percentage of /d/-responses given by experienced listeners to /da/, /ta/ stimuli spliced out from utterances of patients A1 and A2.

Results

In Fig. 1 the percentage of /d/-responses is displayed versus measured VOT for patients A1 and A2. The figure also indicates whether the production target had been /t/ or /d/. For patient A1 most VOT values lie between 5 ms and 50ms. Intended /t/-productions mainly occupy the range above 25 ms. Responses are unanimously /d/ below about 10 ms and /t/ above 30 ms. The category boundary would seem to lie between 20 and 25 ms, but with some ambivalent stimuli in the region below this diverging widely in the number of /d/-responses. The responses were obviously not conditioned by the actual production target, since intended /t/ productions, in particular, were often unambiguously perceived as /d/. There are also several intended /d/ productions with low VOT that are nonetheless almost unanimously identified as /t/.
Productions of Patient A2 covered a much wider range of VOT values, with a general shift to higher values. There are unfortunately some gaps in the continuum; only three productions received 100% /d/-responses. It is also noticeable that several stimuli with a VOT of about 40 ms still received ambivalent responses while for Patient A1 /t/-responses in this range had already risen to 100%. Thus the category boundary seems to be further to the right for A2. The question of whether this might be due to the different VOT ranges was addressed in the second experiment outlined below.
The results of the first twenty stimuli of the continuum from Patient A3 are given in Tab.I with the corresponding VOT value. All further stimuli i.e. nos. 21-70, with VOTs up to almost 200 ms gave 100% /t/-percepts. Perhaps the most surprising aspect of the response pattern is that even at very low VOT values very high numbers of /t/-responses were obtained for some stimuli. The possible acoustical reasons for this will be examined in the discussion.

Table I: VOT and percent /d/ responses for the first 20 stimuli from the continuum of Patient A3.

stimulus-no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
VOT (ms)	14	14	17	17	19	20	22	22	23	23	23	24	24	24	25	25	27	30	33	36
/d/-responses (%)	100	13	23	100	100	28	32	2	12	47	22	32	5	3	100	2	10	2	0	0

3 0323

EXPERIMENT 2

In experiment 2 we wanted to investigate whether response behavior to these kinds of continua is susceptible to range effects, thus possibly explaining the apparent difference in category boundary between A1 and A2.

Listeners and methods remained the same, but the range of stimuli presented from each patient was changed in the following way:

For Patient A1 all stimuli with a VOT longer than 28 ms were discarded, i.e. the range in which only /t/-responses occurred in the first experiment. This left a total of 48 stimuli. In Patient A2, a similar criterion led us to discard all stimuli with VOT above 55 ms, which left a total of 16 stimuli.

These manipulations of the stimulus continuum were expected to lead to more /t/-responses, particularly on the ambivalent stimuli.

The stimuli were randomized with 10 repetitions each and presented to the listeners in 2 sessions for A1 and in a single session for A2. These sessions took place several weeks after those of Experiment 1.

Results

Fig. 2 compares the percentage of /d/-responses for the full and reduced continua, with the responses grouped into 5 ms bins for Patient A1 and 10 ms bins for A2. In both cases there is a slight increase in /t/ responses in the reduced condition. But the effect, while going in the expected direction, is clearly rather weak. It appears somewhat stronger if only the ambivalent stimuli are regarded (i.e. those neither 100% /t/ or /d/ in the first experiment). This is displayed in Table II averaged over all listeners and in Fig. 3 for a single listener.

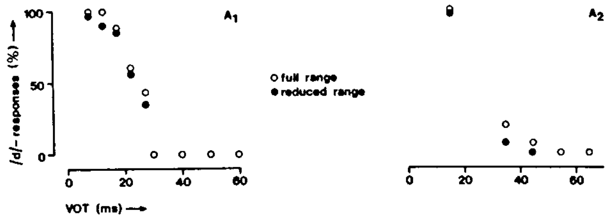


Fig. 2: Change in /d/-response rate from Experiment 1 to Experiment 2 (grouped stimuli, all listeners). Left: Patient A1. Right: Patient A2.

Table II: Percentage of /d/-responses given to ambivalent stimuli of patient 1 (N=19) and patient 2 (N=8) under two different range conditions.

		condition	
		full range	reduced range
patient	1	75	45
	2	14	4

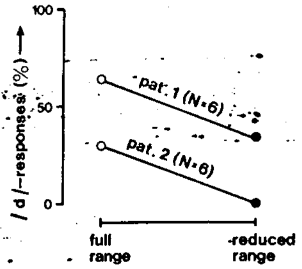


Fig. 3: Change in /d/-response rate from Experiment 1 to Experiment 2 for one listener, restricted to stimuli that were ambivalent in Experiment 1.

DISCUSSION

From the identification results in Fig. 1 and Table I a rough estimate of the category boundary can be made for each patient. However, particularly for Patients A1 and A3 the numerous stimuli not falling on an ideal response curve make it clear that VOT cannot be the only factor conditioning listeners' responses. This is hardly unexpected when one considers that the continua consist of a large number of naturally-occurring stimuli rather than a very restricted number of synthetic or electronically manipulated stimuli. Nevertheless, it is interesting to look closer at those stimuli that fall out of line when ordered according to VOT. Fig. 4 gives one example for each of Patients A1 and A3 in which similar measured VOT values attracted widely differing /d/ response-rates. At the bottom of the figure in the example for A3 both utterances are rather weak in harmonic energy in the high frequencies. The obvious difference, however, is that in the left-hand example the onset of harmonic energy occurs more or less simultaneously at least up to the frequency range of the first formant while in the right-hand example for about 70 ms after voicing onset there is only harmonic energy at the fundamental frequency, presumably because of incomplete glottal adduction. It could be argued that it would be more appropriate to adopt a different criterion for measuring VOT, e.g. the interval from plosive release to onset of harmonic energy in the region of the second or higher formants (cf. Klatt /4/). However, in cases such as the present one even this criterion could be difficult to apply consistently since in the sonagram it is very difficult to discern harmonic energy in the higher formants. This methodological problem would probably be even more acute when measuring in the time-domain. Beyond such technical considerations, however, it seems more important to stress that not all aspects of disturbed glottal adduction can be captured in a single parameter such as VOT, however it is defined.

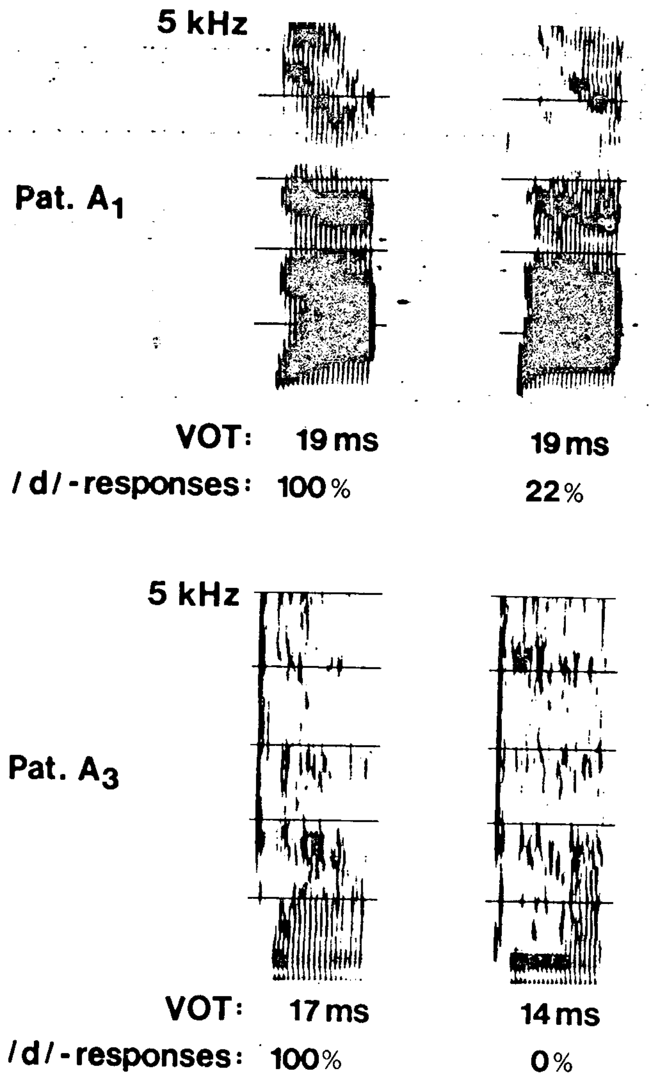


Fig. 4: Two examples of CV-stimulus pairs with similar VOT which attracted different voicing judgments.

The example given in the top half of the figure illustrates a rather different phenomenon. Here it is fairly safe to say that any criterion for VOT measurement would lead to essentially the same value. However, the number of /d/-responses again diverges widely. The most salient difference between the two sonagrams would seem to be the much more visible F1 transition in the production that attracted 100% /d/-responses. This probably reflects differences in the extent of lingual anticipation of the vowel target at the moment of plosive release. One might speculate that a patient like A1 makes active use of such a movement pattern in order to partially compensate for her obvious deficits in the control of VOT. However, the lack of a correlation between intended target and percept in this subject speaks against this view.

Turning to the results of Experiments 1 and 2 with regard to range effects it should first be recalled that for Patient A2 several stimuli still attracted ambivalent reponses in Experiment 1 at a point on the VOT continuum where responses for A1 and A3 were already unanimous (i.e. around 40 ms). Under the assumption that these ambivalent responses to some extent represent the listeners' reaction to the overwhelming number of clear /t/ stimuli it could be expected in Experiment 2 that any manipulation of the VOT range covered in such continua should be reflected in the response-rates, particularly near the category boundaries (cf. /3/). The results are not as clear as one might wish, but there was a trend in the expected direction in both A1 and A2. It is noticeable that the slightly unexpected ambivalent stimuli for Patient A2 almost disappeared in the second experiment.

REFERENCES

/1/ J.E.Trost, G.J.Canter: Apraxia of speech in patients with Broca's aphasia: A study of phoneme production accuracy and error patterns.

/2/ S.E.Blumstein, W.E.Cooper, H.Goodglass, S.Statlender, J.Gottlieb: Production deficits in aphasia: A voice-onset time analysis. Brain & Lang. 9: 153-170 (1980)

/3/ S.A.Brady, C.J.Darwin: Range effect in the perception of voicing. J. Acoust. Soc. Am. 63: 1556-1558 (1978)

/4/ D.H.Klatt: Voice onset time, frication and aspiration in word-initial consonant clusters. J. Speech & Hear. Res. 18: 686-706 (1975)

NORMAL AND APHASIC PROCESSING OF SENTENCE STRUCTURE AND INTONATION[1]

Aita Salasoo
Department of Psychology
State University of New York
Binghamton, NY 13901 USA

Rita Sloan Berndt
Department of Neurology
University of Maryland Medical School
Baltimore, MD 21201 USA

ABSTRACT

The contributions of intonation contour and memory load to performance in an auditory grammaticality judgment (AGJ) task were investigated. College students, agrammatic and other aphasics, and control subjects judged the grammaticality of vocoded utterances with original and flat fundamental frequency (F0) contours. For normal and aphasic listeners, sensitivity and bias differences between seven syntactic structures outweighed smaller benefits from intonation information and few memory demands. Differences emerged according to the severity and selective linguistic deficits among the aphasics. The observed strategic components of syntactic processing have implications for previous interpretations of AGJ data in relation to normal and aphasic language behavior.

INTRODUCTION

While little is known about how syntactic knowledge is used during speech processing, a task that involves explicit grammaticality judgments has been used increasingly to address issues about the language deficits of individuals classified as agrammatic aphasics.[1] The present study attempts to provide baseline data from normal language users and to probe the role of extra-syntactic factors in the task. These goals are prerequisites to the use of the auditory grammaticality judgment (AGJ) task in the study of aphasia. In a recent study[1] four agrammatic patients who failed to use syntactic devices successfully in their language production and in comprehension tasks showed great sensitivity to the violations of those syntactic structures, when asked to judge whether spoken sentences were grammatical. Poor performance was found only in conditions using tag questions and reflexive pronouns. It was concluded that agrammatic aphasics do not have a general syntactic deficit, but that they fail to use syntactic structure in more demanding tasks. The account of the poor performance structures given is in terms of poor semantic encoding of lexical features that cannot support dependent syntactic analyses.

The present study aims to strengthen these conclusions by ensuring that the pattern of results will not generalize to nonagrammatic populations and by addressing possible confounds in stimulus materials. Two factors that may contribute to AGJ performance are intonation and memory load. Normal listeners can use prosodic cues including the pitch or intonation perceived from the fundamental frequency (F0) contour of a sentence in many listening tasks. Agreement does not exist about the dominant source of information when intonation and syntax conflict. [2] [3] Since syntax guides the F0 contours in speech production,[4] listeners may be able to use this information to perceive spoken language. The role of intonation in grammaticality judgments may be studied by removing the information carried by F0 and asking listeners to judge the grammaticality of the resultant utterances. To the extent that performance is worse for these stimuli, intonation cues to grammaticality are implicated.

Syntactic knowledge is relational. In the temporal course of fluent speech, syntactic violations may increase memory demands as the duration (or intervening information load) between violating segments is increased. Adjacent violating elements have fewer memory demands than distant ones. Increased memory demands, in turn, may decrease the detectability of syntactic violations.

The present investigation examines the influence of these properties (available to normal and possibly aphasic listeners) in the AGJ task. Specifically, violations to seven syntactic structures are used to investigate the effects of intonation contour and memory load on performance. In Experiment 1, a group of normal listeners is studied, and in Experiment 2, five aphasics with left-hemisphere lesions and three control subjects are tested.

EXPERIMENT 1

Method

Subjects. The subjects were 48 students from SUNY Binghamton, who were native English speakers aged 18-23 with no known speech or hearing problems.

Materials and design. Seven syntactic violation types were selected to include six which presented few problems to patients in the Linebarger et al. study[1], and one type – verb copying in tag questions – on which patients performed poorly. In each violation type, stimulus pairs were generated whose members differed in grammaticality, and as little as possible on other properties. A variety of lexical and transformational rules are represented in the violation types; they are described fully elsewhere.[5]

Approximately equal numbers of the Linebarger[1] sentences and new sentences were used. The additional sentences employed medium- to high-frequency words to increase the vocabulary of the stimulus set. Sentence length and violated constituent size were controlled. Memory load was operationalized as **violation location**, defined as the point at which the utterance could no longer be completed as well-formed. Nongrammatical stimuli were classified as having **early** (first three words), **middle**, or **late** (last two words) violation locations. Only late violations may have a wide range of distance (in number of words) between the disagreeing sentential elements. In all 156 critical pairs, and 24 practice pairs were employed. The mean length of the grammatical and nongrammatical utterances is 7.88 and 7.86 words, respectively. Overall, 26 pairs have early violations, 67 middle violations, and 63 late violations. Table 1 summarizes the stimuli.

[1] This work was supported by NINCDS grants NS-21054 to the University of Maryland Medical School and K04-NS-00851 to R. Berndt. A. Salasoo is now at Bell Communications Research, Inc., MRE 2G-356, 435 South St, Morristown, NJ 07960, USA.

Table 1: Example Grammatical and Nongrammatical Stimuli (*) and Violation Location Distribution of 7 Syntactic Violation Types. (E, M, and L refer to early, middle, and late violations, and N indicates the total number of stimulus pairs per condition.)

Syntactic Violation Type	Grammatical and Nongrammatical Examples	E	M	L	N
1. Verb control complements	I let Harry pay for the birthday cake. * I let Harry to pay for the birthday cake.	2	16	6	24
2. Missing verb arguments	Before the guests come I'll put the toys in the closet. * Before the guests come I'll put the toys.	3	11	10	24
3. Subject-auxiliary inversion	Hasn't Fred walked the dog yet? * Hasn't Fred hasn't walked the dog yet?	8	4	0	12
4. Verb copying in tag questions	The pies aren't very large, are they? * The pies aren't very large, do they?	0	0	24	24
5. Missing noun phrase elements	I doubt I could afford a month's vacation in Greece. * I doubt could afford a month's vacation in Greece.	8	7	9	24
6. Wh-movement: Fronted noun phrases	Whose broken fence will Gary fix next week? * Whose broken will Gary fix next week fence?	5	11	8	24
7. Gapless relative clauses	She washed the windows that needed cleaning. * She washed the windows that the floors needed cleaning.	0	18	6	24

A male speaker read a random order of the stimuli at normal speed and with normal intonation. To prevent the occurrence of hesitations and other abnormalities in the F0 contour for nongrammatical stimuli,[4] model sentences were employed. The model was a grammatical sentence matched in number of syllables, and in the actual words as much as possible, to the following nongrammatical stimulus. When possible, the grammatical sentence served as the model. Recordings were digitized at 10 kHz, low-pass filtered at 4.8 kHz, and stored on a computer. After signal processing -- an LPC vocoder extracted 14 coefficients from overlapping 30-ms windows of the speech signal – each natural digitized utterance yielded a pair of vocoded stimuli, one with the **natural** F0 contour ranging from 75 - 175 Hz, and the other with a **flat** 90-Hz F0 contour.

Eight experimental tapes were generated for both the natural-F0 and flat stimuli. Each tape contained 3 practice and 39 critical stimuli with equal numbers of grammatical and nongrammatical utterances, and approximately equal representation of each violation type. A warning signal of three 1000-Hz tones preceded each stimulus. It was followed by two presentations of the stimulus, separated by an intersentence interval of one second. The intertrial interval was six seconds.

Procedure. Eight random groups of 6 subjects each listened to 2 natural-F0 and 2 flat-F0 tapes.. A subject heard either the grammatical or nongrammatical version of each stimulus pair in either its natural-F0 or flat condition. The order of tapes and intonation conditions was counterbalanced between groups. Subjects were instructed to listen to the two presentations of each sentence and to decide whether or not the sentence was grammatical, recording their judgment in a response booklet. Stimuli were presented at a comfortable listening level from a speaker situated 1-2 meters away from the subjects.

Results and discussion

For each subject, the mean proportion of hits (correctly accepting a grammatical sentence) and false alarms (FAs – incorrectly accepting a nongrammatical utterance as grammatical) in each experimental condition were computed. Nonparametric signal detection techniques were applied to the measurement of sensitivity to grammaticality and performance bias with the **A**′ (A-prime) and **B**′ (B-prime) statistics, respectively.[6] If $y = p(Hit)$ and $x = p(FA)$, then:

$$A' = 0.5 + \frac{(y-x)(1+y-x)}{4y(1-x)} \qquad 0.5 \leq A' \leq 1.0$$

$$B' = \frac{y(1-y) - x(1-x)}{y(1-y) + x(1-x)} \qquad -1.0 \leq B' \leq 1.0$$

In the present task, higher values of **A**′ indicate greater sensitivity to grammaticality. **B**′ values of 0 indicate optimal criterion that maximizes hits and minimizes false alarms, i.e., no bias. Negative non-extreme values indicate a lax criterion; positive non-extreme values indicate a strict criterion. Table 2 shows mean hit and false alarm rates, **A**′ and **B**′ values for normal listeners making grammaticality judgments as a function of violation type and intonation condition. The reported findings are based on analyses of variance (ANOVAs) and post hoc comparisons and are statistically significant beyond the level of **p** < .01 unless otherwise indicated.

Sensitivity to grammaticality. Overall, listeners were very sensitive to syntactic structure. An ANOVA of **A**′s with violation type and intonation as factors found main effects of both variables. Large differences were observed between the seven violation types, **F**(6,42)=17.95. Post-hoc comparisons revealed three clusterings: Sensitivity was greatest for violation types 3 and 6, intermediate for violations types 2, 4, and 7, and lowest for types 1 and 5. Several accounts of this pattern need exploration, including the possibility that greater F0 (and syntactic) discontinuity existed for type 3 and 6 violations.

Table 2: Grammaticality Judgment Data for Normal Listeners as a Function of Violation Type and F0 Contour.

Violation	F0 Contour	p(Hit)	p(FA)	A′	B′
1	natural	.891	.143	.928	-.116
	flat	.877	.185	.909	-.166
2	natural	.906	.077	.953	.009
	flat	.878	.074	.946	.220
3	natural	.972	.010	.990	.466
	flat	.895	.040	.961	.420
4	natural	.860	.025	.957	.663
	flat	.831	.047	.941	.516
5	natural	.871	.118	.929	.004
	flat	.866	.130	.924	.001
6	natural	.930	.026	.975	.440
	flat	.919	.027	.972	.478
7	natural	.871	.087	.940	.172
	flat	.869	.088	.939	.173

Despite the very high levels of performance, stimuli with preserved intonation contours were judged with greater sensitivity than flat stimuli, $\mathbf{F}(1,47)=21.20$, and importantly, intonation and violation type failed to interact statistically, $\mathbf{F}(6,42)=1.15$, $\mathbf{p}>.35$. Thus, while the presence of F0 information increased sensitivity to grammaticality, it did not differentially affect that sensitivity for the various syntactic structures tested.

Next, the effects of memory load were examined through the violation location variable. A main effect of location was observed, $\mathbf{F}(2,46)=6.97$, but this variable again failed to interact with intonation, $\mathbf{F}<1.0$, $\mathbf{p}>.70$. Post-hoc tests revealed that sensitivity was greatest for early violations (mean $\mathbf{A}'=.957$) and did not differ for midsentence and late violations (mean $\mathbf{A}'$s of .939 and .931, respectively). The result suggests that memory load contributes to sensitivity in grammaticality judgments: The violating segments are either adjacent or separated by a single word in the early conditions and thus may be encoded and stored together in memory, decreasing the memorial demands of the judgment task. Another possible interpretation is in terms of primacy, namely that sentence-initial constituents are attended and encoded better than constituents in later positions, and hence early violations are detected better.

Response bias. Arguments about sensitivity differences hold with greatest strength if subjects maintain a constant bias in performance. Evidence about this was derived from ANOVAs of the $\mathbf{B}'$ measure analogous to those performed for sensitivity. The seven violation types differed in performance bias, $\mathbf{F}(6,42)=6.04$. Only for stimuli in the verb control complement condition (where least sensitive performance was observed) was the mean $\mathbf{B}'$ negative, suggesting a lax criterion: Subjects tended to accept both grammatical and nongrammatical utterances in that condition as grammatical. In contrast, a very strict criterion was observed for the tag questions condition. No main effect of intonation contour on bias was observed, $\mathbf{F}(1,47)=1.40$, $\mathbf{p}>.24$. Notably the bias differences between violation types are more dramatic than those related to intonation contour information. Examined by violation location, the most optimal criterion was found for the midsentence position (mean $B'=.010$), $\mathbf{F}(2,46)=5.21$. Early and late violations did not differ statistically: Both exhibited strict criterion placement. Again, intonation did not interact with the location effect for performance bias, $\mathbf{F}(2,46)=2.20$, $\mathbf{p}>.12$.

In sum, Experiment 1 has provided five major findings about how normal listeners make grammaticality judgments. First, young adults are very sensitive to the grammaticality of spoken sentences. Second, in most cases, unbiased or conservative response criteria are adopted. Third, grammaticality is not unitary knowledge -- performance differed in both sensitivity and bias for different types of syntactic structures. Fourth, memory load as measured by violation locations may play a small part in sensitivity in the AGJ task. Finally, intonation contour information, even when optimally modelled for nongrammatical utterances, has a facilitatory effect on sensitivity to syntactic structure. These results suggest that grammaticality judgments offer valid evidence about the syntactic processes that occur during normal speech processing. The substantial differences in performance observed among our stimuli are not primarily attributable to intonational or memorial factors, but reflect, instead, strategies used by listeners. By placing strict response criteria, listeners are able to judge the grammaticality of sentences accurately; the only exception to this criterion placement resulted in poor judgment performance.

EXPERIMENT 2

In Experiment 2 aphasic patients and control listeners make grammaticality judgments for the speech utterances used in Experiment 1. The study aims to replicate the Linebarger et al.[1] results with listeners diagnosed as agrammatic aphasics and to compare performance across various subject populations. Agrammatics are expected to perform well in all conditions except the tag questions. College students in Experiment 1 had no problems with that syntactic structure; analogous data do not exist for other aphasic, i.e., nonagrammatic, populations nor for normal listeners matched to agrammatic patients in age and education.

Method

Subjects. Five aphasics with left-hemisphere lesions following cerebrovascular accidents participated. One of them, VS, had been tested by Linebarger et al.[1], and all of them had prior experience with the natural utterances on which the stimulus set was based. Full subject descriptions, including detailed reports of comprehension tasks and of performance with those natural speech stimuli may be found in Berndt et al.[5] For present purposes, the subjects are informally grouped as agrammatic (FM, VS), nonagrammatic aphasic (JD and JS have mild and severe auditory comprehension deficits, HY is anomic), and 3 normal control listeners of similar age and background.

Materials and procedure. The stimuli and tapes from Experiment 1 were used. All subjects were tested with stimuli preserving the F0 contour before flat stimuli over the course of 4-8 weeks. Subjects were tested individually and heard 2 tapes per session. The experimenter recorded the subject's verbal grammaticality judgment. Patient and control data are reported individually, since agrammatic and nonagrammatic labels may not reflect homogeneous deficits.

Table 3: Grammaticality Judgment Data for 2 Agrammatic Listeners

Violation Type	F0 Contour	FM				VS			
		p(Hit)	p(FA)	A′	B′	p(Hit)	p(FA)	A′	B′
1. Verb control complements	natural	.70	.30	.79	.00	.88	.68	.71	-.35
	flat	.70	.35	.76	-.04	.96	.70	.78	-.69
2. Missing verb arguments	natural	.83	.39	.81	-.26	.88	.23	.90	-.25
	flat	.75	.35	.79	-.10	.96	.41	.88	-.73
3. Subject-auxiliary inversion	natural	.67	.18	.83	-.20	.75	.45	.74	-.14
	flat	.75	0.0	.94	-	.92	.36	.87	-.52
4. Tag questions	natural	.91	.83	.64	-.27	.04	.04	.50	0.0
	flat	.87	.83	.57	-.11	0.0	.04	.50	-
5. Missing NP elements	natural	.48	.35	.62	.05	.83	.58	.72	-.27
	flat	.65	.26	.78	.08	.96	.74	.77	-.67
6. Wh-movement: Fronted NPs	natural	.83	.27	.86	-.17	.83	.25	.87	-.14
	flat	.78	.24	.85	-.03	.87	.43	.82	-.37
7. Gapless relative clauses	natural	.70	.14	.86	.27	.91	.38	.86	-.48
	flat	.65	.36	.72	-.01	.91	.32	.88	-.45

Table 4: Grammaticality Judgment Data for 3 Nonagrammatic Aphasic Listeners

Type	F0 Contour	JD				JS				HY			
		p(Hit)	p(FA)	A′	B′	p(Hit)	p(FA)	A′	B′	p(Hit)	p(FA)	A′	B′
1	natural	.91	.62	.77	-.48	.65	.50	.63	-.05	.70	.48	.68	-.08
	flat	.96	.40	.88	-.72	.48	.40	.58	.02	.78	.50	.73	-.19
2	natural	1.0	.42	.90	-	.46	.35	.60	.04	.64	.26	.78	.09
	flat	1.0	.35	.91	-	.42	.41	.51	0.0	.92	0.0	.98	-
3	natural	.92	.45	.84	-.54	.92	.82	.67	-.33	.83	.36	.83	-.24
	flat	1.0	0.0	1.0	0.0	.83	.64	.69	-.24	.92	0.0	.98	-
4	natural	.22	.04	.75	.63	.91	.74	.71	-.40	.26	.04	.77	.67
	flat	.09	.04	.65	.36	.65	.78	.30	-.14	.56	.04	.87	.73
5	natural	.96	.63	.81	-.72	.43	.37	.56	.03	.46	.18	.74	.25
	flat	.96	.68	.79	-.70	.56	.53	.53	-.01	.74	.26	.82	0.0
6	natural	.91	.14	.94	-.19	.70	.52	.66	-.08	.74	.14	.88	.23
	flat	.96	.14	.95	-.52	.65	.43	.68	-.04	.74	0.0	.94	1.0
7	natural	.86	.36	.84	-.31	.65	.41	.69	-.03	.55	.09	.84	.50
	flat	.91	.41	.85	-.49	.65	.32	.75	.02	.65	.14	.84	.31

Results and discussion

The control subjects performed well in the AGJ task. For them, mean sensitivity was computed and compared to the performance of young adults in Experiment 1, with the seven violation types as repeated measures. No difference in sensitivity was seen for the natural-F0 stimuli, $\mathbf{t}(6)=1.14$, $\mathbf{p}>.10$, but the older control subjects outperformed those in Experiment 1 with the flat stimuli, $\mathbf{t}(6)=5.60$, $\mathbf{p}<.01$. One reason for this result may be that in the present study improvements due to stimulus repetition are masking possible decrements in performance caused by the absence of intonation information in the flat stimuli. In fact, contrary to Experiment 1, these subjects showed no effect of intonation on sensitivity to grammaticality, $\mathbf{t}(6)=1.35$, $\mathbf{p}>.10$. Sensitivity was greatest for violation types 3 and 6, consistent with Experiment 1.

Table 3 shows AGJ performance of two agrammatic listeners. Both patients performed better than chance, but worse than the control listeners, and the subjects tested by Linebarger et al.[1] In particular, VS who was tested in the earlier report showed less sensitivity to syntactic structure in the present study. Several factors of our stimuli[5] may have contributed to this result. More importantly, both agrammatic patients showed least sensitivity to manipulations in tag questions, as predicted by the Linebarger et al.[1] results. Moreover, this syntactic structure elicited dramatic changes in decision criteria for both patients, but in opposite directions: FM accepted the majority of type 4 stimuli as grammatical, while VS rejected virtually all of them. Notably, FM also performed poorly with the two structures causing normal listeners most difficulty. Another difference between the two agrammatics is that FM, but not VS, showed normal effects of memory demand. AGJ performance was consistent across intonation conditions, suggesting that previous AGJ results with aphasics are probably not based on reliance on prosodic cues to syntax.

The final part of this study compares performance of other aphasic patients (see Table 4) to that of agrammatics. JS who has an asyntactic comprehension deficit but is fluent[5] could not perform the AGJ task. Only for the subject-auxiliary inversion manipulations was JS's performance above chance. Both JD and HY have milder deficits, and indeed, performed almost as well as normal listeners. Notably, JD had a lax decision criterion in all conditions except the tag questions -- there, he rejected most stimuli as nongrammatical, and showed least sensitivity. The similarity to the agrammatic pattern seen for VS exists, despite the overall high levels of performance. A similar bias shift was seen for the anomic HY with the tag questions, but it was not accompanied by decreased sensitivity. Instead, the verb control complement violations involving lexical rules were most difficult for HY. Also, JD and HY, but not JS, showed an advantage for early violations, suggestive of the normal pattern of memory effects. Most importantly, removal of F0 information did not alter the pattern among the syntactic structures for each patient.

CONCLUSIONS

In sum, signal detection methods allowed detailed investigation of how aphasic listeners differ (from one another and from normal listeners) not only in selective sensitivity to syntactic structures, but also in response bias in the AGJ task. Experiment 2 has replicated and extended the results of Linebarger et al.[1] by reporting similar findings from 2 agrammatics and 1 nonagrammatic aphasic and different patterns for 2 other aphasics and for normal listeners. Intonation information did not influence the syntactic differences observed for both normal and aphasic listeners. For aphasics with nonsevere deficits and for normal listeners, performance suffered when memory load was increased. These results suggest a complex relationship between sentence intonation contour, memory demand, and response bias in grammaticality judgment tasks.

REFERENCES

1. Linebarger, M.C., Schwartz, M.F., Saffran, E.M. (1983) Sensitivity to grammatical structure in so-called agrammatic aphasics. **Cognition, 13**, 361-392.
2. Berkovits, R. (1984) A perceptual study of sentence-final intonation. **Language and Speech, 27**, 291-308.
3. Nespor, M. & Vogel, I. (1983) Prosodic structure above the word. In A. Cutler & D.R. Ladd (Eds.) **Prosody: Models and measurements** (pp.123-140). Berlin: Springer-Verlag.
4. Sorensen, J.M., & Cooper, W.E. (1980) Syntactic coding of fundamental frequency in speech production. In R.A. Cole (Ed.) **Perception and production of fluent speech** (pp. 399-440). Hillsdale, NJ: Erlbaum.
5. Berndt, R.S., Salasoo, A., Mitchum, C.C., & Blumstein, S.E. (1986) The role of intonation cues in the grammaticality judgment task. Submitted for publication.
6. Grier, J.B. (1971) Nonparametric indexes for sensitivity and bias: Computing formulas. **Psychological Bulletin, 75**, 424-429.

AN EXPERIMENTAL INVESTIGATION OF SPEECH PERCEPTION IN MOTOR APHASIA

ALEXANDER Y. PANASYUK

Dept. of Phonetics
Leningrad University
Leningrad, USSR 199034

IRINA V. PANASYUK

Dept. of Foreign Languages
Extramural Polytech. Institute
Leningrad, USSR 191041

ABSTRACT

The perception of sentences with directional prepositions was studied in motor aphasia in two experimental conditions: in Exprriment 1 aphasics had to identify prepositions in spoken sentences, while in Experiment 2 the sentence as a whole had to be understood and identified with a pictorial representation. The type of testing had a significant effect on the subjects' error number, suggesting that they chose different strategies. The types of perceptual confusions also depended on the testing procedure. The phonetic factor is of primary importance in preposition identification, whereas the semantic meaning of prepositions plays a major role in sentence identification.

I. INTRODUCTION

This report reviews a series of experiments investigating speech production and perception disturbances in motor aphasia. Our previous studies have shown that speech difficulties are caused by the inability or reduced ability to process and reproduce particular combinations of phonetic features in Russian (e.g. voicedness and softness in plosives) while at the same time retaining the correct use of other features (e.g. voicedness and softness in fricatives) / 1, 2 /. Motor aphasics seem to be quite sensitive to the phonetic realization rules, i.e. they produce and perceive those phonemes successfully which are "strong" from the point of view of production or perception and fail to pronounce and perceive "weak" phonemes which are characterized by a comparative complexity on either level. Another difficulty arises when individual phonemes have to be combined in speech utterances. As motor aphasics are unable to produce certain phoneme sequences they tend to simplify the sound chain of a word to form a CVCV q sequence.

It is well known that a typical feature of motor aphasia on the syntactic level is so called motor agrammatism which is characterized by the omission of prepositions and other function words /3,4,5/. Prepositions are normally unstressed and therefore lack prominence compared to open class words. It is this peculiarity which, in the opinion of some researchers accounts for the exceptionally rare occurrence of prepositions in the free speech of patients with motor aphasia /6/. Observations made in the course of remedial speech work with motor aphasics have led us to suppose that the omission of prepositions is caused, among other factors, by their inability to realize certain phonemic sequences which may occur at the preposition-noun boundary, namely, consonant clusters. Similarly, one may suppose that the perception of phonetically strong or more prominent prepositions (with a CV or VC structure) will be quite different from phonetically weak ones (consisting of one consonant).

In the present study we tried to test experimentally the hypothesis of the phonetic nature of prepositional agrammatism and to find out whether aphasics' performance and perception was affected by test conditions. Another aim was to evaluate the ability of motor aphasics to make use of the semantic features of prepositions.

II. METHOD

In order to investigate motor aphasics' ability to comprehend sentences with prepositions as well as perceive these prepositions, a program was constructed which consisted of 5 series of sentences. Each series contained 6 sentences with the following directional prepositions: "k" (to, towards), "v" (in, into), "na" (onto), "ot" (from), "iz" (out of) and "s" (from). The sentences described simple spatial situations which could easily be realized in the form of line drawings, for example, "The bird is flying up to the cage", "The bird is flying out of the cage", etc.

The prepositions in question could be divided into two groups according to semantic meaning: 1)prepositions denoting movement towards an object and 2) prepositions describing the reverse movement, i.e. away from the object.

Phonetically, these prepositions can be ascribed to two opposed classes: 1) syllabic and 2) non-syllabic. In the Russian language prepositions ending in a consonant (or consisting of one consonant only) have at least two variants due to the regressive assimilation of voicedness/voicelessnees. Thus, the preposition "k" may be spoken as "k" or "g" depending on the quality of the initial consonant in the subsequent noun or adjective. Besides, when the following word begins with a "k" or a "g" the resulting cluster is pronounced as "kk" or "gg". In this experiment all nouns had a voiceless initial consonant so that all final consonants in prepositions were pronounced as voiceless, e.g. /k, f, at, is, s /.

The program consisting of 30 sentences was read aloud by a male speaker of Russian and recorded on magnetic tape. Each sentence was pronounced twice.

A special apparatus was devised and constructed by one of the researchers to conduct audio-visual matching experim - ents. A detailed description of the apparatus is given in / 7 /.

The subject was seated in front of a screen onto which either one or four pictures could be projected. The pictures represented schematically different spatial situations.

In Experiment 1 the subjects were shown a picture simultaneously with the auditory presentation of a corresponding sentence and asked to press one of the six panels on a special board which contained the preposition that matched the sentence and the line drawing. In other words, the subjects had to choose the right preposition out of the six given alternatives.

In Experiment 2 four pictures were projected on separate panels constituting the screen. One picture matched the orally presented sentence, a second required an opposite direction preposition, a third showed a closely related movement in the same direction but slightly different in its final stage, e.g. not "into" but "onto", and the last one was chosen randomly from the remaining situations. The subject had to press one of these panels which completely corresponded to the sentence.
In both experiments the time of non-verbal reaction was measured automatically.

There were 4 subjects with motor aphasia in the study. Two subjects with normal speech were also investigated. Two aphasics were tested twice in Experiment 1 and thus the total number of responses was 180. In Experiment 2 two subjects were tested twice and one subject three times, giving 240 responses in total.

III. DISCUSSION OF THE EXPERIMENTAL DATA

First of all, it should be mentioned that the average error number was much higher in Experiment 1 (28%) than that in Experiment 2 (16%). This indicates that it is much more difficult for aphasics to detect (in the spoken utterance) a particular segment (a preposition) and recognize it from a number of alternatives than to identify this utterance with one of the given visual stimuli (line drawings). Normal subjects made no mistakes in either experiment.

Figure 1 gives the matrices of preposition substitutions in Experiment 1 and 2 (in per cent).

(A)

prepositions perceived

prepositions presented

	k	s	v	na	ot	iz
k	57	13	13	7	3	
s		73		3	3	21
v	7	10	70	7	3	3
na			3	77	17	3
ot	7	3	7		73	10
iz	3	7	3	3		84

(B)

	k	s	v	na	ot	iz
k	95		5			
s	2	68		7		23
v	8		85	2		5
na			2	95	3	
ot	2				95	3
iz		15	4		13	68

Fig.1 Confusion matrices of prepositions in Experiment 1 (A) and Experiment 2 (B). Each line of the matrix shows

the distribution of prepositions (in per cent) the aphasics perceived in sentences presented as indicated in the left margin. Cells forming the diagonal are correct perceptions. Prepositions are transliterated (the prepositions "v" and "iz" were actually pronounced as "f" and "is", the transliteration of the other consonants in the prepositions coincides with the transcription).

For most prepositions (4 out of 6) the number of correct responses was higher in Experiment 2 than in Experiment I. There are two exceptions, i.e.prepositions "s" (from) and "iz" (out of). Both prepositions belong to the "away from the object" group. It seems to be worth mentioning that the preposition which was the easiest to identify ("iz") is phonetically the longest, while the preposition which turned out to involve the largest number of errors was the one with shortest duration ("k"). It was quite natural to expect that the phonetic structure of a preposition was crucial for its correct identification. To test this hypothesis, we calculated the percentage of correct identifications of one-phoneme versus two-phoneme prepositions. The data obtained confirmed this hypothesis. One-phoneme prepositions were identified correctly in 67% of cases whereas two-phoneme (syllabic) prepositions were correctly identified in 78 per cent of cases.

At the same time, it seemed sensible to suppose that the phonetic structure of prepositions would affect the comprehension task in Experiment 2 to a lesser degree than in Experiment I. This hypothesis proved to be true. The total count of correct responses in Experiment 2 to one-phoneme prepositions was approximately the same as to two-phoneme prepositions (83% and 88%, respectively).

In order to investigate further the role of the phonetic and semantic factors in speech perception and comprehension, we analyzed the types of most frequently encountered mistakes from considering the length and semandic meaning of prepositions. The experimental data have shown that there is a "universal substitutor", namely, the preposition "iz" (out of) which replaces other prepositions in most cases in both experiments. As mentioned above, this preposition ranked first in the list of identification accuracy in Experiment I.

It was rather tempting to look for a phonetic tendency in the perceptual confusion pattern.

In Experiment I one-phoneme prepositions revealed a slight tendency to be perceived as two-phoneme prepositions. In the case of two-phoneme prepositions, the mistakes were determined by chance.

In Experment 2 the aphasics tended to perceive one-phoneme prepositions as two-phoneme ones much more frequently than vice versa (71% and 29%, respectively). The two-phoneme prepositions, however, were randomly confused with prepositions of either class.

Our next task was to test the semantic hypothesis of prepositional agrammatism according to which the perceptual im - pairments in aphasics reflect the complexity of semantic features associated with a certain preposition.

Experiment I was designed to evaluate the ability of motor aphasics to process the speech flow and identify its segments As expected, the semantic factor is not of primary importance for the aphasic listener and, consequently, the percentage of correct responses to prepositions denoting movement towards an object was roughly equal to that of prepositions having the opposite meaning (74% and 77%).

Conversely, in Experiment 2 the role of the semantic factor increases: the prepositions that denote movement towards an object are 1.2 times more frequently perceived correctly than those which describe movement away from it. The total number of correct answers is 92% and 77%, respectively.

An analysis of the mistakes has shown that motor aphasics understand the directional meaning of a preposition and are able to ascribe it to one of the two basic groups but fail to recognize the more subtle features. Thus, in Experiment 2 the "towards an object" prepositions are confused with each other in 75% of the cases of erroneous responses and the "away from the object" prepositions undergo substitutions within the group in 75% of all cases. There was no uniformity in the subjects' performance in respect of this semantic feature in Experiment I.

The reaction time during the two perceptual experiments was found to be much longer in motor aphasics than in normal subjects and to vary significantly from stimulus to stimulus. There seems to be no relationship between the type of preposition and the latency period. The latter appears to be a function of the severity of motor aphasia.

IY. CONCLUSIONS

Our data support the hypothesis that disturbances in production and perception of prepositions so often encountered in motor aphasia are related to the phonetic structure of these function words.

The relative importance of the phonetic factor in the perception of prepositions depends on the type of the perceptual task. Thus, it seems to be of primary importance in tasks involving the phonemic analysis of speech flow (as in Experiment I) and is replaced by other factors, e.g. semantic, when experimental conditions provoke the subject to listen to an utterance not merely for its segment analysis but for its comprehension as a whole (as in Experiment 2).

ACKNOWLEDGEMENTS

We wish to thank Dr Yu.I.Kuzmin, Head of the Speech and Voice Pathology Laboratory at the Research Institute of Diseases of Ear, Throat, Nose and Speech in Leningrad for his help in experimental work in the field of speech pathology. We are also grateful to Miss Margaret Inglis for her helpful suggestions in reading the English text of the present report.

REFERENCES

/I/ I.V.Panasyuk, "Sound Disturbances in Patients with Motor Aphasia", Cand.Sci.Dissertation(Applied Linguistics), Leningrad University, 1980.

/2/ G.M.Sumchenko, I.V.Panasyuk, "Use of Two Types of Testing for the Evaluation of Phonemic Perception in Aphasia" (Russian text), In: Neuropsychological Investigation in Neurology, Neurosurgery and Psychiatry, Publications of the Bechterev Institute, vol.XCVII, Leningrad, I98I, pp.89-92.

/3/ E.S.Beyn, T.G.Vizel, F.M.Hatfield, "Aspects of agrammatism in aphasia", Language and Speech, 1979 vol.22, pt 4, pp.327-346.

/4/ A.D.Friederici, "Production and Comprehension of prepositions in Aphasia", Neuropsychologia, I98I, vol.I9, № 2, pp.I9I-I99.

/5/A.Yu.Panasyuk, I.V.Panasyuk, "Aphasic Disorders in the Perception of Sentences with Spatial Prepos-

/6/ M.C.Kean, "The Linguistic interpretation of aphasic syndromes: agrammatism in Broca' aphasia, an example", Cognition, I977, 5, pp.9-46.

/7/ A.Yu.Panasyuk, I.V.Panasyuk, "An experimental technique of training and testing speech perception abilities in aphasia", In: Problems of Voice and Speech Pathology, Publications of the Research Institute of Diseases of Ear, Throat, Nose and Speech(Russian text), Moscow, 1983, pp.97-I03.

PHONEME DISCRIMINATION AND THIRD LANGUAGE LEARNING

MARIJKE LOOTS

Faculty of Letters
Catholic University Brabant
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

ABSTRACT

Bilingual children appeared to make fewer errors in discriminating between the phonemes of a foreign language than their monolingual peers. This difference is explained on the basis of a greater awareness on the part of bilinguals of the fact that phonemic boundaries may differ per language.

INTRODUCTION

In August 1986 the teaching of English at Dutch primary schools was made obligatory. One of the questions that this decision has raised is whether the addition of English to the curriculum will constitute an extra burden for immigrant children, who are already in the difficult position of having to keep up two languages, the command of one of them often being inferior to that of the other. A research project was set up to study the impact of bilingualism on third language learning, including various tests on the learning of grammar, vocabulary and phonology. What this paper on phonological contrast has in common with the whole study is that it is not limited to achievement, but rather explores the strategies bilingual children develop to cope with unfamiliar linguistic input.
Most of the research on bilingualism comes from the United States, which has a long history of bilingual education. For a long time bilingual children were considered to have a handicap in comparison with their monolinguistic peers [1][2]. The poor performance on a number of tasks was ascribed to a negative influence of bilingualism on the children's intellectual development. The most serious flaw in these early investigations seems to be that they did not control for the fact that bilingual children often originated from families with a comparatively low socio-economic status. Since the sixties, studies into bilingualism have shown that "balanced" bilingual children show a wide range of advantages rather than disadvantages when compared with monolinguals from a comparable socio-economic background, the most conspicuous of these being the ability to separate word sound from word meaning. Some authors have claimed that bilingualism also helps individuals to learn additional languages more easily ([2], p. 180). To our knowledge, bilingualism has hardly ever been studied in relation to foreign language learning, the one exception being an achievement study in Sweden. In a nationwide investigation with standardized tests on vocabulary, grammar, reading and listening comprehension, it appeared that bilingual children performed slightly better than monolingual children when compared on the level of parental education [3]. In the same way that bilingual children may have a greater cognitive flexibility than monolingual children, and are better able to distinguish between form and meaning, they might also be more aware of the fact that the relation between sound and phoneme need not always be the same. As they already have two phoneme sets at their disposal and consequently a larger number of phoneme boundaries, they might be more capable of discriminating between the phonemes of a third language than their monolingual peers. This hypothesis formed the starting point of the present experiment into the relation between bilingalism and sound discrimination in foreign (third) language learning.

THE TEST

Subjects

Forty children who were pupils of the two final grades of ten different schools took part in the experiment. Their age varied from 9 to 12. At the moment of testing they had just started their English lessons. The mother tongue of the monolingual children was Dutch, the bilingual children spoke Turkish and Dutch. The children were matched on socio-economic status on the basis of a standard classification of their father's profession. They were matched on intelligence with the help of the non-verbal Raven's progressive matrices. To determine the degree of bilingualism, the Turkish-speaking children were asked to do a standard editing test for Dutch and cloze test for Turkish. Furthermore, the Turkish teachers and the Dutch teachers of the children were asked to evaluate the children's performance in Turkish and Dutch, respectivily, with the help of four categories describing their command of the two languages. For both monolingual and bilingual children we determined whether they spoke any language apart from Dutch, or Dutch and Turkish. As a result, we were left with twenty "truly" monolingual children and twenty children who, having a fair command of both languages, may be called "balanced" bilinguals.

Material

The test contained a Dutch and an English part, the Dutch test being used as a check whether the two groups scored in a comparable way on a language they were both supposed to have mastered. The choice on the items included in the Dutch test was based on experience from foreign language teaching. For back vowels, the most difficult phonemic contrasts for Dutch learners of English lie in the area between close and half close, and half close and half open. For front vowels, they lie around the line dividing the open area from the half open area. The contrasts used in the test can be inferred from Figure 3 and 4.

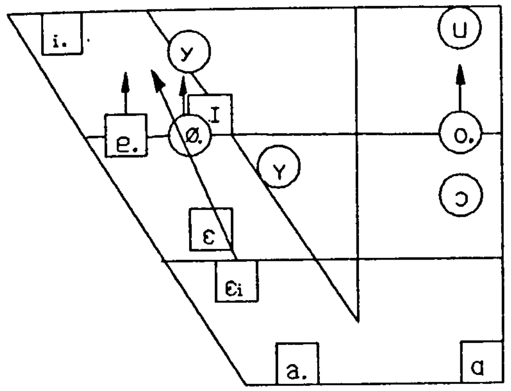


Figure 1. Dutch Vowel Diagram. Circles and boxes stand for rounded and unrounded vowels, respectively. The number of dots refer to phonemic duration. Arrows indicate direction and extent of diphtongization.

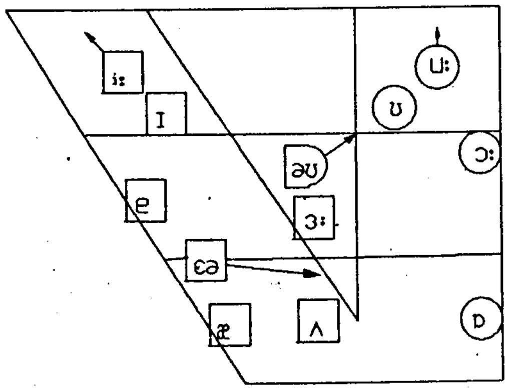


Figure 2. English Vowel Diagram. For futher details, see caption Figure 1.

The test made use of an ABX paradigm in which a nonsense word X had to be compared with two nonsense words A and B, in which X did not have to be acoustically identical to A of B, there being two speakers for each language. For each phonemic contrast there were 8 items depending on the order in which the phonemes were contrasted, whether X was like A or B, and the speaker of X.

Method

The test was administered to individual children with the help of a cassette recorder and headphones. The numbers 1-80 were typed on four consecutive sheets followed by the sequence A B ?. The children had to draw a circle around the A or the B, depending on the similarity they thought to have discovered between the third word and A or B. They were forced to give an answer, even if they heard no difference at all. Both the Dutch and the English part contained a number of trial items.

Results

There was no effect of the order of presentation (t = 0.74, df = 38, p = 0.462). Whereas there was no difference between the number of correct answers on the Dutch test (t = .51, df = 38, p = 0.613), the bilingual children scored significantly higher on the English test (t = 2.86, df = 38, p = 0.007). Both the monolingual children and the bilingual children scored higher on the Dutch test than on the English test (*mono :* t = 8.25, df = 19, p > 0.001; *bi :* t = 6.51, df. = 19, p < .001). The scores of the individual phonemic contrasts can be inferred from Figures 3 and 4.

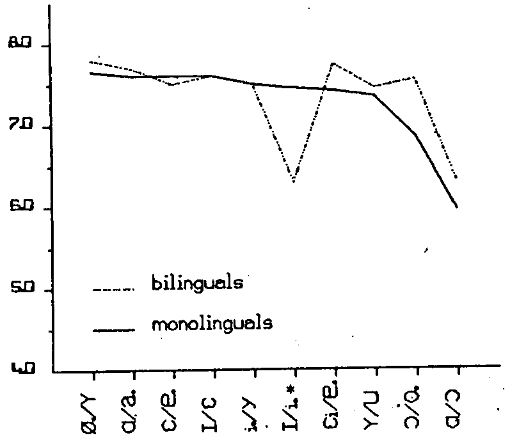


Figure 3. Mean scores for the Dutch test on the 8 items per contrast. An asterisk indicates a significant t-value for the difference between the two groups of listeners.

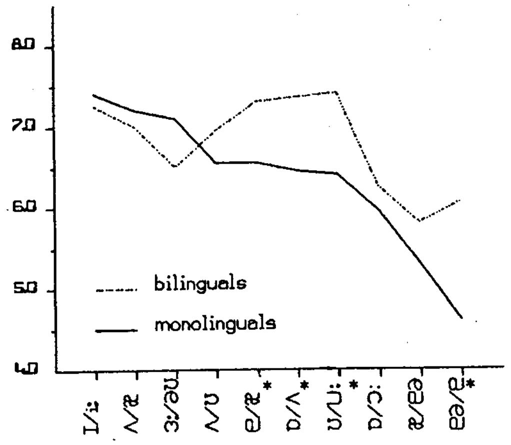


Figure 4. Mean scores for the English test. For further details, see caption Figure 3.

In the Dutch test, two contrasts appeared to be more difficult than the other items. For the monolinguals these were /ɑ/-/ɔ/ and /ɔ/-/o./ for the bilinguals they were /ɑ/-/ɔ/ and /I/-/i./ In a comparison of the phonemic contrasts between mono- and bilinguals, the contrast /I/-/i./ appeared to have been more difficult for the bilinguals, the contrast /ɛi/-/e./ for the monolinguals. In the English test, both mono- and bilinguals seem to have perceived the ten contrasts as three relatively easy ones, three relatively difficult ones and some that were neither the one nor the other, the relatively difficult ones being the same for mono- and bilinguals.

Discussion

The fact that the contrast /ɑ/-/ɔ/ in Dutch proved difficult for both groups of listeners shows that its low scores are not related to the linguistic history of the testees. For the other contrasts, this influence cannot be dismissed offhand.
There is some literature available on the production and perception of Dutch vowels by Turkish adults who are monolingual [4], [5], [6]. Here, all data indicate that the contrast /I/-/i./ is difficult for these subjects. It may be that our bilingual listeners suffered from interference from their first language.
Why the English phonemic contrast /I/-/i:/ is not significantly more difficult for the bilinguals than for the monolinguals may be ascribed to the fact that in English duration is a more important phonemic contrast than it is for Dutch [7]. We have no explanation why the contrast /ɔ/-/o./ should have been relatively difficult for Dutch listeners, nor why the contrast /ɛi/-/e./ should have received higher scores from the bilingual listeners than from the monolinguals. It may be that the bilingual listeners are more alert to the diphthongal character of /ɛi/.
In the English test, the monolinguals made a large number of mistakes in the areas known to cause problems (see section Material). The three most difficult items were the same for both groups, although the monolinguals scored significantly lower than the bilinguals on one of them. The low scores on the contrast /ɜ:/-/əʊ/ for bilingual listeners may perhaps be explained from the absence of central vowels in the Turkish vowel diagram [8]. Any attempt at explaining the relative ease of the contrast between /ʊ/-/u:/, /ɒ/-/ʌ/ and /e/-/æ/ on the basis of a comparison of the Turkish and English vowel diagrams fails, however.
It would seem that bilinguals are better able to distinguish between the sounds of a foreign phoneme inventory. That their high scores cannot be explained with reference to their first language argues in favour of the hypothesis that they perform better because they have become aware of the fact that a sound falling within one phonemic area in one language may fall outside it in another language.

REFERENCES:

[1] Diaz, R.M., Thought and Two Languages: The impact of Bilingualism on Cognitive Development, in: Gordon, E.W., <<Review of Research in Education>>, 1983, 10, 23-55.

[2] MacLaughlin, B., <<Second Language Acquisition in Childhood>>, 1984, Vol. 1.

[3] Balke-Aurell, G., and T. Lindblad, <<Immigrant children and their languages>>, Research Bulletin, University of Gothenburg, 1983.

[4] Boeschoten, J.A. van, Intelligibility of sounds in isolated Dutch words spoken by Turks, in Bennis, W.U.B., and M. van Lessen-Kloeke, <<Linguistics in The Netherlands>>, Foris Publications, Dordrecht, 1984, 23-33.

[5] Boeschoten, H.E., and L. Verhoeven, Integration niederländischer lexikalischer Elemente ins Türkische - <<Linguistische Berichte>>, 1985, 98: 347-364.

[6] Heuven, V.J. van, Some acoustic characteristics and perceptual consequences of foreign accent in Dutch spoken by Turkish immigrant workers, in: Oosen, J. van, and J. Snapper, <<Proceedings of the Colloquium on Dutch Linguistics>>, Berkeley Linguistic Society, 1986.

[7] Gussenhoven, C., and A. Broeders, <<The pronunciation of English; a course for Dutch Learners>>, Wolters-Noordhoff-Longman, Groningen, 1976.

[8] Demircan, Ö., <<Türkiye Türkcesinin ses Düzeni Türkiye Türkçesinde sesler>>, Ankara University Press, 1979.

I am grateful to Vincent van Heuven, who persuaded me to go on with this line of research when my enthusiasm was flagging, and to Roeland van Hout, who helped me in methodological matters.
Rik Boeschoten and Ludo Verhoeven developed the cloze test.

THE INFLUENCE OF TARGET-LANGUAGES ON VOCALIC SPACE IN TEN-MONTH-OLD INFANTS

B. de Boysson-Bardies
Lab. de Psychologie Expérimentale

L. Sagart
C. R. L. A. O.

P. Hallé
Lab. de Psychologie Expérimentale

C. N. R. S. - E. H. E. S. S.
54, bd Raspail 75006 - PARIS

We have hypothesized that incipient linguistic differenciation in babbling children could be first reflected in vowel production. Twenty ten-month old infants from Paris French, London English, Hong-Kong Cantonese and Algiers Arabic backgrounds were recorded in the cities of origin. 1047 non-nasalized vowels extracted from syllables containing at least one consonant were spectrally analyzed. Formant frequencies were plotted on F1-F2 charts by infant and by language group. Statistical analyses provide evidence of an early differenciation in vowel production between infants from different language backgrounds.

The hypothesis that prelinguistic productions are unrelated to any specific language is supported by similarities found in the statistical distribution of consonants in the productions of infants from different linguistic backgrounds. We postulate that articulatory gestures for vowel production could offer the first indication of linguistic differentiation in prelinguistic infants.

The development of vowel space has been shown to be a gradual process. However,at 10 months, children's vowel spaces have reached a relatively large extension (1) (2) and there is a relative stabilisation of the anatomical configuration of the vocal tract around 38-40 weeks . We thus decided to investigate whether language-specific effects on vowels could be shown to exist in the babbling of 10-month old infants. We investigated the vowels of ten-month old infants from French, English, Algerian and Cantonese linguistic backgrounds.

The two positions about babbling in relations to language background entail different predictions:

1. If the babbling forms depend mainly upon biological mechanisms and maturational processes, there should be no systematic differences between infants and the distribution of vowel formant frequencies should be independent of the language of the environment.

2. On the contrary, if by 10 months infants have already begun to be influenced in their productions by the language of their environment, we may expect to find systematic differences between children as a function of the differences existing between the corresponding adult languagês in the domain of vowel production.

Vowel production in the adult languages

Frequency counts of vowels in running speech (3-4-5) show English and Cantonese to be sharply contrasted in the frequency of occurence of certain vowel sounds. Specifically the following feature ratios are markedly higher for English than for Cantonese: front/back, high/low, unrounded/rounded. In each case French stands between English and Cantonese. In terms of the distribution of vowel formant frequencies in running speech these preferences entail the following:

- Cantonese prefers high F1's and low F2's (compact vowels).
- English shows the reverse preference and favors low F1's and high F2's (diffuse vowels).
-On both the F1 and F2 dimensions, French stands between English and Cantonese.

No acoustic description or phoneme frequency count is available for Algiers Arabic. Given the general information on Maghreb Arabic we may conjecture that Algiers Arabic may favor more central values of F1 and F2,possibly with a bias, due to pharyngealization, toward higher central values for F1 and lower central values for F2.

If similar preferences are found in the babbling of infants that will support the claim of an early influence of target languages on babbling.

Recording Procedures

Five ten-month old infants were separately tape-recorded during a single one-hour session in each of the following cities: Paris, London, Algiers and Hongkong. Children were paired for age and sex across language groups.

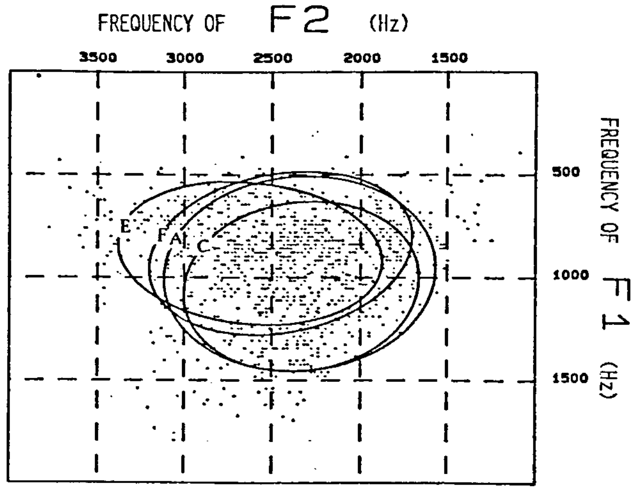
Acoustic analyses.

The tapes were transcribed into narrow IPA transcription by one transcriber. These transcriptions were then simplified by reducing the number of vowel symbols to nine. The frequency of occurrence of each wide vowel class in the babbling productions of each child were then established. Acoustic analyses were run on non-nasalized vowels from canonical or variegated babbling (6). For each babbling utterance, one token of each of the different vowel symbols was analysed such that the same repartition of vowels in the acoustic analyses as found in the wide transcriptions was maintained.
The audio signal was analogically low-pass filtered at 4.5 KHz, then sampled at a rate of 10 KHz. After visually displaying the resulting signal, a relatively steady-state portion of 30 msec. was selected for each vowel to be analyzed. F1 and F2 were estimated from 20Hz resolution short-time spectra cepstrally smoothed by means of the "True Envelope" method (7).The first and second formants were plotted at the corresponding frequencies on F1xF2 formant charts.

Results and statistical analysis.

The overall ranges of formant variation for all children are approximately 400-1800 Hz for F1 and 1250-3800 Hz for F2.
All vowels by all twenty subjects are plotted in the formant chart in fig.1. The ellipses enclose 75% of the vowels in each of the four language groups.

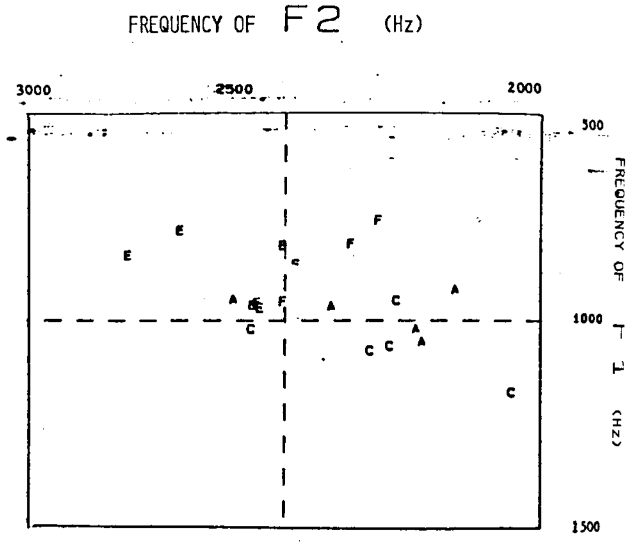
Fig.1.



F1-F2 plot of all vowels by all infants. Ellipses enclose 75% of the vowels in the French (F), English (E), Cantonese (C) and Algerian (A) groups.

Taken together for each child, mean F1 and mean F2 define a "mean vowel" which lies at the center of the area of the formant chart occupied by the vowel productions of a child. The "mean vowels" of each child are shown in fig. 2.

Fig. 2



Plot of mean F1 and F2 frequencies in Hz by subjects. Mean vowels of French, English, Cantonese and Algerian infants are marked F,E,C,and A.

The mean vowels of five infants - 2 French, 2 English and 1 Algerian - are very close together, with formant frequencies close to F1 = .95 KHz and F2 = 2.55 KHz. With the rest of the children, the "mean vowels" occupy different sectors of the formant chart as a function of language background.
Mean formant frequencies and standard deviations by language group are listed in Table I

Table I

	French	English	Cantonese	Algerian
F1	878	876	1047	976
	(239)	(207)	(241)	(282)
F2	2456	2628	2343	2341
	(439)	(452)	(381)	(459)

An overall analysis of variance (ANOVA) was run on these data with Language as factor and mean F1 and mean F2 as observations. The effect of language environment was found to be significant for F1: F = 6.651 (3,16) p<.01, and for F2 as well: F = 4.728 (3,16) p<.05.

A second set of analyses was performed based on estimated distances between vowel sets. If we assume that the vowel sets of infants all belong to a single population, in other words, that there are no significant differences between infants from different language backgrounds (hypothesis H0), we can then characterize that population by a covariance matrix computed on all the data, and characterize the distance between the vowel sets of any two infants by means of Mahalannobis distances:

$$d_{ki,lj}{}^2 = \Delta\Sigma_a^{-1}\Delta^T$$

where k and l are indexes for the languages, i and j indexes for infants within a given language, and Δ is the vector of differences between F1 means (i=1 to 2). Average intra-language distances are computed according to:

$$(1) \qquad D_k = (\sum_{i=1}^{4}\sum_{j=i+1}^{5} n_{ki}n_{kj}\,d_{ki,kj}{}^2)/(\sum_{i=1}^{4}\sum_{j=i+1}^{5} n_{ki}n_{kj})$$

where n_{ki} is the degree of freedom of the corresponding vowel sets and Dk the average intra-language distance within the kth language. Similarly, the averaged inter-language distance between the kth and lth languages is given by:

$$(2) \qquad D_{k,l} = (\sum_{i=1}^{5}\sum_{j=1}^{5} n_{ki}n_{lj}\,d_{ki,lj}{}^2)/(\sum_{i=1}^{5}\sum_{j=1}^{5} n_{ki}n_{lj})$$

Table II summarizes the results under the H0 hypothesis. Clearly, infants differ more between different languages than within any single language. This precludes H0 and in itself warrants the conclusion that infants' vowels in babbling differ according to language background.

Table II

Intra-language and inter-language distances under H0

	French	English	Cantonese	Algerian
French	0.31	0.83	0.86	0.83
English		0.38	2.00	1.99
Cantonese			0.45	0.86
Algerian				0.42

The mean intra-language distance is 0.392; the mean inter-language distance is 1.231.

We also tested the hypothesis H1 according to which several populations -possibly one per language- are to be considered. The kth language is characterized by the covariance matrix Σk. Intra-language distances are

$$d_{ki,kj}{}^2 = \Delta\Sigma_k^{-1}\Delta^T,$$

while inter-language distances are obtained by averaging two different covariance matrices, e.g.:

$$d_{ki,lj}{}^2 = \Delta\Sigma_{k,l}^{-1}\Delta^T, \text{ where } \Sigma_{k,l} = (n_k\Sigma_k+n_l\Sigma_l)/(n_k+n_l),$$

and n_k is the degree of freedom of the entire vowel set of the kth language.

Averaged intra- and inter-language distances are computed as in (1) and (2). Table III summarizes the results obtained under H1: these results are still consistent with the claim of an early influence of linguistic environment.

Table III

Intra-language and inter-language distances under H1.

	French	English	Cantonese	Algerian
French	0.29	0.47	1.00	0.51
English		0.46	1.36	0.80
Cantonese			0.60	0.46
Algerian				0.40

Mean intra-language distance is 0.440 and mean inter-language distance is 0.767.

Specific influence of target languages.

English group.
F1 values for English infants are generally low (mean is 876 Hz, lower than with the Cantonese and Algerian groups) and little dispersed. Individual English children depart only slightly from this general tendency. This illustrates the preference of English children for diffuse vowels as predicted.
Cantonese group.
Mean F1 in Cantonese children is higher than in any other group and mean F2 is lower than in any other group. The preferred formant frequencies of Cantonese infants are those characterizing compact vowels.
French group
As expected,mean F2 is higher than with Cantonese infants and lower than with English infants. Mean F1 (878 Hz) is also lower than mean F1 with the Cantonese and Algerian groups but identical with that of English children.
Algerian group
The distribution of formant frequencies is indeed characterized by a preference for central frequencies: the mean formant values of F1: 976 Hz and F2: 2346 Hz, although the vowel space of Algerian infants as a whole is not noticeably less spread out than that of the other groups.

Conclusion

Our acoustic study of formant frequencies in vowel productions by 10-month olds has shown that certain characteristics of the target-languages are reflected in the productions of babbling.
When speaking about production most investigators frequently refer to constraints on performance only. However in production as well as in perception the programs involved require underlying representations. With respect to production, the preference of babbling children for vowels situated in an

acoustic space that is consistent with the yowel space of the corresponding adult language indicates that target-language oriented articulatory procedures begin to be available. In an earlier study (8) we claimed that setting articulatory limits to tongue and lip movements in vowel production in babbling could be the first step towards acquiring the vowel system of a target language. This implies that the notion of "representation for production" must be considered for 10-month olds.
According to Locke (9) language acquisition begins "when a child moves away from what would continue to be his pattern and closer to the ambient one". In the present article we have shown that the buildup of target language-oriented articulatory skills is already under way at the end of the first year.

References

(1) Lieberman,P.,(1980) On the development of vowel production in young children. In G.H. Yeni-Komshian, J.F. Kavanagh,& C.A.Ferguson (eds),Child Phonology, vol 1 Production. New-York: Academic Press 113-142.

(2) Kent, R.D. & Murray, A.D. (1982) Acoustic features of infant vocalic utterances at 3,6 and 9 months. Journal of Acoustic Society of American.72, 353-365.

(3) Wioland, F., (1972). Estimation de la fréquence des phonèmes en français parlé. Travaux de l'Institut de Phonétique. 4, 177-204.

(4) Delattre, P. (1965) Comparing the phonetic features of English, German, Spanish and French. Heidelberg: Julius Groos Verlag and Philadelphia: Chilton Books.

(5) Fok, A. (1979) The frequency of occurence of speech sounds and tones in Cantonese. In R. Lord (Ed.), Hongkong Language papers. Hongkong: Hongkong University Press.

(6) Oller, K. (1980) The emergence of the sounds of speech in infancy.In G.H. Yeni-Komshian, J.F. Kavanagh & C.A.Ferguson (eds),Child Phonology, vol 1 Production. New-York:Academic Press 93-112.

(7) Imai,S. & Abe,Y.(1979) Spectral envelope extraction by improved cepstral method IECE vol.J62-A (4), 217-223.

(8) Boysson-Bardies B.de, Sagart,L.,Halle P &Durand C. (1986) Acoustic investigation of cross-linguistic variability in babbling.In B.Lindblom &R.Zetterstrom (eds),Precursors of Early speech Werner-Green International Series, Vol 44, 113-166.

(9) Locke, J.L. (1983) Phonological Acquisition and Change New-York:Academic Press.

PHONETIC AND PHONOLOGICAL PROPERTIES OF CHILD LANGUAGE

JAROSLAVA PAČESOVA

Department of Phonetics
Brno University
602 00 Brno, Czechoslovakia

ABSTRACT

The present paper attempts to be a small exploratory study in phonology, adducing data from the child's acquisition of his mother tongue, i.e. Czech, with the intention of contributing to our understanding of phonological organization and phonological processes in language development. Particular reference is made throughout to motivating examples from "substitution paths" which children follow in acquiring the phonemes, especially as far as the nature of distinctive features, the active organizational role of marking in determining the structure of child language, the strategies in phonology acquisition and the interrelations between language levels are concerned.

INTRODUCTION

The bulk of linguistically oriented research on child phonology since Jakobson's "Kindersprache"has mostly been concerned with discovering a universal order of phonemic development:/1/. The principle of maximum contrast and various priorities, e.g. that of unmarkedness as against markedness, stop as against fricative and semiocclusive, front as against back and simple as against complex nuclei formations have repeatedly been dealt with in studies in this field:/2/.

The position of a phoneme in a phonemic system is, however, determined not exclusively by possible oppositions but also by the extent to which alternates with other phonemes and by realization of its potential to be an exponent of linguistic meaning. One kind of individual variation is the apparent chance following of alternate paths in the acquisition of particular sounds:/3/.

Our study is based on the data dealing with the linguistic development of a Czech-speaking boy recorded since his first utterances at the age of ten months and ending at two-and-a-half years when his phonological system was well established and language use quite fluent. The data were collected daily and notes were made of the context of use. In analyzing them, systematic confrontation with specialized literature and with results gained during our longitudinal research in language development of fifteen children - aged from one to three years, most of them being the boy's playmates at the nursery home - was done.

The observation in the development of consonants have let us to take the position that there is a consistent phonemic patterning within the speech of each child which is often strictly idiosyncratic. A widely recognized principle is exemplified in our study: that of progress from a simple beginning to greater complexity: /4/. An attempt is made to illustrate this at the phonetic and phonological levels, showing at the same time how such development is related to increases in vocabulary size, awareness of grammatical system and amount of language use. In a brief paper of this nature it is not possible to give a full description of the child's phonetics and phonology. There - fore, a restricted aspect, viz. the acquisition of fricative phonemes - with special view to the lateral and vibrants has been selected for consideration.

It is generally acknowledged that at the earliest stage of language development, the child's vocabulary is very small and plosives, nasals and vowels are the only sounds used. Most of them are familiar from babbling and their perception and production is thus well practiced. The continuant sounds, on the other hand,require more skill not only in perceptual discrimination but also in production.The articulators have to be in a posture of close approximation to achieve friction, neither completely in contact as in plosives and nasals, nor well clear of each other as in vowels. The appearance of fricatives - at certain stage of phonologic development - is, nevertheless, not seen as resulting primarily from a production difficulty but as arising from the boy's increased perceptual discrimination, the need for their production not having arisen until the child is able to discriminate and to recognize them as functional. In our subject, first attempts at the friction and sibilance were made with varying degrees of success. Friction was achieved at the following points of articulation: labial (= bilabial IφI and labiodental IfI, the former having predominance, in spite of the fact that it does not exist in Standard Czech); palatal (= IjI) and velo-glottal (= velar IxI and glottal IhI). Sibilance, on the other hand, was produced at the alveolar, postalveolar, palato-alveolar and palatal areas (= IsI, IšĭI, IʃI and IʃI, of these only /s/ and /ʃ/ have their counterparts in adult language).

The new sounds, however imperfect, brought new consonantal contrast into use so that a higher level of complexity was reached and, simultaneously, the new contrasts resulted in new word structures. In addition to the already established structures such as plosive+vowel (e.g. Ibu:I), vowel+plosive (IopI), plosive+vowel+nasal (IbumI), there were now also nasal+vowel+sibilant (ImiʃI), plosive+vowel+sibilant (Iba:sI), plosive+vowel+fricative (IbafI), fricative+vowel+ nasal (IhamI) and fricative+vowel+fricative (IhafI).

As for the frequency of use with fricatives, only /j/ can compete with the frequency of plosives and nasals. And it is the substitutive function of /j/ which accounts for most of its occurrences, cf. the fact that the observed child utilized /j/, simultaneously, as a substitute for five fricative phonemes, viz./v/,/z/,/3/, /l/ and /r/.The child's phonological system had thus a phoneme with very high functional load, the degree of its integration in the fricative subsystem was, on the other hand, minimal: having been used as a substitute for five fricative phonemes, it was not opposed to them by means of related distinctive features. With gradual mastering distinctive features of the so far substituted phonemes viz. labiodentality in /v/, alveolarity and sibilance in /z/, postalveolarity and sibilance in /3/, alveolarity and laterality in /l/ and alveolarity and vibrativity in /r/ and /ř/ - the phoneme /j/ lost its former phonemic territory and became a fully integrated phoneme opposed to the above mentioned phonemes by its otherness:/5/.

Interesting steps in the phonological maturation process can be seen in the boy's mastering the fricative phonemes /l/, /r/ and /ř/. The former two, commonly classified as liquids, in Standard Czech split in two subcategories, viz. the lateral /l/ and the vibrant /r/. They differ in various properties, the most important being the kind of conjunction of closure and opening: while in the lateral the closure and opening occurs simultaneously but in different places, in the vibrant these two attributes alternate but occur in one and the same place, i.e. at the point of alveolars. As mentioned before, both /l/ and /r/ were - at earliest stages of phonology development - replaced by the phoneme /j/: in terms of features, both the lateral and vibrant character were ignored and so was the point of articulation. The next output, the palatalized /l'/, signalled the child's mastering the feature of laterality as well as the contrast simple vs. complex nuclei in fricative articulation. By additional loosing the palatal character and the obstruence the output became an intrinsic allophone of the Standard phoneme /l/, distributively, however, it still represented the two liquids namely, /l/ and /r/. The output /r/- identified with the Standard phoneme /r/- was then telescoped by the additional feature of having become a discontinuant vibrant opposed to continuant lateral, and, at this stage, the demand of language system, viz. the semantic contrast should be encoded phonologically, was fulfild.

A glance at the boy's dealing with the syllabic allophones of the two liquids that are firm elements of the phonemic system of Standard Czech, displays another interesting phenomenon in the process of phonology development. At the stage in which the boy has not yet mastered the feature of laterality and in which, naturally, the even more difficult feature of vibrance is also absent, he uses the vowels /u/ and /e/ as substitutes for both the syllabic Il̩I and syllabic Ir̩I thus showing his awareness of the feature of syllabicity. Nevertheless, a question arises as to what makes him use two vowels in substituting one syllabic allophone and whether the alternation of the two substituting vowels is purposeful or merely accidental. A deeper insight into the boy's behaviour reveals rather surprising phenomenon: the alternation takes place even within the inflected or derived forms of an identical word unit, cf. Ipuʃi:I = prší (it rains) with IpeʃejoI = pršelo (it rained); IvunaI = vlna (wool) with IveɲenejI = vlněný (woollen). This observation of ours certainly runs counter to the boy's habit in the domain of morphology where alternation - in Standard Czech obligatory - is ignored by him and the preservation of the same vowel (or consonant) is one of the most typical feature of his early phonotactic system. For illustration cf. the following examples: the child forms Iku:ɲI = nom. (horse) - Iku:ɲaI = accus.; IbloukI = nom.sg. (beetle) - IbloukiI = nom.pl. with the corresponding adult forms Iku:ɲI - IkoɲeI, IbroukI - IbroutsiI. From what has been said follows

that what seems to be universal at the morphological level does not hold good at the phonological level. Instances such as IketʃekI = nom. (neck) - Ido kutʃkuI = accus. (into the neck), IvunaI = noun (wool) - IveɲenejI = adj. (woollen) bear evidence of this. The discrepancy might be, in our opinion, accounted for in the following way: at that stage of language development where morphophonemics has the upper hand, the child pays little attention to morphology including the origin and function of inflections and derivations. The acquisition of word forms takes place at first by processing each item as an individual unit. Their phonological patterning, deletions or substitutions of phonemes correspond to adopting the strategy which seems fundamental in the communicative act, namely the application of the principle of least effort and maximum economy in articulation. This shows up in various types of assimilations (vowel or consonant harmony being the most frequent), contact and distant dissimilations, metatheses and other phonetic changes. Systematic simplifications both in vocalic and consonantal contrasts phonetic variations and instability in the proper distribution of the phonemes are the next markers which reveal the, as yet, non matured developmental stage.

In analyzing the acquisition of the phoneme /ř/, many a peculiarity stands out. Before dealing with its implementation in the observed child, a few comments should be made on this consonantal phoneme as far as its realizational and distributional characteristics in the adult speech is concerned.

As probably known, Czech is the sole of the Slavonic languages which has not only developed this phoneme but has also preserved it as a fixed, though structurally isolated element of its phonemic system. The structural isolation shows both from the phonetic and phonemic point-of-view. After much speculation about classifying this sound from both the articulatory and acoustic standpoints it was ranked as a trill with two allophones, voiceless and voiced, occurring in complementary distribution:/5/

The first output of this no doubt most unusual and difficult consonant in the observed child was the phoneme /c/, i.e. the voiceless palatal plosive utilized for both the voiceless and the voiced allophone of the phoneme /ř/.

The lenition process, viz. the mastering of the binary feature change /-voice/ - /+voice/ had the occurrence of the pho - neme /ɟ/ as its result. Hence both allophones of /ř/ found their substitutive sounds. Unlike the other voiced phonemes which - due to neutralization of the feature of voice word-finally obligatory in Standard Czech - are devoiced in this position, with /ɟ/ its devoicing also in initial and intervocalic position is not exceptional. This fact fits well with inherent physically caused variation in speech, namely the tendency to devoice is the greater the smaller is the surface area of the oral cavity, that is, the less is the capacity to absorbe the glottal air-flow:/6/.

The next step on the substitution path was the application of the phoneme /ʃ/ in place of the voiceless Ir̥I and of the phoneme /3/ in place of the voiced IřI - an evidence of the fact that the child had already mastered the feature of fricativity. The feature of voice, too, seems to be well established - with few exceptions concerning the lesser stability of the voiced member word-initially. As next followed the output of intrinsic allophone IrʃI and Ir3I, telescoped by the additional feature of vibrativity. And, finally, when IrʃI merged with the voiceless allophone Ir̥I and Ir3I with the voiced allophone IřI, the phoneme /ř/ enters the child phonemic system, being opposed to the other trill, viz. /r/ by its stridency. The adult specific realization of the voiced allophone initially, intervocalically and when preceded by the voiced consonant, and, conversely, the realization of the voiceless allophone in the unvoiced neighbourhood and word-finally, was adopted by the child as one of the obligatory phonological rules in accordance with adult language system:/7/.

We have mentioned before that interrelations between languge levels may play an important role in our understanding and explaining the child language. Here are further examples to illustrate the fact: the phoneme /x/ is, evidently because of difficulty concerning velar fricativity, changed into /s/ by a process of fronting, cf. the child's realization InestsuI instead of the proper InextsiI (I don't want). An alternative realization of this rather difficult consonantal cluster in the same child is InektsuI - where a process of stop formation and dissimilation takes place. The application of /k/ in substituting ItsI and ItʃI in the child's forms IklukiI = nom.pl. (boys) and Ikluki:kI = the diminutive form of the same noun, viz. IklukI = nom. sg. instead of the proper forms IklukI - IklutsiI - Iklutʃi:kI is, on the other hand, not only an illustration of the child's preference of stop articulation as against semi-occlusive articulation but also confirmation of the fact that he has not yet gained the awareness of the palato-alveolar formation as an index of morphological plural formation in the first instant and of diminutive formation in the second instant. On the other hand, his active approach in mastering the grammatical sys-

tem is manifested. He does not passively borrow word forms from adults but creates his own plurals and diminutives (and of course also other inflections and derivations within given grammatical categories) in accordance with his own grammar the typical feature of which is the high degree of regularity and avoidance of exceptions.
The next specimen, viz. the realization of /x/ in the plural form IuxiI instead of the correct IuʃiI (ears) is explainable not only on the ground of the boy's preference of regular formation, viz. IuxoI = nom.sg. - IuxiI = nom.pl. instead of the proper IuxaI, but, simultaneously, but also of his ignorance of the adult convention, that is the Standard restriction of the word form "ucha" to ear-shaped things, especially handels of a vessel, versus the word form "uši" used in reference to the organ of hearing.
To sum up, the child language system is different both for quantity of information and for its organization from that of adult language system at all its levels, and as such should be interpreted.

In concluding our paper we would like to emphasize that - unlike earlier models of phonological acquisition which assumed that the child either passively awaits the maturation of physiological control system /8/ or passively waits untill he can limit or suppress the natural processes reducing his output to pabulum /9/ - we take the phonology development as an extremely creative process in which the child "intentionally" partakes and uses a variety of strategies as his guide: /10/. The fact that he, at one moment, pronounces a phoneme correctly and in its proper position, does not mean that from that time onwards he definitely masters the phoneme. It has been said elsewhere that the acquisiton of the phonological system is a process. In this place it should be stressed that this process involves steps both forwards and backwards, although, of course, the general trend is progressive. The child's language system as a whole is for a certain period in a state of flux and fuzziness with a fairly strong element of unpredictability as regards phonemic distinctions. And, not exceptionally, the actual distinction between two phonemes may be in free variation with the lack of this distinction. There is no doubt, still a long way to go in establishing and in testifying a detailed model of phonological acquisition. Further studies of a larger number of children and in relation to different languages showing the interrelations of development at various levels of lanaguage are needed in order to get deeper information on what we have tried to deduce from the language behaviour of a Czech-speaking child.

REFERENCES

/1/ R. Jakobson,"Kindersprache, Aphasie und allgemeine Lautgesetze". In: Selected Writings I., Phonological Studies. The Hague, 1962, 328-401

/2/ J. Pačesová, "The Development of Vocabulary in the Child", Brno, 1968

/3/ N. Waterson, "Growth of Complexity in Phonological Development". In: The Development of Communication, N. Waterson and C. Snow eds., New York, 1978, 415-442

/4/ Ch.A. Ferguson, "Phonology as an Individual Access System: Some data from Language Acquisition", Academic Press, Inc., 1979, 189-201

/5/ W.U. Dressler, "A Semiotic Model of Diachronic Process Phonology". In: Perspectives of Historical Linguistics, W.P.Lehmann and Y.Mankel eds. Amsterdam, 1982, 93-131

/6/ J. Vachek, "Situace ve fonologickém podsystému souhlásek: Problémy kolem fonému /ř/". In: Dynamika fonologického systému současné spisovné češtiny. Academia Prague, 1968, 92-102

/7/ J.J. Ohala, "The Application of Phonological Universals in Speech Patology". In: Speech and Language: Advances in Basic Research and Practice. Vol. 3, N.J. Lass ed., New York, 1980, 75-97

/8/ M. Romportl, "On the Czech System of Consonants". In: Studies in Phonetics, Academia Prague, 1973, 105-117

/9/ M. Templin, "Certain Language Skills in Children: Their Development and Interrelationships. Minneapolis, Univ. Minnesota Press, 1957

/10/ D.Stampe, "The Acquisition of Phonetic Representation". Papers from the 5th Regional Meeting of the Chicago Linguistic Society, 1969

/11/ G.Drachman, "Generative Phonology and Child Language Acquisition". In: Phonologica, München, 1972 235-251.

DEVELOPMENT OF CHILD SPEECH HEARING AT THE ONSET OF SPEECH

YELENA ISENINA

Dept. of English Language
Ivanovo State University
Ivanovo, USSR 153025

Abstract

The results of the investigation of speech hearing indicated the ability of children aged from 18 to 24 months to recognize the words before they are able to differentiate the phonemes; the general tendency of phoneme differentiation revealed by N.H. Shvachkin is supported; the influence of the acoustic and speech motor analysers over the formation of phonemic. hearing depends on individual differences in the development of one of them in the child.

Introduction

There are two important problems in the investigation of child speech hearing at the onset of speech: the order of phoneme differentiation and the influence of the speech-motor and/or acoustic analysers on the development of the phonetic hearing.
The stages of phonemic hearing development were revealed by N.H. Shvachkin /7/. The drawbacks in the methods used were criticized by O. Garnica and the results were partially confirmed in the experiments of M.G. Edwards /9/. That is why the problem was open to investigation.
There are three points of view regarding the influence of speech motor and auditory analysers on the development of phonemic hearing. The adherents of the acoustic theory of speech perception suppose the phonemic hearing development to be based on the operations of the auditory analyser /8/. The supporters of the motor theory believe the formation of the phonemic hearing is impossible without the activity of articulatory organs /1/. Some scholars believe that only the interaction of hearing and articulation define the formation of phonemic hearing /7/.Before describing our experiments devoted to these two problems it is necessary to define the concept "speech hearing" differentiating it from the concept "phoneme differentiation"or "phonemic hearing". Speech hearing includes: 1. phoneme differentiation ability, 2. speech recognition. The mechanisms of differentiation and recognition that underlie these abilities in adults were investigated by N.I Zinkin /3/ and E. Esenina /4/. While differentiating one singles out all the differential features of an object (the phonemes of the word, for ex.), the relations between them and in this way the image of an object is formed. The process of recognition is based on the image which has been already formed and "makes use" of some features referring to an object as a whole (the word structure features, length, different features of some phonemes).

Experiments

In our first preliminary experiment /6/ the general problems of child phoneme differentiation and recognition were investigated. The problems were: to find adequate methods of infant speech hearing investigation; to find out the relation between child phoneme differentiation and word recognition; to find out what sounds in the words of different length are the most informative for recognition. Two groups of children: (3 children in each group), speaking and non-speaking at the age from 1,8 to 2 years took part in the experiment.
The results pointed to the absence of difference in compared groups both in the number of phonemes differentiated and in the number of words recognized. The features used in recognition changed according to the word context. The words had been recognized before all the phoneme differentiations were achieved. The most informative for the recognition were: the stressed vowel, the vowel after the stressed syllable, the first sound. (The same results were obtained with the grown-ups /5/.
In our next preliminary experiment /6/ the succession of phoneme differentiation and the recognition process were under investigation. The subjects were 4 non--speaking children and 5 speaking children at the age of about 2 years old. The experiments confirmed the succession of phoneme differentiation achieved by N.H. Schvachkin. The experiment also suggested that non-speaking children could achieve the same level of phoneme differentiation as speaking children. While not being

able to differentiate phonemes the children used acoustic features in the recognition of 49 names of objects.

In the concluding experiments /6/ the following problems were under investigation: the ability to differentiate phonemes by speaking and non-speaking infants to follow the role played by motor and hearing analysers in phonemic hearing; the succession of phoneme differentiation in the groups of speaking and non-speaking infants. We also wanted to find out whether the time of the first word of non-speaking infants depended on the previous success with phoneme differentiation.

We supposed (according to N.H. Shvachkin /7/) that the differentiation of one and the same phoneme by different children may be based either on the acoustic or upon the motor analyser depending upon the individual degree of its development. But there may be some regularities of the interaction of both analysers that depend upon non equal difficulties in the pronunciation or recognition of different sound groups. In order to define these regularities we supposed that three sequences should be compared: 1) the sequence of phoneme differentiation /7/; 2) the sequence of articulatory differentiation which is obtained from the observation of the children's sound pronunciation acquisition /1/; 3) the sequence of the recognition of the same pairs of sounds by speaking children which was obtained in the experiments with some noise interference /1/.

The sequence of acoustic and motor differentiations was defined by the order of sound appearance in the pronunciation of the cild or by the order of its recognition. This was done on the supposition that after the given child had mastered the pronunciation of the latest sound in the pair, or had recognized it we would take the differentiation (acoustical or motor) of this pair of sounds for granted. The relative order percentage of phonemic, acoustic and motor differentiation acquisition of the same pairs of sounds is reflected in diagram N 1. (p. 3). Judging from the data in the diagram, some conclusions can be drawn about the interaction of acoustic and motor differentiations which influence the formation of the phonemic ear.

1. In the case of phonemes [р, с, з, ш, ж/the articulatory differentiation lags behind the auditory very considerately. That means that this phoneme differentiation may be based on some acoustic differences in the sounds which calls forth some articulatory changes leading to the articulatory differentiation of these sounds. In other cases difficulties in the pronunciation lead to a lag in phoneme differentiation of soft and hard sibilants (in Russian), sonorants and non-articulatory voiceless sounds.

2. In the case of phonemes /п, т, к, б, д, г, в, б', в', м/ the development of the phonemic ear is based upon their articulation, with acoustic differentiation following it. It should be born in mind that these tendencies are likely to change under the influence of the child acoustic and motor individual development. For the purpose of testing some of these suppositions, the following experiment has been carried out.

Subjects. 17 children aged from 1 year 5 months to 1 year 9 months. Four from 5 speaking children of the control group could produce sentences. The experimental group consisted of 12 non-speaking children who could pronounce only: "мама папа, бах, биби, дай". All the children had parents speaking only one language (Russian). Their hearing was normal. They were on the 6th stage of sensory-motor intellectual development (according to Piaget). Their ability to understand speech and to speak was checked up. Our hypothesis was that the group of non-speaking children bases their phoneme differentiation on the acoustic properties. of the speech sounds given in diagram N 1. The experiment lasted for two and a half months.

Material. In the experiment 35 control cards and 28 experimental cards were used. The control cards presented pictures of objects familiar to the children. The exp. cards presented objects having monosyllabic names. and differing only in one phoneme. They reflected the 9 stages of phonemic development given by N.H. Shwachkin /7/.

Procedure. First, the Exp. taught the children to recognize and differentiate the control cards shown to the child in chance order (3 or 4 at a time) until they all could recognize and point to the card (when asked) very quickly and without any mistakes. After this the Exp. passed to presenting control pairs of cards. A pair of cards differing in one phoneme was put before the subject. Four commands had to be fulfilled with each word. The commands were: "Point to..., give it, put it into the box, take it out of the box". Each pair was presented 8 times. Every subject had to differentiate 4 or 5 pairs of words at each experiment. Every child participated in the experiment once a day for 20 minutes, five times a week.

Discussion of the Results. In O.K. Garnica's experiments /10/ only the fact of the differentiation on non-differentiation of the phonemes of the pair was

1. The acoustic qualities of the sound include not only phonemic but all other characteristics.

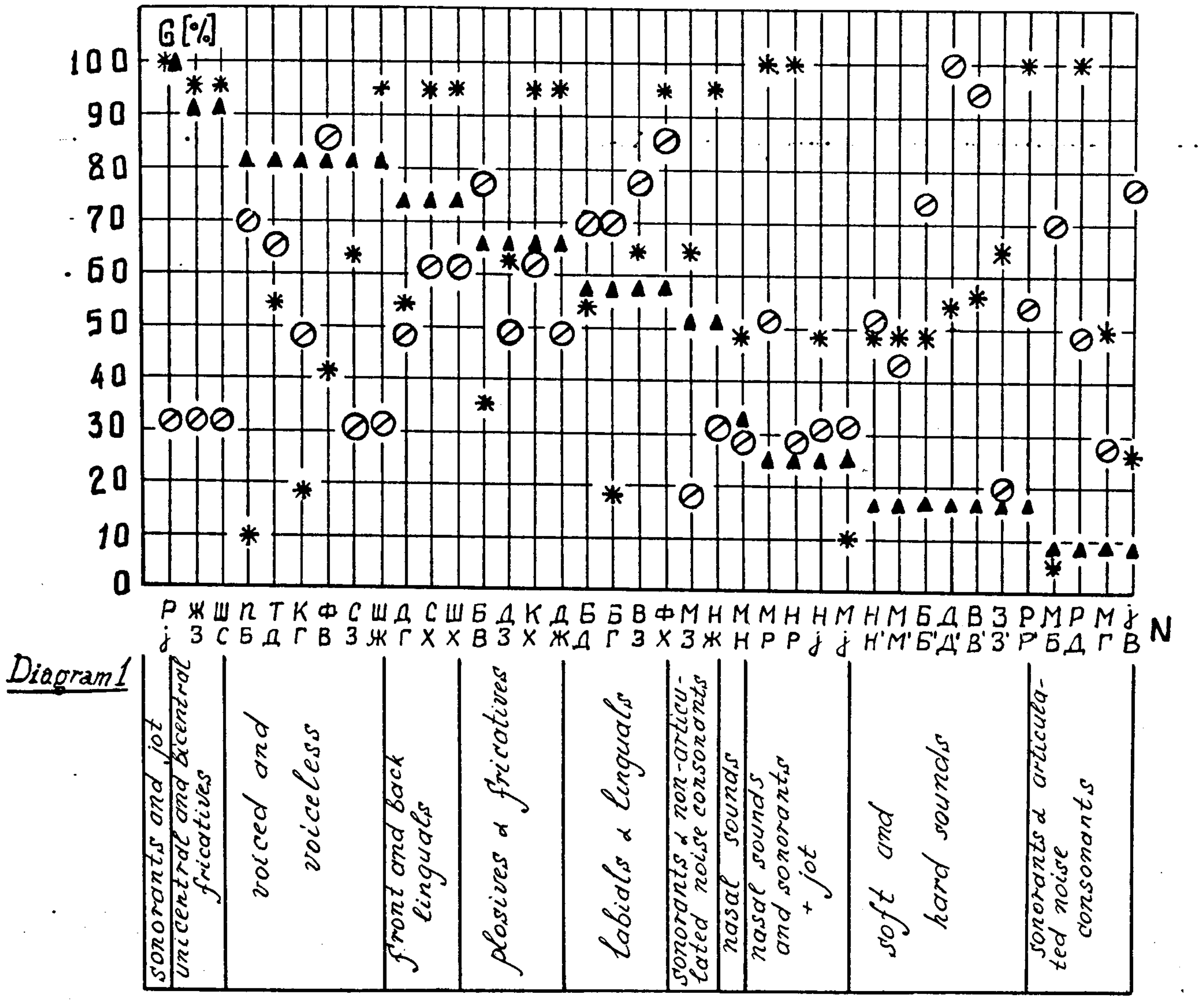


Diagram 1

✱ - motor differentiations

▲ - phonemic differentiations

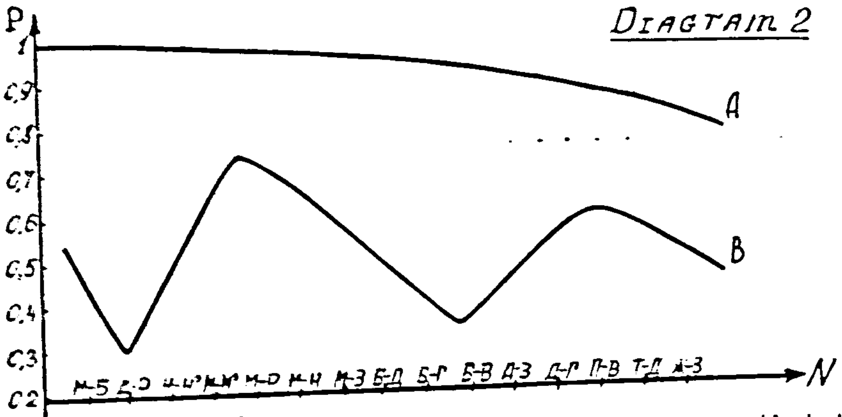
⊘ - acoustic differentiations

G - relative order percentage of phonemic, acoustic, motor differentiations .

N - phonemic, acoustic and motor differentiations

into account. We supposed that the figures showing the ratio of the number of the subjects' correct choices to the number of all the choices which define the degree of phonematic differentiation can also be taken into account.

Diagram N2 (p.4) shows the dependence of the probability of correct choices by every subject on the stages of Shvachkin's scheme of phonemic development (19 pairs). According to their results all the subjects were divided into two groups: with low probability of correct choices (less than 50% - nine subjects of the exp. group; and with high probability of correct choices (more than 50%) - 8 subjects,3 from exp. group and 5 from contr. group. The data of three experimental and five contr. subjects is shown on the same upper curve A of a monotonous character. The results show that the success of the first 8 points of phonemic differentiation come close to 95%-100%. From the 9th point an almost linear lessening of the differentiation of the given phonemes is observed. The same held true for vowel-phonemes. This proves Shvachkin's scheme of phonemic development. The results also show that three non-speaking subjects differentiated all the given pairs of phonemes as well as all the speaking subjects. For the other 9 subjects of the exp. group, the curve of differentiation -B has a polyextremal character, with periodic rises



and falls in the distinguishing of some phonemes.

The Spearman correlation range coefficient between the probability of phoneme differentiation and the order of acoustic (CKA) and motor (CKM)differentiating was calculated. CKA=0,54 – correlation is statistically meaningful. CKM=0,01 (statistically non meaningful. That means that the better is the recognition of the most difficult sound of a pair, the higher is the probability of phoneme differentiation of this sound for the given non-speaking subject.

In some months after the experiment we asked the parents of our non-speaking subjects about the time they began to speak. We calculated the Spearman coefficient between the children level of phonemic differentiation and the number of months which were necessary for them to begin speaking. It appeared to be 0,72 –the correlation was statistically evident.

Conclusion

Summing up our results of child speech hearing investigation the following conclusions can be drawn: before being able to differentiate the phonemes the child can recognize the words properly, this recognition is evidently based on non-phonemic acoustic features of the word; the most informative for recognition elements are: the stressed vowel, the vowel after the stressed syllable and the first sound of the word; the general tendency of phoneme differentiation coincides with that discovered by N.H.Shvachkin; speaking children may differentiate phonemes better than non-speaking ones, it points to the positive influence of articulation over phoneme differentiation; non-speaking children can differentiate phonemes quite well making use of acoustic but not motor characteristics of the phonemes; the time necessary for non-speaking children to master the pronunciation of the first words in positively correlated with the level of phoneme differentiation, that is one can see the positive influence of phoneme differentiation over the articulation of non-speaking children.

That is why we can say that the development of hearing and speech motor analysers is of euristic character. The part these analysers play in the formation of phonemic hearing changes in connection with what plays the leading part (develops faster) in the individual development of the child: his articulation or phoneme differentiation.If the hearing and motor analysers develop simultaneously, the development of the phonemic hearing may depend on the difficulties in motor and acoustic differentiations of sounds.

References

1. Бернштейн С.И. Вопросы обучения произношению. М., I937.
2. Бельтюков В.И., Салахова А.Д. Об усвоении ребёнком звуковой (фонемной) системы языка. – Вопросы психологии, I975, № 4, с. 7I-80.
3. Жинкин Н.И. Механизмы речи. М., I958.
4. Исенина Е.И. К вопросу о формировании образа слова. – Вопросы психологии, I967, № I, с. 5I-65.
5. Исенина Е.И. О признаках слова, необходимых для его узнавания при слушании. – Новые исследования в педагогических науках. Вып. XII. М., I968.
6. Исенина Е.И. Дословесный период развития речи у детей. Саратов, I986.
7. Швачкин Н.Х. Развитие фонематического восприятия речи в раннем возрасте. – Изв. АПН СССР, вып. I3, I948.
8. Jacobson R. Child language, aphasia and phonological universals. Mouton,1968.
9. Edwards M.G. Perception and production in Child phonology. The testing of 4 hypothesis. – Papers and reports of Child language development, 1974, April, p.67-84.
10. Garnica O.K. The development of Phonemic Speech perception. – In: Cognitive Development and the acquisition of language. N.Y. 1873, p.215 – 222.

ON SPECTRAL DIFFERENCES IN THE CRY OF NEWBORNS OF DIFFERENT NATIONALITIES

ZURAB N.JAPARIDZE
Institute of Linguistics
Tbilisi, Georgia,USSR
380007

YURI A. STRELNIKOV
Institute of Linguistics
Tbilisi, Georgia, USSR
380007

IGOR Y. STRELNIKOV
Georgian Polytechnical
Institute, Tbilisi,
Georgia, USSR 380075

ABSTRACT

A statistically reliable difference in the cry of newborns of different nationalities is revealed. This is observed both auditorily and through a spectrum analysis and is assumed by the authors to be connected with the specificity of the mother tongue of ancestors. Newborns are supposed to be inheriting certain features of the articulatory basis of this tongue.

Experiments on perception revealed a statistically reliable difference in the cry of newborns of different nationalities. The Georgian auditors were to listen in pairs to the cry of the Georgian and Russian newborns and asked to mark which - the first or the second - of the two stimuli (an isolated cry) was closer to the Georgian vowel a.
Similar experiments showed that the auditors could be assigned (with necessary breaks) to make not more than 100 comparisons. Therefore two groups of stimuli for two different series were made with 100 oppositions of the cry of newborn children of the same sex but of different nationalities. In order to reduce the influence caused by the position of stimuli in pairs on perception of similarity, in 50 cases of each series first on the list came the cry of a newborn of one nationality while in 50 other cases that of other nationality. Naturally the auditors would not know the cries of which - Russian or Georgian - newborns they were to asses.
The study of the auditors' records (6400 assesments) revealed that the auditors had more often found the cry of the Georgian newborns closer to the Georgian vowel a than that of the Russian newborns. This was noted in the cry of both boys (36 auditors) and girls (28 auditors). None of the auditors found the cry of the Russian newborns closer to the standard vowel more often than that of the Georgian newborns. The cry of the latter was considered to be closer to the standard vowel in 76 cases out of 100 on the average and 81-82 cases with individual auditors. Such a constantly one-way deviation being several times greater than the standard deviation of ±5 allows to assert with a high degree of reliability that there is a considerable nonrandom difference (/1/, at greater length /2/) in the studied stimuli of the Georgian newborns in one aspect and those of the Russian newborns in another aspect.
However, the actual picture might have been slightly changed in the assessment of the auditors due to their tiredness, distraction, illusive perception or other factors that occur when an analysis is carried out not by a machine but by a man. The aim of the present investigation was the instrumental study of the spectral structure of the sounds under review and matching up of the revealed differences with the picture of their perception and formation. For this purpose a band-pass spectrum analysis of newborn cries through 30 band-pass filters over the range of 80 to 8300 Hz was carried out. Over 1000 Hz the spectrum constituents were boosted by 6 dB per octave as is usually done in phonetic investigations. As is known, a negligible base frequency difference between sounds in a band-pass spectrum analysis may lead to considerable amplitude discrepancies on band-pass filter outlets. To avoid this the spectrum was smoothed out with the given frequency band range substituted by the arithmetic mean of the given and several neighboring values. Account of values from 7 adjacent channels turned out to be optimum for coping with the task: the values of the given band and three lower and three higher bands were taken account of. In the case of the first and last three bands the arithmetic mean could be naturally deduced only from the smaller number of the bands.

Prior to statistical data processing the intensity of every cry had been brought to one and the same value and computed as a sum of square values from each frequancy band without boosting by 6 dB per oct.

A total of 322 cries was analyzed out of which 173 cries belonged to the Russian newborns and 149 cries to the Georgian ones.

By use of Willcockson's criterion which is applicable to any kind of distribution a difference between the compared groups was elucidated and rated on statistical reliability.

The analysis allowed to assert with a degree of reliability of 0.99 that there is a nonrandom difference between these groups (i.e. the cries of the Russian newborns on the one hand and those of the Georgian newborns on the other hand) in a number of bands. Namely such a difference was observed in frequency ranges 80-100 Hz, 300-1500 Hz, 1900-2700 Hz, 5200-8300 Hz. Out of these in the first and third ranges the expectations were greater in the cries of the Georgian newborns while in the second and fourth ranges they were greater in the cries of the Russian newborns.

Such a picture complies sufficiently well with the perception data described above. The two seperated (second and fourth) frequency ranges 300-1500 Hz and 5200-8300 Hz in which the values were greater in the cries of the Russian newborns revealed a definite similarity to the first two formants of the diffusive vowels e, i. As to the spectrum of the cries of the Georgian newborns, it has greater values in the middle third range and is closer not to the diffusive but, on the contrary, to the compact vowel a (in comparison with the cries of the Russian newborns).

The difference in the low-frequency range band 80-100 Hz deserves a special mention. This band is situated lower than the base frequency of the stimuli and its vibrations cannot, at first glance, be related with the cry of newborns. However, the spectrum constituents of different intensity in this band observed in the cries of newborns both nationalities are supposedly related with the formation of sound vibrations in the back of the oral resonator: with a friction noise in the narrowed passage in the velar and laryngeal regions and with low-frequency vibrations of the soft palate and the uvula. The most favourable conditions for such vibrations are created through a rather backward placement of the tongue that is permissible for the Georgian newborns. A possible influence of glottalization should also be considered.

The above-mentioned assumption is connected with our earlier suggestion that individual features of the articulatory basis (articulatory habits) of mother tongue is somehow transmitted to newborns. At least by the moment of birth they are observed to possess certain features of such a "starting placement" of the speech organs that is convenient for passing in the acts of speech to the pronunciation of the sounds of the given language /1/, /2/. But what "starting placement" of the speech organs may be characteristic for the Russian and Georgian languages? A great functional load of the contradistinction of soft and hard consonants in Russian and its absence in Georgian, and the realization of more advanced and raised vowels in Russian in comparison with the Georgian ones make different starting placements more convenient for the Russian and the Georgian speakers: a more advanced and raised placement of the tongue for the Russians and a more retracted and lowered placement of the tongue for the Georgians. Such a difference in placements of the speech organs in the cry of newborns might lead to a greater similarity of the cry of the Georgian newborns to the a-type vowel, i.e. to the difference ascertained both in the spectrum of the stimuli and in their perception /1/, /2/.

It is not quite clear how could the features of the starting placement of the speech organs be transmitted to newborns. This requires further investigations. But it should also be noted that the difference in the cry of newborns of different nationalities cannot so far be satisfactorily linked with any other nonlinguistic factor. The assumption that this difference could be related to the anatomical difference in the structure of the speech tract was not confirmed by specialist investigations. The anatomical difference could but cause a converse effect: the cry of the Russian newborns should have been closer to the vowel a but not that of the Georgian ones /3/.

The obtained results of the investigation allow to assert with certainty a timbral difference in the cry of newborns of different nationalities. This is confirmed by data of both auditory and spectrum analyses and may be related, as stated above, to the phonetic specificity of the mother tongue of ancestors of the newborns.

/1/ Z.N.Japaridze, Y.A.Strelnikov, On Linguistic Characteristics of the Cry of Newborns. XXIV Scientific session of the Institute of Linguistics. Working plan and report theses (in Georgian), Tbilisi, 1978.

/2/ З.Н.Джапаридзе, Ю.А.Стрельников, О различиях в плаче новорожденных разной национальности и пола, Сб.Экспериментально-фонетический анализ речи, Ленинград, 1984,

/3/ М.Г.Абдушелишвили, З.Н.Джапаридзе, Ю.А.Стрельников, К оценке влияния антропологических факторов на характеристики плача новорожденных, Сб.Проблемы фонетики и фонологии Материалы всесоюзного совещания (ноябрь 1986 г.) Москва - 1986.

CONTINUOUS VARIATION OF THE VOCAL TRACT LENGTH IN A KELLY-LOCHBAUM TYPE SPEECH PRODUCTION MODEL.

Hui Yi WU   Pierre BADIN   Yan Ming CHENG   Bernard GUERIN

Laboratoire de la Communication Parlée (ICP, UA CNRS 368)
E.N.S.E.R.G. - I.N.P.G.
46, av. Félix Viallet - 38031 GRENOBLE Cédex, FRANCE.

ABSTRACT

The KELLY-LOCHBAUM reflexion-type line analog (K-L model) is a temporal speech production model which has the advantages of a low computational cost and of a simple and clear physical interpretation. But it has an important drawback : it is not designed to handle a continuous variation of the vocal tract length. In this paper we present a strategy to solve this problem : the vocal tract length variation is dealt with as a variation of the working sampling frequency and then this variable sampling frequency is converted into a constant one.

The sampling frequency conversion is achieved by means of a time-varying FIR filter, designed to minimize the computational cost. The performances of the algorithm are evaluated with simple sinewave signals and with synthetic vowels. Finally, since for a practical application we should use frames within which the sampling frequency is constant, we extend our algorithm to a frame to frame basis and solve the problem of the FIR filter definition when moving across frame boundaries.

INTRODUCTION

In human phonation, the phonological informations are encoded by both vocal tract configuration and its dynamic variations. For synthesis purposes, a vocal tract configuration is described as an area function, including naturally the information of vocal tract length. Several models have been proposed for the acoustic simulation of the vocal tract in the time domain (/2/, /4/). For computational cost reasons, we use a KELLY-LOCHBAUM (K-L) reflection-type line analog to develop our research. Since the K-L model and even its recent developments (LILJENCRANTS, 1985) do not take into account the possibility of variation of the vocal tract length, we have made a attempt to develop this feature. In the first section of this paper, we present the basic idea for the spatially continuous length variation of the vocal tract, i.e. sampling frequency conversion, and we test the method. In the second section, we extend this algorithm to a frame to frame based temporal variation.

1. SPATIALLY CONTINUOUS VARIATION OF THE VOCAL TRACT LENGTH

1.1 The Problem

Since the vocal tract length vary rough between 16 and 19 cm during speech, we need include this feature in any vocal tract acoust simulation. For a K-L type of vocal tract tempor simulation (or improved versions, /3/), all t tubes have the same length (spatial sampling step and the sampling frequeney of the temporal sign produced is inversely proportional to this length. continuous variation of the vocal tract length c be achieved by a continuous variation of the tub length around a given value, which leads to related variation of signal sampling frequenc Therefore, if we wish a signal sampled with constant frequency, we need a system to convert t signal sampled with the variable frequency into signal sampled with the constant frequency. In th first section, we reduce the problem to the me conversion from a constant input sampling frequen F_i to a constant output sampling frequency F_o.

1.2 The Sampling Frequency Conversion

In a classical way (/1/), we decompose th sampling frequency conversion into two steps : firs we convert the discrete input signal x(n) sample with F_i into a continuous signal $x_c(t)$, and then w sample this continuous signal with F_o.

To reconstruct $x_c(t)$ from x(n), we only nee to low-pass filter x(n) with a cutoff frequency F equal to $F_i/2$. The theoretical formula c interpolation by an ideal low pass filter is :

$$y_c(t) = \frac{1}{F_i} . \sum_{n=-\infty}^{+\infty} x(n).h(t-nT_i) \qquad (1$$

where

$$h(t) = 2.F_c . \frac{\sin 2\pi F_c t}{2\pi F_c t} \qquad (2$$

is the impulse response of the filter, an $T_i = 1/F_i$. This impulse response beeing infinit (I.I.R.) and non causal, we need to approximate th filter by a F.I.R. (Finite Impulse Response) filte by using a windowing function w(t). This leads to a certain distortion of the frequency response of th filter.

In the second step, we need to sample y(t) at F_o, and thus, to avoid aliasing we need to insure that F_c is lower both than $F_i/2$ and than $F_o/2$. Then we obtain the formula :

$$y(m) = \frac{2F_c}{F_i} . \sum_{n=N1}^{N2} x(n).w(mT_o-nT_i) . \frac{\sin(2\pi F_c.(mT_o-nT_i))}{2\pi F_c.(mT_o-nT_i)} \qquad (3)$$

where w(t) is the windowing function, and N1 and N2 are determined as functions of F_i and F_o, and of the length of the window. Finally, the global system described by eq. (3) is a time-varying low-pass digital filter (/1/), implemented as a F.I.R. filter. The method of windowing the impulse response of an I.I.R. filter for F.I.R. filter design provides the avantage that the F.I.R. length, and thus the computational cost of the filter, can be easily and independently varied. At the same time, it makes it easy to compute the coefficients of the filter for each frame.

We know that the type and the length of the window used influences the properties of the filter. Therefore we need to evaluate quantitatively this influence.

1.3 Evaluation of the Transformation

In order to evaluate the performance of the sampling frequency conversion, we have made tests with sinewaves of different frequencies, and with synthetic vowels generated by our K-L line analog.

Sinewaves analysis

The influence of the transformation on a single sinewave has been analyzed : for different fundamental frequencies, two sinewaves with the same fundamental frequency, amplitude and phase, S_{Fi}, sampled with the system input frequency F_i, and S_{Fo}, sampled with the system output sampling frequency Fo have been generated. Then S_{Fi} has been converted into S'_{Fo} by the system, and finally the following parameters have been compared for S_{Fo} and S'_{Fo} : (1) the difference of amplitude between the sinewaves, (2) the difference of phase, and (3) the Signal/Distortion (S/D) ratio.

Because of the nature of the low-pass filter, an undulation is introduced in the pass band of the filter transfer function : it is always smaller than ±1 dB, which can be considered negligeable. Since the window we use is symetric around the origin point, the impulse response is symetric and thus a linear phase filter is insured : the transformation has no effect on the signal waveshape.

As expected, the S/D ratio increases with the window length. An informal analysis (by visual inspection of the FFT spectrum of S'_{Fo}) has shown that the distortion is mainly due to harmonic components corresponding to frequencies such as $F_0 + n.(F_o-F_i)$ or $F_0 + n.(F_o-F_i)/2$, where F_0 is the frequency of the sinevawe, and that the non-correlated noise is very much below this distortion. Therefore the S/D ratio is defined as the ratio between the energy measured in a 300 Hz band centered on the sinewave fondamental frequency and the energy outside this band (up to 5 kHz). Fig. 1 shows the evolution of the S/D ratio as a function of the number of points for the window, for a rectangular and for a Hamming window, for two different sampling frequency conversions. For short windows (i.e. 4-5 points), there is less scattering in the S/D ratio for a rectangular window than for a Hamming window, and for longer windows, the opposite phenomenon happens : we conclude that rectangular windows lead to better results than Hamming windows for short windows, and that Hamming windows give better results for longer windows.

Synthetic vowel analysis

The transformation has also been tested with vowels synthesized with our K-L model. The signals for the synthetic vowels [a], [i] and [u have been converted into signals sampled with various frequencies ; the corresponding spectra (obtained by the Cepstrum method) have been compared with the original spectra visually and by means of a "distance" defined by :

$$D = \sum_{N=1}^{1024} \frac{A_{dB}(N\Delta F) - A_{dBref}(N\Delta F)}{1024} \qquad (4)$$

where the missing points of $A_{dB}(N\Delta F)$ are evaluated by linear interpolation (since the frequency steps for the two spectra are different, due to different sampling frequencies). On the curves (see example in Fig .2) we can see that the system retains the formant characteristics very well, the errors appearing mainly in the spectrum valleys.

The error measured by eq. (4) converges toward a non zero value when the window length increases, depending on the vowel configuration and on the sampling frequency change. Since we know that for a long window the error must be very small, we conclude that this bias is due to our "distance" and to the linear interpolation, and we normalize the results in relation to this convergence value for each case. Fig. 3 shows that finally, there is no obvious difference between a rectangular and a Hamming window. In every case, there is a rather abrupt decrease of the scattering of the normalized error for windows longer than 4 points : we conclude that a rectangular window with 5 points is optimal.

2. FRAME TO FRAME TEMPORAL VARIATION OF THE VOCAL TRACT LENGTH

2.1 The problem

In real speech, the length of the vocal tract varies continuously with time, for instance in transitions from rounded vowels to unrounded ones. Following a classical approximation, we suppose that the vocal tract area function is constant during a short time interval, and thus its length. This is the basis for a frame by frame simulation of the vocal tract in most of the models. Thus, we should apply the frequency conversion developped in the previous section on a frame to frame basis : the input sampling frequency F_i must be considered constant for each frame, but may vary from one frame to the next one, according to the length variation of the vocal tract. This leads to the problem of the realization of the filter defined by eq. (3) when the window w(t) overlaps the boundary between two frames.

2.2 The solution

We first try to solve theoretically the problem for a system with only two input sampling frequencies. We suppose that the input signal is given with a sampling frequency F_{i1} from $-\infty$ to 0, and with a sampling frequency F_{i2} from 0 to $+\infty$. Thus, we can suppose that the corresponding

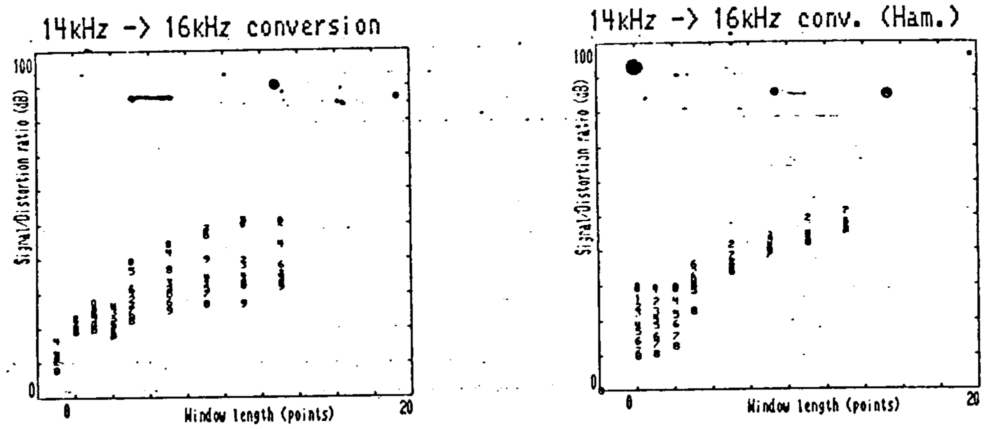


Fig.1 . Signal/Distortion ratio against window length (expressed as a number of points) for sinewaves wi frequencies ranging from 50 Hz (symbol 1) to 4.5 kHz (symbol 9) by 500 Hz steps (Ham. = Hamming windowing, otherwise rectangular windowing).

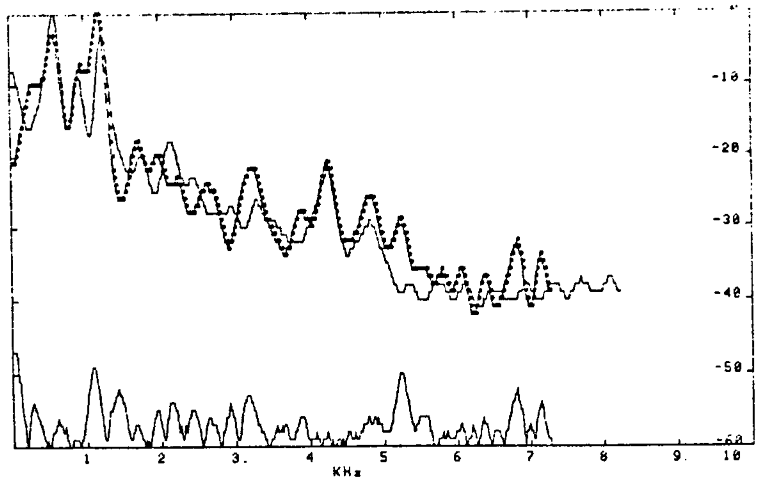


Fig.2 : Spectra of a synthetic vowel /a/ (continuous line, F_i=16.55kHz), of the signal resulting from th transformation (dotted line, F_o=14.55kHz), and difference of the spectra (bottom line).

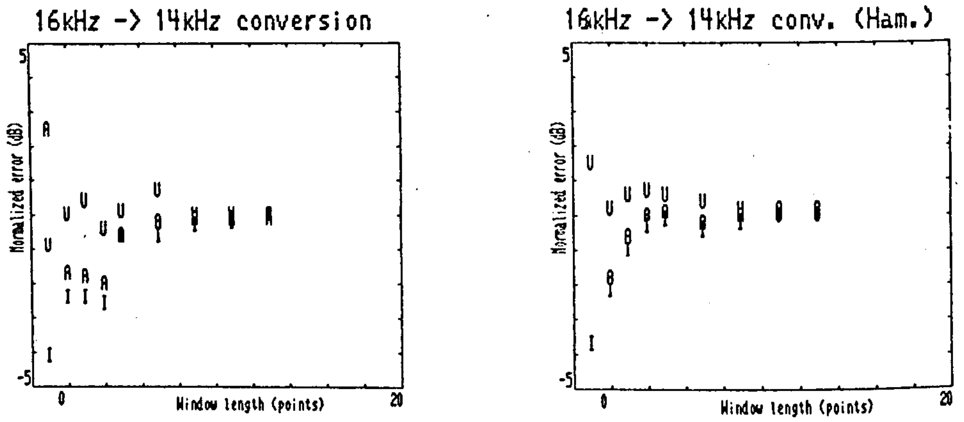


Fig.3 : Spectral error against window length (expressed as a number of points) for 4 vowels (Ham. = Hamming windowing, otherwise rectangular windowing).

continuous input signal $x_c(t)$ has been decomposed, by an appropriate-step windowing into two signals $x_{c1}(t)$ and $x_{c1}(t)$ defined by :

$$\text{for } -\infty < t < 0,\quad x_{c1}(t) = x_c(t) \text{ and } x_{c2}(t) = 0,$$
$$\text{for } 0 \leqslant t < +\infty,\quad x_{c1}(t) = 0 \text{ and } x_{c2}(t) = x_c(t).$$

Then, we suppose that $x_{c1}(t)$and $x_c2(t)$ are sampled respectively with F_{i1} and F_{i2} into $x_1(n)$ and $x_2(n)$. Thus, we can apply eq. (3) to these two discrete signals, with the corresponding $2.F_c/F_{i1}$ and $2.F_c/F_{i2}$ factors,. and then sum up the results to reconstruct the complete signal, knowing that the filtering is a linear operation. In the case where the window is entirely included in one of the two frames only, eq. (3) applies directly. Otherwise, taking into account the instants where x_1 and x_2 are zero, we obtain the following equation :

$$y(m)=\frac{2F_c}{F_{i1}}.\sum_{n=N1}^{NO} x(n).w(mT_o-nT_{i1}).\frac{\sin(2\pi F_c.(mT_o-nT_{i1}))}{2\pi F_c.(mT_o-nT_{i1})}$$
$$+\frac{2F_c}{F_{i2}}.\sum_{n=NO+1}^{N2} x(n).w(mT_o-nT_{i2}).\frac{\sin(2\pi F_c.(mT_o-nT_{i2}))}{2\pi F_c.(mT_o-nT_{i2})} \quad (5)$$

where NO is the last sample of the first frame and NO+1 the first sample of the second frame.

We see that eq. (5) means that for output samples close to the frame boundary, it is just needed to take into account the contributions from the samples in the two frames with the suitable coefficients. Nevertheless, we understand that the abrupt step windowing defined above will introduce some spectral distorsion in the two filters, which leads practically to some distortion for the signal near the boundary. We can anyhow check that eq. (5) reduces to (3) in the case where $F_{i1} = F_{i2} = F_i$, which could be expected.

The problem beeing solved for one boundary, we can extend the method to the frame to frame basis mentioned above, as far as the window length is shorter than the frame length, to avoid to include two boundaries in the same window. We have chosen to keep the cutoff frequency F_c of the interpolation filter independant of the input sampling sampling frequencies : we take the half of the smallest of all the input and output frequencies.

2.3 Evaluation of the method

In this section, we give describe the tests that we used in order to check the validity of the algorithm.

Triangle waveform

In order to evaluate the distortion of the signal at the frame boundaries, we have generated constant frequency triangle waveforms with sampling frequencies varying from one frame to the next one. Then, we have applied our frequency conversion algorithm to obtain a signal with a constant sampling frequency. Thus, whenever a segment of straigth line crosses a frame boundary, it is easy to measure the departure from linearity. We have made this evaluation directly on the processed signal, and also on its second derivative which can be expected to be an series of Dirac impulses corresponding to the inversions of slope in the triangle waveform : when the signal departs from a straigth line, the second derivative is not null any longer and shows up as a noise.

In different experiments, we have shown that the amplitude of the distortion at the boundary :
(1) depends very little on the window length (the length of the segment where distortion appears is longer for longer windows) ;
(2) is roughly proportional to the amplitude of the signal at the boundary ;
(3) is proportional to the sampling frequency variation between the two frames ;
(4) is roughly independant of the input and output sampling frequencies.

In the reality, the vocal tract length does not vary quickly : thus the sampling frequency variation from one frame to the next one never exceeds one or two percent. For these type of variation, the departure from linearity is less than one percent.

Vocalic transitions

Finally, we have elaborated a few vocalic transitions such as [i] → [u], [a] → [u]. The signal produced is high quality, and it is impossible to detect any boundary problems either by listening or by visual inspection of the signal.

We conclude that our algorithm is well suited to our practical speech application.

CONCLUSION

We have shown that it is possible to solve the problem of the spatially continuous variation of the length of the vocal tract by a sampling frequency conversion method. This method leads to good results even with rather short windows (4-5 points). It has been succesfully extended to a simulation based on a frame to frame decomposition. Thus our K-L model is not limited any longer by a constant vocal tract length. The study has been done for a "quasi-static" model : now it is needed to extend our algorithm to a fully dynamic model.

BIBLIOGRAPHY

/1/ **CROCHIERE R.E. & RABINER L.R. (1983)**, "Multirate Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey.
/2/ **KELLY J.L. & LOCHBAUM C.C. (1962)**, "Speech Synthesis", 4th Int. Congr. Acoust., G42.
/3/ **LILJENCRANTS J. (1985)**, "Speech Synthesis with a Reflexion-Type Line Analog", Doctoral dissertation, R.I.T., Stockholm.
/4/ **MAEDA S. (1982)**, "A Digital Simulation Method of the Vocal-Tract System", Speech Comm. 1, 199-229.

NASALIZATION OF FRENCH VOWELS
Contribution of the nasopharyngeal tract and the sinuses

Gang FENG & Christian ABRY

Institut de Phonétique de Grenoble
Institut de la Communication Parlée
38031 Grenoble Cedex, FRANCE

ABSTRACT

In this paper, a complete simulation study on nasalization of 11 French vowels will be presented. All simulations, starting from different oral vowel configurations finally attain a nasality target : the nasopharyngeal tract. We will describe several rules for the pole-zero evolution structure. Afterwards, a study situating the "true" nasal vowels in the nasalized vowel frame will be presented. Finally the effects of sinuses on these simulations is discussed.

INTRODUCTION

The conception of a nasality target has been proposed in a preceding study [1,2]. This target, articulatorily represented by the nasopharyngeal tract, can be characterized by its first two spectral peaks : 300 and 1000 Hz. In order to simulate more realistically these acoustic characteristics, we suggested using a small acoustic equivalent nostril for the nasal tract. It has been shown that this approach can provide a better match for the first spectral peaks by comparing it with the nasal tract (alone) and nasopharyngeal tract sweep-tone measurements.

Since we proposed that all nasal vowels should be considered as dynamic trends towards the nasality target, a complete simulation of transfer functions from oral vowels to their corresponding nasopharyngeal configurations seems necessary : i.e. we should examine dynamically all of the coupling degrees.

In this paper, we will present the simulation results for the nasalization of 11 French vowels. For each vowel, we start from the oral configuration, which corresponds to the highest velum position. Then we change progressively the velum position, representing different coupling degrees. Finally the nasopharyngeal configuration with the velum lowered to the tongue is attained.

Such simulations give us a continuous transfer function evolution for all the French vowels. We tried to show several rules from these simulations that should be useful for analysis and synthesis of nasal vowels.

On the other hand, we tried to situate the "true" nasal vowels in this nasalized vowel frame, especially the nasal vowel [ɑ̃], which is often considered as the most difficult to synthesize.

After examining the simulation results that establish the basic pole-zero evolutions, we shall consider the sinus in these simulations. It will be shown that the main effect of the sinus is to make the spectrum more complex.

1. THE SIMULATION MODEL AND AREA FUNCTIONS

We adopted the classical electrical-line vocal-tract model in our simulations to calculate transfer functions in the frequency domain [3]. The area functions used for the 11 French vowels were taken from BOE [4] and MRAYATI [5], slightly modified by FENG [6].

In our simulations, one crucial problem was to determine the area function in the velum region for different coupling degrees. Without sufficient physiological data, we were obliged to make an approximation, which appeared reasonable. We simply divided the area at the coupling point (i.e. around the extremity of the velum-uvula) into two parts : one corresponding to the input of the nasal tract and the other to the input of the oral tract. For the different velum positions, we chose a series of area-ratio (partial area / total area) as follows : 0.0, 0.025, 0.05, 0.1, 0.3, 0.7, 0.9, 0.95, 0.975, 1.0. The two sections (section length : 1cm) just after the coupling point in both the nasal and oral tracts are then determined by a linear interpolation.

2. THE POLE-ZERO EVOLUTION STRUCTURE

Before discussing real vowel simulations, it should be useful to show here a simplified simulation that will provide a pole-zero evolution structure.

In this simulation, a parallel L-C(-R) circuit is used to study the coupling problem. In the system, each of the branches is composed of a 2-order L-C(-R) circuit, having two resonance frequencies and a zero-impedance frequency. One can adjust the access-coefficient of each branch by modifying its impedance, thus changing the coupling degree of the system. The transfer function of this system is defined by the ratio : the sum of the two output currents / input current.

When the access-coefficients of the two branches are changed continuously, one obtains an evolution of the system transfer functions. Figure 1 is a typical example. The highest and the lowest transfer functions of this evolution correspond to the two extreme situations in which one of the two branches is totally connected and the other cut off. So the system transfer function corresponds to only one of the two branches, having two poles and no zero. Between these two extremities, the transfer functions present three poles and one zero, due to coupling. The middle of the evolution shows maximal coupling since the two branches possess the same access-coefficients.

One can obtain different evolutive images when the parameters of the two branches are changed. It is not difficult to prove that, for this system, all evolutions can be classified in several structures, shown in Figure 2. Here F11, F12 and Z1 denote respectively the two resonance frequencies and the zero-impedance frequency for one branch, and F21, F22, Z2 for the other branch.

Naturally, the real coupled vocal tract model is more complex than this system. However, it has been found that impedance variations of the nasal tract and the oral tract, during different coupling degrees, are similar to the above system. Moreover, we are mainly interested in the first two formant region. So the above pole-zero evolutive structure remains instructive for the study of real simulations.

3. NASALIZATION OF THE 11 FRENCH VOWELS.

We will now examine the simulation results for nasalization of the 11 French vowels. We will begin with the extreme cardinal vowels [i], [u] and [ɑ], since they determine the limits of the vocalic space.

For each vowel, 10 transfer functions have been calculated and presented, corresponding to the series of area-ratios described above (Fig. 3). The lowest one represents the transfer function for the oral vowel and the highest that for the corresponding nasopharyngeal tract. Between these two extremes, figure transfer functions for the nasalized vowels with different coupling degrees.

For [i], starting from the first two formants 240 - 2270 Hz, we attain the nasopharyngeal structure with 240 - 1090 Hz. The evolution firstly shows the appearance of a "shoulder" to F1 for [i], and then this peak finally forms the second pole of the nasopharyngeal tract (the second typical pole). As for F2 for [i], it ends by joining the zero and then disappears. This evolution is very similar to the structure presented in Figure 2-h.

In such a transition, analysis and perceptual knowledge provide for us a domain in where the vowel can be validated as nasal [7].

Concerning the production for [u], the evolution towards the nasopharyngeal tract is carried out like a slight elevation of the second pole, the first one remaining - as is the case for [i] - relatively stable. Between these two poles evolves the first pole-zero pair : this corresponds to the usual structure (e) of Fig. 2.

We can obtain, here also, a rather large domain of validation for nasality [7].

The transition for vowel [ɑ] is in the same category as for vowel [i] (but corresponding to the structure in Fig. 2-d) :

- an addition of a peak, as is the case for [i], but here it concerns a low nasopharyngeal peak ;
- a disappearance of a formant, like for [i], but it concerns the first formant here.

The second formant shifts downwards and becomes the high nasopharyngeal peak.

Sweep-tone measurement data [8] confirms here that the typical vocalic domain for [ɑ̃] (for perception, cf. [9]) is rather close to a configuration presenting a maximum velopharyngeal opening - contrary to [i] and [u] that offer a wider validation domain. These simulation results thus correspond to the fact that the low vowels demand a greater velopharyngeal opening in order to be categorized as nasal vowels (reviewed in [10]).

Disparity of nasal correlates for these three types of extreme vowels - already cited in the literature - may seem rebelious to any attempt in simplifying the articulatori-acoustic correspondence.

Addition of poles (so-called "nasal poles"), a high one for [i], a low one for [ɑ] ; evolution of the second pole, lowering of "F2" for [ɑ], elevation of "F2" for [u]..., all of these effects are coherent only when considered as one tendency towards a single objective : acquisition of two essential nasopharyngeal tract characteristics or of one only if the vowel already possesses the other.

The remaining vowels would be situated between these three extremes, with the following major modifications :

- acquisition of a high pole and disappearance of F2 for [i], [y], [e], [ø], [ɛ], and [œ] ;
- lowering of F1 and/or eleva_tion of F2 for [u] and [o] ;
- acquisition of a low pole, disappearance of F1 and lowering of F2 for [a],

[ɑ] and [ɔ].

To summarize further we notice that we are confronted with two topological criteria : - F1 becomes a low pole, the case in [i,...,œ] and [u,o], or disapppears [a, ɑ,ɔ] ; - F2 becomes a high pole, the case in [a, ɑ, ɔ] and also [u,o], or disappears [i,...,œ].

Consequently, when F1 disappears the vowel acquires a low pole [a,ɑ,ɔ] ; if such is the case for F2, the vowel acquires a high pole [i,...,œ]. These acquisitions are, in the final analysis, the most crucial properties. The topology of these acquired poles thus remains the most useful in categorizing our vowels.

It is certain that a continuous passage from one of these categories to the other is possible : or a discontinuous one if emphasis is made on topological "catastrophies", namely the appearance of a pole on the right or left of a formant (the case of [ɛ] vs. [a], etc.).

4. THF NASAL VOWEL [ɑ̃] IN THE NASALIZED VOWEL FRAME

We shall now examine how the true French nasal vowels are inserted in the nasalized vowel frame. The vowel [ɑ̃] will be taken as an example since it is related to : - the realizations of [ɛ̃], or rather the nasalization of [ə] or [ǣ] [11,12] ; - the realizations of [ɔ̃] : more open than [o] (at times even more than [ɔ]) concerning the tongue but very close to the latter with regards to the lips [13,14].

In its oral part, the area function of [ɑ̃] (according to [14]), is close to [ɔ] as to vocal tract length (protrusion-closion of the lips, lowering of the larynx), but different when it comes to pharyngeal constriction size. [ɑ̃] seems to be different from [ɑ], both for vocal tract length ([ɑ] is less protruded for the two speakers in [14]) and for pharyngeal constriction (the same speakers have a tendency to narrow this part of the vocal tract).

These available radiographic data [14] were converted to area functions for their oral part (using coefficients from [15]). The corresponding transfer functions were then calculated. We present here one transfer function for [ɑ̃] to compare with two oral vowels [ɑ] and [ɔ] (Fig. 4).

The oral configuration of [ɑ̃] presents the smallest distance between the two first poles. The [ɔ] formants are lower than those for [ɑ] due to a different lip opening.

[ɑ] compared with [ɔ] shows a decrease of the constriction area in the pharynx. This explains - the second formant having a great sensitivity in this zone (cf. FANT's nomograms [16]) - the relative lowering of F2. The elevation of F1 seems to be a result of a slight continuous retreating of the constriction (ibid.).

A transition from the oral configuration of [ɑ̃] to its nasopharyngeal one is then simulated with the same procedure as presented in section 3.

Figure 5 shows a similar evolutive structure as those of the nasalized vowels [a,ɑ,ɔ]. A lower pole appears below the first formant, then tends towards the first nasopharyngeal peak. Only the evolution of the zero presents a little difference, having a (g)-type structure in Figure 2. However, the acquisition of a low pole on the left of F1 puts this vowel in the phonetic category where [a,ɑ,ɔ] are already situated.

5. THE EFFECTS OF SINUSES

The preceding similations without sinuses provide a clear image of pole-zero evolutions for the nasalized vowels. But knowing the complexity of the nasal tract labyrinth, it seems unrealistic to neglect the sinus cavities (mainly the maxillaries). Here, we will present a simulation result (Fig. 6) in which the maxillaries sinuses were simulated by a Helmholtz resonance [17,18]. Due to the large variability of sinus sizes, and consequently of their acoustic properties [19], we took just as an example a sinus having a volume of 18cm3 and a resonance frequency of about 650 Hz.

Comparing these results with Figure 3, we can see that the main characteristics of the pole evolution structures are preserved in the simulations with a sinus. But a pole-zero pair, resulting from the presence of the sinus, is added to transfer functions, thus changing slightly their structures. So the main effect of this sinus seems only to add more complexity to the nasal spectrum.

CONLUSION

We have tried to establish a complete pole-zero evolution structure for the nasalization of the 11 French vowels. All evolutions, starting from different vowels finally attain a nasality target : the nasopharyngeal tract. This trend versus the complexe pole-zero evolutive structures allows us to propose a common strategy for the nasalization of all vowels. We can place the "true" nasal vowels for French in this frame. To make the simulations more realistic we took into account the complex effects introduced by the sinuses. But the complexity of the real nasal vowel spectrum seems to demand more elaborate simulations, in which the nasal tract would be better modeled, and the effect of the source (the glottal formant) be taken into account [20].

REFERENCES

[1]. FENG G., ABRY C. & GUERIN B. (1985), How to cope with nasal vowels ? Some acoustic "boundary poles". - Proc. of French-Swedish Seminar on Speech, Grenoble.

[2]. FENG G., ABRY C. & GUERIN B. (1986), The nasopharyngeal tract : A target for nasality. - 12th ICA, Toronto, paper A3-8.

[3]. CHARPENTIER F. (1982), Application of an optimisation technique to the inversion of an articulatory speech production model. - Proc. IEEE ICASSP,1984-1987.

[4]. BOE L.J. (1973), Etude acoustique du couplage larynx-conduit vocal (frequence laryngienne des productions vocaliques). - Revue d'Acoustique 6, 235-244.

[5]. MRAYATI M. (1976), Contribution aux études sur la production de la parole. - Thèse de Doctorat d'Etat, INP Grenoble.

[6]. FENG G. (1986), Modelisation acoustique et traitement du signal de parole, le cas des voyelles nasales. - Thèse de Doctorat, INP Grenoble.

[7]. HAWKINS S. & STEVENS K.N. (1985), Acoustics and perceptual correlates of the non-nasal - nasal distinction for vowels. - J. Acoust. Soc. Am. 77, 1560-1575.

[8]. FUJIMURA O. & LINDQVIST J. (1971), Sweep-tone measurements of vocal-tract characteristics. -J. Acoust. Soc. Am. 49, 541-558.

[9]. BEDDOR P.S. & STRANGE W. (1982), Cross-language study of perception of the oral-nasal distinction. - J. Acoust. Soc. Am. 71, 1551-1561.

[10].REENEN Van- P. (1982), Phonetic feature definitions. Their integration into phonology and their relation to speech. A case study of the feature nasal. - Dordrecht, Cinnaminson.

[11].MARTINET A. (1969), Le français sans fard. - PUF, Paris.

[12].LONCHAMP F. (1979), Analyse acoustique des voyelles nasales françaises. - Verbum II, 1, 9-54.

[13].BRICHLER-LABEYE C. (1970), Les voyelles françaises. - Klincksieck, Paris.

[14].ZERLING J.P. (1984), Phénomènes de nasalité et de nasalisation vocaliques : Etude cineradiographique pour deux locuteurs. - Travaux de l'Inst. de Pnonétique de Strasbourg 16, 241-266.

[15].SANCHEZ H. & BOE L.J. (1984), De la coupe sagittale à la fonction d'aire du conduit coval. - Bull. de l'Inst. de Phonétique de Grenoble 13, 1-24.

[16].FANT G. (1960), Acoustic theory of speech production. - Mouton, The Hague.

[17].MAEDA S. (1982), The role of the sinus cavities in the production of nasal vowels. - Proc. IEEE ICASSP, Paris, 911-914.

[18].FANT G. (1985), The vocal tract in your pocket calculator. -in Phonetic Linguistics, FROMKIN V.(ed.) New York.

[19].LINDQVIST-GAUFFIN J. & SUNDBERG (1976), Acoustic properties of the nasal tract. - Phonetica 33, 161-168.

[20].MAEDA S. (1984), Une paire de pics comme correlat acoustique de la nasalisation des voyelles. - 13e JEP du GALF, Bruxelles, 223-224.

ACKNOWLEDGMENTS

S. MAEDA, G. FANT & L.J. BOE for suggestions ; R. SOCK for translation.

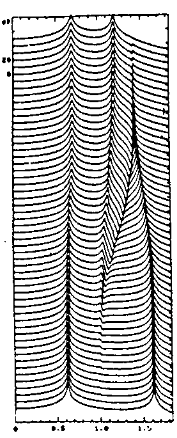


Fig 1

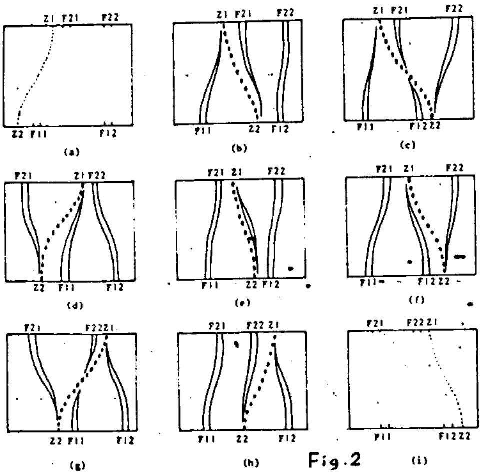


Fig.2

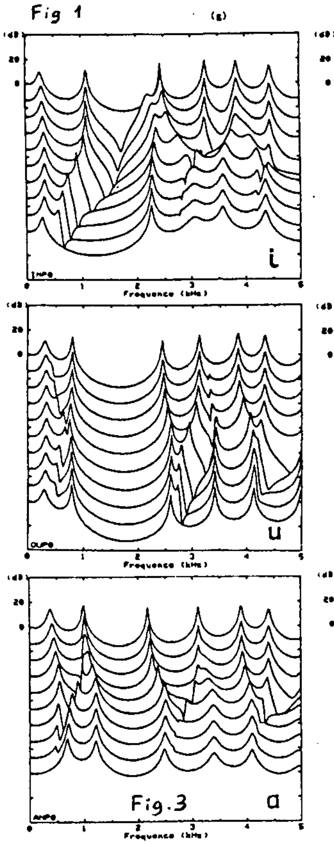


Fig.3

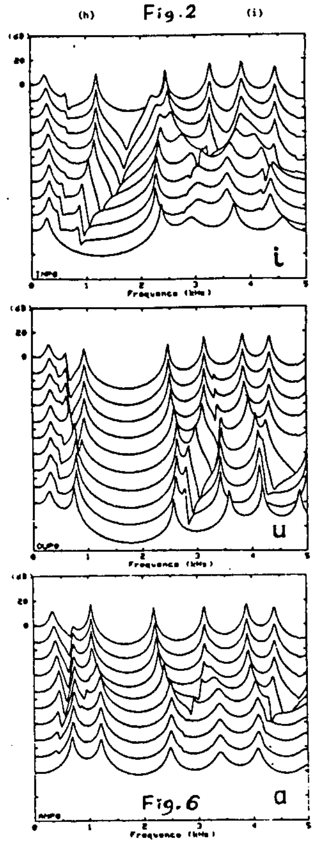
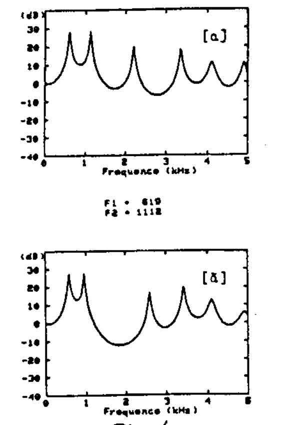


Fig.6



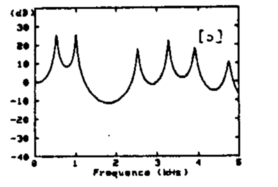


Fig.4

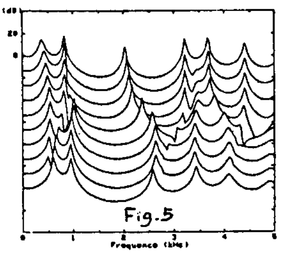


Fig.5

CONTRIBUTION A LA CLASSIFICATION ARTICULATORI-ACOUSTIQUE DES VOYELLES : "ETUDE DES "MACRO-SENSIBILITES A L'AIDE D'UN MODELE ARTICULATOIRE.

R. MAJID[1], C. ABRY[1], L.J. BOE[1] & P. PERRIER[2]

Institut de la Communication Parlée. UA CNRS 368, FRANCE.
[1] Institut de Phonétique, Univ. Langues & Lettres, BP 25X 38040 GRENOBLE CEDEX.
[2] Laboratoire de la Communication Parlée, ENSERG/INPG, 46 Av. Félix Viallet, 38031 GRENOBLE CEDEX.

RESUME

Le domaine de validité des fonctions de sensibilité de FANT & PAULI (1974) est trop limité pour qu'elles puissent être exploitées dans le cadre de la synthèse articulatoire. Avec le modèle articulatoire de MAEDA (1979), sont proposées les "macro-sensibilités" articulatori-acoustiques de treize voyelles (dont les onze du français) qui décrivent la totalité du triangle vocalique. Les résultats sont analysés à la lumière de la théorique quantique de STEVENS et nous proposons deux paramètres articulatoires qui permettent une orthogonalisation du triangle vocalique interprétable en termes articulatoires.

INTRODUCTION

L'étude des conséquences de petites perturbations de la fonction d'aire du conduit vocal sur le signal acoustique correspondant, a fait l'objet de nombreux travaux. Citons parmi les plus importants ceux de FANT & PAULI [1] qui, à la suite de SCHROEDER [2] qui avait évalué les énergies cinétique et potentielle mises en jeu, ont proposé les "fonctions de sensibilité" associant les perturbations de la fonction d'aire et les variations des fréquences formantiques correspondantes.

Pour intéressantes qu'elles soient, ces fonctions présentent, à notre avis, trois limitations importantes :

* la fonction d'aire n'est pas une donnée directement interprétable en termes de commandes articulatoires ;
* les perturbations introduites dans la fonction d'aire ne tiennent pas compte des contraintes intra- et inter-articulateurs ;
* le domaine de validité des perturbations (10% max.) est faible, en regard des variations de forme du conduit vocal pouvant intervenir lors de la production réelle de parole.

Pour cet ensemble de raisons, nous avons étudié les "macro-sensibilités" articulatori-acoustiques, c'est-à-dire les conséquences sur les fréquences formantiques de variations, éventuellement importantes, de paramètres articulatoires. Les résultats que nous avons ainsi obtenus, peuvent être interprétés en termes de stabilité articulatori-acoustique [3], et font apparaître deux paramètres permettant une "orthogonalisation" non arbitraire de l'espace vocalique F_1/F_2 [4].

LA METHODE UTILISEE

Pour ce travail, nous avons adopté une modélisation globale du processus de production vocalique, susceptible d'intégrer au mieux les connaissances physiologiques et acoustiques.

La clé de voûte de cette étude est constituée par le modèle articulatoire proposé par MAEDA [5]. Notre choix trouve sa justification dans la conception même du modèle : il résulte d'une analyse statistique effectuée, dans un premier temps, sur des radiographies pour la langue, et, dans un deuxième temps, sur des labiofilms pour les lèvres, et il intègre a priori des connaissances sur les articulateurs, ce qui autorise à la fois une bonne adéquation avec la réalité physiologique et une grande facilité d'interprétation en termes de phonétique articulatoire. Les premiers tests globaux [6] ont très largement confirmé la validité de ce modèle. Il est commandé par cinq paramètres : les "lèvres", la "mâchoire", le "corps", le "dos" et la "pointe" de la langue. Il fournit une coupe sagittale, à partir de laquelle est évaluée la fonction d'aire [7] qui n'est ici qu'une donnée intermédiaire.

La réponse harmonique est ensuite calculée à l'aide d'un analogue électrique du conduit vocal [8] implanté par CHARPENTIER (1982) [9], qui inclut les pertes par vibrations des parois, par viscosité, par chaleur et par rayonnement.

LA SELECTION DE STANDARD VOCALIQUES.

A partir d'un "dictionnaire articulatori-acoustique" contenant environ 200.000 formes vocaliques potentielles générées selon la procédure présentée ci-dessus, et décrivant la totalité du triangle acoustique F_1/F_2 [6], nous avons déterminé les configurations susceptibles de correspondre aux voyelles orales du français, décrites par des ellipses de dispersion formantiques [10]. De cet ensemble, par approximations successives, nous avons déduit des valeurs standard en nous astreignant à respecter une quadruple cohérence structurelle au niveau : des paramètres articulatoires, des coupes sagittales, des fonctions d'aire et des formants.

Pour compléter ce triangle F_1/F_2, nous avons générés deux voyelles supplémentaires qui remplissent l'espace laissé libre entre le [y] et le [u] :

* [ɯ], voyelle obtenue à partir du [u] par un avancement de la langue et une plus grande aperture aux lèvres ;
* [x], une voyelle "inconnue", située dans l'étroite zone comprise entre [ɯ] et [u] ; MAEDA [11], FENG et al. [12], ont avancé l'hypothèse que cette zone correspond à la cible des voyelles nasales ; [x] est obtenue à partir du [u] par un abaissement du dos de la langue et de la mâchoire.

LES MACRO-SENSIBILITES PARAMETRIQUES.

Dans le modèle de MAEDA, les paramètres de commande évoluent autour d'une valeur moyenne, dans les limites de - 3 à + 3 fois l'écart-type. Nous avons décidé de faire varier, un à un, chacun des paramètres de commande, avec une dynamique totale de 2 fois l'écart-type, répartie, quand c'était possible, symétriquement de part et d'autre de la valeur cible de chaque paramètre, et ceci pour chaque voyelle.

Examinons maintenant un à un l'effet de chacun des paramètres articulatoires.

Les voyelles françaises.

Le paramètre "lèvres". (Figure 1.a.)

On constate une tendance générale de toutes les voyelles à converger vers la zone du [u]. Les macro-sensibilités du [o], du [ɔ] et du [u] à ce paramètre sont particulièrement importantes. Pour la voyelle [u], en particulier, la rétraction des lèvres, liée à l'augmentation de l'aperture labiale, induit un important déplacement vers le centre du triangle acoustique, via le [o].

C'est en revanche pour [i] et pour [e], et pour les voyelles d'avant en général, que le paramètre "lèvres" a le moins d'influence. Notons cependant qu'il est loin d'être négligeable, puisque pour le [i] la fermeture-protrusion des lèvres permet d'atteindre la zone du [y]. Enfin, pour les voyelles [a] et [o] et [ɔ], on constate essentiellement une variation du premier formant F_1, dont on sait qu'il est sen sible à l'aperture labiale.

De manière générale, on obtient un résultat déjà souligné par LINDBLOM & SUNDBERG [13] et PERKELL & NELSON [14] : les fréquences des formants diminuent avec la fermeture des lèvres.

Le paramètre "mâchoire". (Figure 1.b.)

Dans tous les cas, l'abaissement de la mâchoire provoque une augmentation de F_1 et une diminution plus ou moins sensible de F_2 (sauf pour le [u]). Mais c'est avec [y], [i] et [œ] qu'apparaissent les variations maximales. Ces résultats confirment pleinement ceux de LINDBLOM & SUNDBERG [13] et les précisent. Remarquons que, dans le cas du [a], l'effet de l'abaissement de la mâchoire n'est pas une fonction monotone. On peut expliquer ce phénomène par le fait que la position du lieu de constriction, situé dans la région pharyngale, est affectée lors de l ouverture de la mâchoire : on observe ainsi à partir d'un certain point, les effets conjoints de l'ouverture de la mâchoire et du déplacement du lieu d'articulation.

Notons enfin la grande insensibilité du groupe [u]. [o], [ɔ] aux variations du paramètre "mâchoire".

Le paramètre "corps de la langue" (Figure 1.c)

Le déplacement de la langue vers l'avant provoque, de manière générale, une tendance à la convergence vers la zone du [i]. Ceci est très marqué pour les voyelles d'avant qui sont de loin les plus sensibles.

La précision nécessaire sur [i] est très importante puisque dès qu'on recule la langue, on passe à un [e], puis à un [œ], en se dirigeant rapidement vers le centre du triangle vocalique.

Le paramètre "dos de la langue". (Figure 1.d)

Ce sont [i], [y] et [e] qui sont les voyelles les plus sensibles aux variations de ce paramètre. En revanche, les voyelles d'arrière sont relativement peu affectées. On observe que l'abaissement du dos de la langue fait tendre le [i] vers le [e], alors que [u] évolue légèrement dans la direction du [y]. Les déplacements dans le plan F_1/F_2 se font en gros selon deux directions quasi orthogonales, de façon beaucoup plus nette que les simples tendances mises en évidence par GOLDSTEIN [15]. L'abaissement du dos de la langue peut avoir quatre conséquences différentes :

* pour [i,e,ɛ] F_1 augmente et F_2 diminue ;
* pour [y,ø,œ] F_1 augmente et F_2 augmente ;
* pour [u,o,ɔ] F_1 est stable et F_2 augmente ;
* enfin, pour [a], F_1 reste stable et F_2 diminue.

Le [a] se différencie des autres voyelles d'arrière par le sens de la variation de F_2. Là encore, l'explication réside dans le recul du lieu de constriction associé à l'aplatissement de la langue.

Le paramètre "pointe de la langue". (Figure 1.e)

L'abaissement de la pointe de la langue affecte, comme on peut facilement le prévoir, essentiellement les voyelles d'avant fermées [i,e,y]. Il s'accompagne généralement d'une diminution de F_2 et d'une augmentation de F_1.

Conclusions.

De manière générale, il ressort de l'observation des figures 1.a-1.e que [i,e,y] sont les voyelles les plus sensibles , avec en particulier de grosses variations de F_1, et qu'inversement, les voyelles d'arrière présentent une assez bonne résistance à la variation des paramètres.

Les voyelles [ɯ] et [x].

L'observation de la figure 1 montre que le comportement de la voyelle [ɯ] est en grande partie comparable à celui des voyelles d'avant, sauf pour le "dos de la langue" analogue à celui du [y]. Cette voyelle se situe donc un peu à mi-chemin entre les voyelles d'avant et les voyelles d'arrière du français.

Mis à part pour les "lèvres", la voyelle [x], ne présente pas de grandes sensibilités aux paramètres articulatoires : elle se rapproche en cela du [u]. Le fait que, pour l'ensemble des langues, peu de voyelles fréquentent cette zone du plan F_1/F_2 [11], ne semble donc pas pouvoir être expliqué par une grande instabilité articulatori-acoustique. S'agit-il d'instabilité articulatori-perceptive, de "mauvaise forme perceptive" [16], ou faut-il introduire des contraintes articulatoires encore plus sévères sur le modèle de MAEDA ? Le débat reste ouvert...

ORTHOGONALISATION DU TRIANGLE VOCALIQUE.

Deux paramètres articulatori-acoustiques non ambigus : lèvres et corps de la langue.

En 1983, FANT [4] a proposé deux paramètres spectraux, "spread" et "flatness", correspondant dans l'espace acoustique à F'_2-F_1 et F'_2+F_1, permettant une optimisation de la représentation du système vocalique du suédois. Ainsi la distinction entre les voyelles d'arrière et les autres, d'une part, et les voyelles arrondies et les autres d'autre part, est bien mise en évidence. Cependant la rotation ainsi effectuée dans le plan F_1/F'_2 présente ùn aspect arbitraire. Nous pensons que l'utilisation des resultats de l'etude des macrosensibilités peut être à cet égard très intéressante.

Observons, en effet, les figures 1.a et 1.c : la fermeture-protrusion des lèvres tend à déplacer toutes les voyelles dans la direction du [u]; tandis que toutes ont tendance à converger vers le [i] lors d'un mouvement du corps de la langue vers l'avant. Ce phénomène n'apparait que dans ces deux cas : les paramètres "lèvres" et "corps de la langue" sont donc les seuls paramètres non-ambigus dans la relation articulatori-acoustique. Utilisons cette propriété en traçant les droites de régression, dans le plan F_1/F_2, de l'ensemble des points des trajectoires obtenues pour les "lèvres" d'une part (droite D_1), et pour le "corps de la langue" d'autre part (droite D_2). Considérons alors le plan orthogonal défini par les droites D_1 et D_2. Les positions cibles se placent alors sur un triangle orthogonal défini par les droites [i-a] et [a-u] (Figure 2). Cette première représentation souligne l'importance du "corps de la langue" et des "lèvres" pour le classement articulatoire des voyelles. Voyons maintenant quelle peut être l'interprétation géométrique de ces deux axes.

Interprétation géométrique.

Nous proposons de relier : l'axe [i-a] à la notion de postériorité (Backness [17]), que nous définissons comme la distance entre l'incisive inférieure et le centre d'un cercle ajusté à la plus grande courbure de la langue ; l'axe [u-a] à la notion de profondeur (depth [18]), correspondant à la distance entre l'extrémité des lèvres et le point d'élévation maximal de la langue. La figure 3 montre bien la possibilité d'une adéquation satisfaisante entre la réprésentation des positions cibles des voyelles dans cet espace et leurs positions dans l'espace décrit plus haut (D_1/D_2).

Relation avec l'acoustique.

Dans le plan acoustique, on peut faire la correspondance entre ces axes et les axes F_1 et F''_2, où $F''_2=0.6*F_2$, ainsi que le confirme la figure 4. Notons qu'articulatoirement, la postériorité est une grandeur absolue (référence fixe) et la profondeur ~~une~~ grandeur relative (référence mobile), alors que, curieusement, c'est l'inverse dans le plan acoustique.

Nous avons ensuite examiné les relations entre les paramètres articulatoires, les paramètres géométriques et les paramètres acoustiques ainsi proposés. Les résultats sont donnés figure 5. On remarque la bonne corrélation entre les données géométriques et acoustiques, et les paramètres articulatoires et acoustiques. La corrélation entre "lèvres" et "profondeur" est moins bonne, mais cela est sans doute dû aux caractéristiques du modèle articulatoire, dans lequel, rappelons le, les "lèvres" ne sont pas liées statistiquement à la mâchoire et à la langue.

Nous pouvons donc proposer, dans le plan F''_2/F_1, une rotation optimisant la représentation du sytème vocalique parfaitement interprétable en termes articulatoires et reliée au rôle joué par les lèvres et surtout par le corps de la langue.

CONCLUSIONS.

Nous avons défini pour chaque voyelle des positions cibles et pour chacune d'elles nous avons présenté leurs macro-sensibilités aux variations des paramètres articulatoires. Les résultats que nous avons obtenus confirment les grandes tendances déjà mises en évidence par LINDBLOM & SUNDBERG [13] et GOLDSTEIN [15], et les affinent. Ils vont dans le sens général de la théorie quantique de STEVENS [3] : il existe, en effet, très clairement des régions acoustiques qui sont peu sensibles aux variations de certains paramètres articulatoires. Enfin l'exploitation de ces résultats nous a permis d'orthogonaliser l'espace vocalique, selon une rotation dans l'espace acoustique, parfaitement interprétable en termes articulatoires. Il nous reste cependant à affiner la définition de la notion de profondeur.

REFERENCES.

[1] G. FANT & PAULI S., "Spatial Characteristics of Vocal Tract Resonance Modes.", Speech. Comm. Seminar 2, 121-132, 1974.

[2] SCHROEDER M.R., " Determination of the Geometry of the Human Vocal Tract by Acoustic Measurement.", J. Acoust. Soc. Am. 41, 1002-1010, 1967.

[3] K.N. STEVENS, " The Quantal Nature of the Speech : Evidence from Articulatory-Acoustic Data.", in Human Communication, Mc Graw Hill, New York, 1972.

[4] G. FANT, "Speech Production : Feature Analysis of Swedish Vowel. A revisit.", STL QPSR 2-3, 1-15, 1983.

[5] S. MAEDA, " Un modèle articulatoire de la langue avec des composantes linéaires.", 10èmes JEP, GALF-GCP, 152-162, 1979.

[6] P. PERRIER, L.J. BOE, R. MAJID & B. GUERIN, "Modélisation articulatoire du conduit vocal : exploration et exploitation.", 14èmes JEP, GALF-GCP, 55-58, 1985.

[7] H. SANCHEZ & L.J. BOE, " De la coupe sagittale à la fonction d'aire du conduit vocal.", 13èmes JEP, GALF-GCP, 23-25, 1984.

[8] J.L. FLANAGAN, K. ISHIZAKA & K.L. RIPLEY, "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract.", B.S.T.J. 54, 485-506, 1975.

[9] F. CHARPENTIER, "Un logiciel de simulation électrique du conduit vocal.", C.N.E.T. Lannion, Comm. Perso., 1982.

[10] C. ABRY, BOE L.J. & R. DESCOUT, "[i,a,u] ? Pas si fou ? Ou les lèvres des consonnes maximisent elles l'espace des voyelles ?", 13èmes JEP, GALF-GCP, 205-207, 1984.

[11] S. MAEDA, " Une paire de pics spectraux comme corrélat acoustique de la nasalisation des voyelles.", 13èmes JEP, GALF-GCP, 223-224, 1984.

[12] G. FENG, C. ABRY & GUERIN B., "The Nasopharyngeal Tract : A Target for Nasality. Acoustic Simulation vs. Sweep Tone Measurements.", Proc. 12th. I.C.A., A 3.3, 1986.

[13] B.E.F. LINDBLOM & J.E.F. SUNDBERG, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement.", J. Acoust. Soc. Am. 50, 1166-1179, 1971.

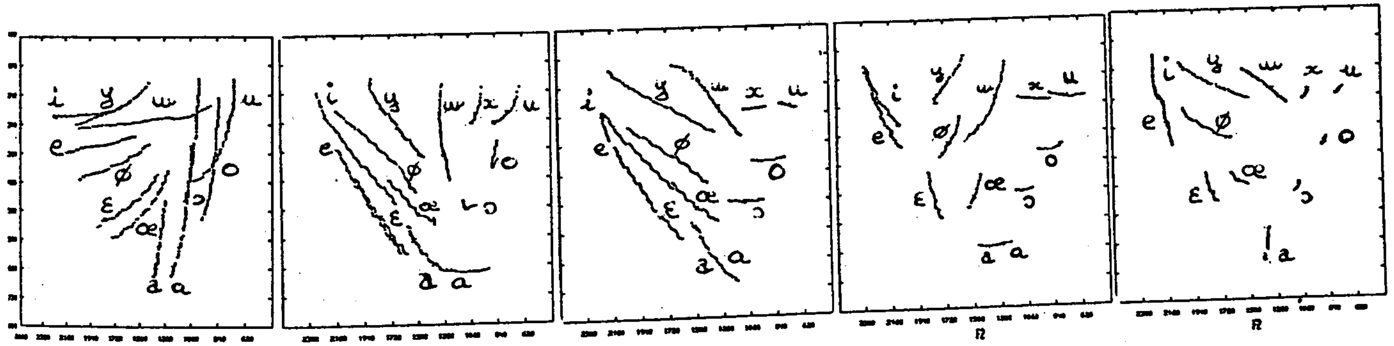
[14] J.S. PERKELL & W.L. NELSON, "Variability in Production of the Vowels /i/ and /a/.", J. Acoust. Soc. Am. 77, 1889-1895, 1985.

[15] L. GOLDSTEIN, "Vowel Shifts and Articulatory-Acoustic Relations.", 10th Int. Congr. Phonetic Sci. IIA, 267-273, 1983.

[16] C. ABRY & J.L. SCHWARTZ, "Quelques éléments pour une théorie des objets phonétiques du langage.... Autour d'une voyelle inconnue.", Revue de l'I.C.P., Bulletin du L.C.P., 1 ,1987.

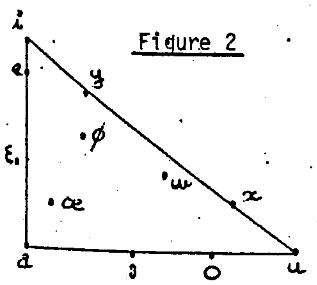
[17] N. CHOMSKY & M. HALLE, "The Sound Pattern of English.", Harper & Row, New York, 1968.

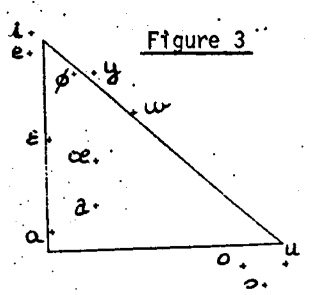
[18] H. TRAUNMULLER, " Some Aspects of the Sounds of Speech Sounds.", Workshop of the Psychophysics of Speech Perception, Utrecht, 1986.

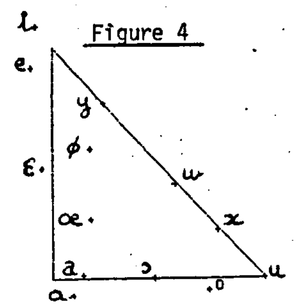


1.a : Les lèvres 1.b : La mâchoire 1.c : Le corps 1. d : Le dos 1.e : La pointe

Figure 1 : LES MACRO-SENSIBILITES







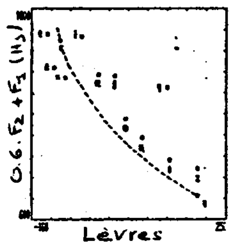
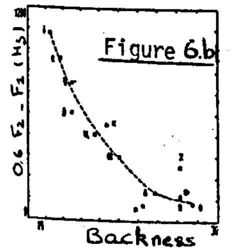
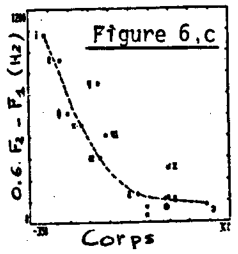
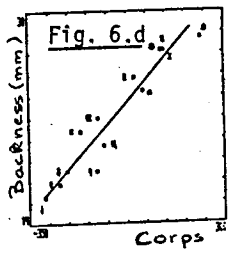
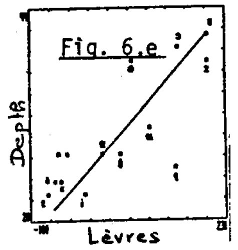


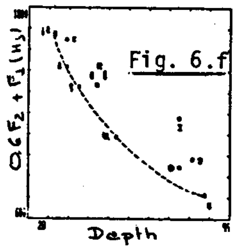
Figure 6.a











VOCAL TRACT VOCALIC NOMOGRAMS : ACOUSTIC CONSIDERATIONS.

A Crucial Problem : Formant Convergence.

Pierre BADIN

Louis-Jean BOE

Institut de la Communication Parlée (U.A. CNRS 368)

Laboratoire de la Communication Parlée
E.N.S.E.R.G. - I.N.P.G.
46, av. Félix Viallet - 38031 GRENOBLE Cédex

Institut de Phonétique de Grenoble
Université des Langues et des Lettres.
B.P. 25 - 38040 GRENOBLE Cédex, FRANCE

ABSTRACT

Presented by FANT in 1960, nomograms have not been thoroughly exploited for studying vocalic productions. It took a long time before we noticed the utilisation of this kind of tool with STEVENS' quantal theory (/7/) and, more recently, with the attempts by LADEFOGED and BLADON to reproduce FANT's nomograms (/5/). In spite of its simplicity, the four-tube model can be used to interpret the articulatori-acoustic relations (formants-cavities affiliations) and to describe the main vocalic types.

We first make a revisit of FANT's explanations for the "affiliations" between formants and cavities, and we try to refine his findings, in relation with losses in the vocal tract (especially at the glottis). We study the phenomenon of formant convergence ("focalization"), and more precisely in the case of |i|. Finally, we compare the results of our simulations with natural speech.

INTRODUCTION

Presented by FANT in 1960, nomograms have not been thoroughly exploited for studying vocalic productions. It took a long time before we noticed the utilisation of this kind of tool with STEVENS' quantal theory (/7/) and, more recently, with the attempts by LADEFOGED and BLADON to reproduce FANT's nomograms (/5/). We think that nomograms are still very powerful tools in the field of articulatori-acoustic relations for vowel production, for vocalic system interpretation and prediction, as well as for formant measurement up to F5 (e.g. F'2 estimation).

In the first section, we give a brief description of FANT's nomograms, and we recall FANT's explanations about the "affiliation" phenomenon. Then, we study the effect of losses, especially at the convergence point between F2 and F3, in the case of an intermediate lip opening.

1. FANT'S VOCALIC NOMOGRAMS : A REVISIT

1.1 The Four Tubes Model : Basic Resonances

In order to mimic in a simple way the acoustical behavior of the vocal tract when the constriction is moving along the midline from glottis to lips, FANT defined a four-tube model (/4/, p.71-79). Here, we retain the following configurations for the 4 sections :

Pharynx cavity : $A_4 = 8.cm^2$, L_4 varying ;
Tongue constriction : $A_3 = 0.65cm^2$, $L_3 = 5.cm$;
Mouth cavity : $A_2 = 8.cm^2$, L_2 varying ;
Lips : A_1=0.16, 4.cm^2 or no lips, L_1=1.cm.

The constriction center coordinate X_c (measured from the glottis position) can vary from -2.5cm to 17.5cm, keeping : $L_2 + L_3 + L_4 = 15cm$.

In order to understand the affiliation phenomenon, we have made nomograms for the different cavities alone. We have supposed very low heat and viscosity losses, no wall vibration, no lip radiation load, and we have used the acoustic model developped by BADIN & FANT (1984).

Fig.1a shows the resonances of the back cavity alone : integer multiples of the half wavelength resonance $c/2L_4$ (solid lines). The dashed line at the bottom corresponds to a "HELMHOLTZ resonance" between the back cavity and the constriction neck.

Fig.1b shows the resonances of the "mouth + lips" front cavity alone. In the case of no lip section, the front cavity produces resonances at odd integer multiples of the quarter wavelength resonance $c/4L_2$ (dotted lines). In the case of a very small lip opening ($A_1 = 0.16cm^2$), we obtain the same behavior as for the back cavity : resonances at integer multiples of the half wavelength resonance $c/2L_2$ (upper dashed lines), plus a "HELMHOLTZ resonance" between the mouth cavity and the lip constriction (bottom dashed line). For the intermediate case ($A_1 = 4.cm^2$), the resonances have an intermediate behavior (solid lines).

Fig.1c shows the behavior of the constriction resonances : when the constriction tube is open at both ends, it produces the half wavelength resonance $c/2L_3$ (middle and right part of the figure), whereas when it is closed at the glottis, it produces resonances at odd integer multiples of the quarter wavelength resonance $c/4L_3$ (left part).

1.2 Affiliation and Coupling : focal points

Fig.2 (a, b, c) represent, for the same simplified boundary conditions as in 1.1, the resonances of the complete four-tube system (solid lines) and the contributions from the different cavities (dashed lines) deduced from Fig.1. This nomogram demonstrates clearly the affiliation phenomenon : whenever a solid line is very close to a dashed line the resonance of the whole system

depends mainly on the cavity corresponding to the dashed line ; on the opposite, when the solid line departs from the dashed line, there is a little affiliation and coupling phenomena occur.

When two dashed lines cross each other, i.e. a resonance associated with the front cavity and a resonançe associated with the back cavity have close frequencies for a given position of the tongue constriction, we observe a focalization of the formants : we call **"focal point"** (/2/) this convergence region where the affiliation of the lower and upper resonances switches from one cavity to the other when the tongue constriction is shifted.

Coupling between two resonant systems is a classical phenomenon : it always spreads apart the natural frequencies of the two systems. More precisely, the greater the coupling, the larger the frequency spreading, and conversely, the smaller the coupling (i.e. the constriction cross area), the more prominent the formant convergence. The case of a small lip opening (|u|, Fig.2c) shows a good example of large coupling : the HELMHOLTZ resonances associated with the "back cavity + tongue constriction" resonator and with the "mouth cavity + lip constriction" are spread into F1 and F2 when coupled through the constriction tube. The case of a middle lip opening (|i|, Fig.2b) shows a good example of F2/F3 convergence : on the glottis side of the convergence point, F2 is clearly associated with the front cavity, and F3 with the back cavity, whereas it is the opposite on the lip side. We call an |i| configuration close to the convergence point "focal" |i|, and configurations on the glottis side and on the lip side respectively "prefocal" and "postfocal" |i|'s.

By generalization of the above examples, it is possible to define other formant convergences and other vocal types. The vowel |a| with open lips shows a F1/F2 convergence between the HELMHOLTZ back cavity resonance and the front cavity first resonance (Fig.2b). The vowel |y| has a F2/F3 convergence corresponding to the HELMHOLTZ "mouth cavity + lips" resonance and the back cavity half wavelength resonance in the case of constricted lips (Fig.2c).

The constriction presents an interesting behavior in the central region of the nomogram corresponding to |u| (Fig.2c). By chance (because L2 is exactly one third of the total length of the configuration without lips), there is a point where three dashed lines cross each other : the half wavelength resonances of the front and back cavities, and of the constriction itself. The constriction resonance is not modified, and the resonances of the front and back cavities are spread apart. This focal point is one of the two convergence points for |u|, but we should mention that, owing to the lip constriction, the amplitude for F3, F4 and F5 is rather low. For more open lips, the situation is different : this triple point does not exist, and the resonance frequency of the constriction is much modified by the coupling. We can induce that the formants F3 and F4 must be rather sensitive to the location and to the size of the constriction.

1.3 Losses effects

To have a better insight into the coupling phenomena we have neglected in section 1.2 the losses and the boundary effects such as wall vibration or lip radiation. We now include these effects in the simulation. The contribution of different types of losses to formants and bandwidths has already been discussed somewhere else (/4/, /1/). We just recall that most of the losses are due to the radiation at the lips and to the glottis resistance. The effect of the losses due to the lip radiation or to the glottis is to decrease the amplitudes and to broaden the bandwidths of the associated resonances. The losses at the lips increase with lip opening, and the losses at the glottis increase with glottis opening.

Because of affiliation, the lip and glottis opening conditions have a selective influence upon the resonances of the associated cavities. It is interesting to analyze the nomogram in a region of formant convergence. The selective effects of the glottis losses are depicted on Fig.3 : a 3-D representation of the nomogram, for a middle lip opening (A_1=4.cm^2), is given for a "small" glottis opening (a) (glottis area A_g=10.mm^2, resistance R_g=43Ω) and for an "moderate" glottis opening (b) (A_g=4.6mm^2, R_g=100Ω) . We clearly see that when the glottis opens, the amplitude of the formant related to the back cavity decreases and the bandwidth increases, due to tne increase of glottis losses. The same phenomenon happens for a variation of the lip opening (which would be associated with a displacement of the convergence point, because of the correlated variation of the length end correction at the mouth, /4/, p.36).

A more detailed analysis of Fig.3 leads to a notion of **"bandwidth inversion"**, especially clear for the small glottis opening : on the glottis side of the focal point, the bandwidth of F2 (which is associated with the front cavity) is greater than that of F3, whereas on the lip side the bandwidth of F3 (which is then associated with the front cavity) is greater : this inversion is the consequence of affiliation. When the glottis opening increases this effect decreases, but for a moderate glottis opening, the inversion effect occurs again, the role of the two cavities being then exchanged (wider bandwidths for the back cavity resonance, Fig.3b).

2. EXPERIMENTAL ILLUSTRATION

In this section, we compare some results of the above simulation study with equivalent situations for real speech.

2.1 F2/F3 convergence for |i|

A recent and interesting attempt to reproduce FANT's nomograms has been the one by LADEFOGED & BLADON (/5/). They present a series of sonagrams corresponding to sounds for which the tongue constriction is progressively shifted from the pharynx to the teeth, every other parameter being supposed constant : they noticed regular shifts of the formant frequencies corresponding to what was expected from FANT's nomograms, but they mentioned that, for a vowel in one of their series, "F3 would appear to have suddenly assumed a value comparable

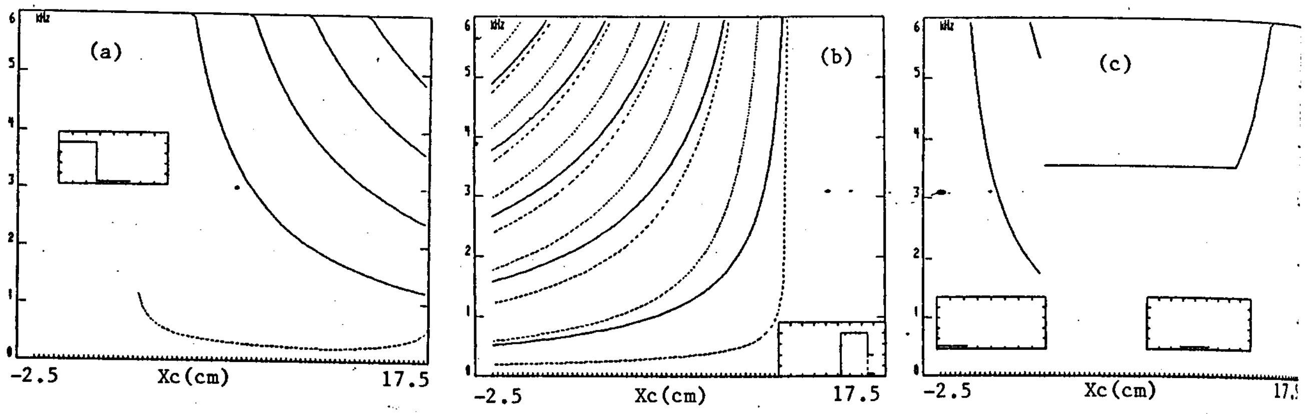


Fig.1 : Resonances of the different cavities vs. tongue constriction location.
(a) back cavity : pharynx + tongue constriction (b) front cavity : mouth + lip constr. (c) tongue constr.

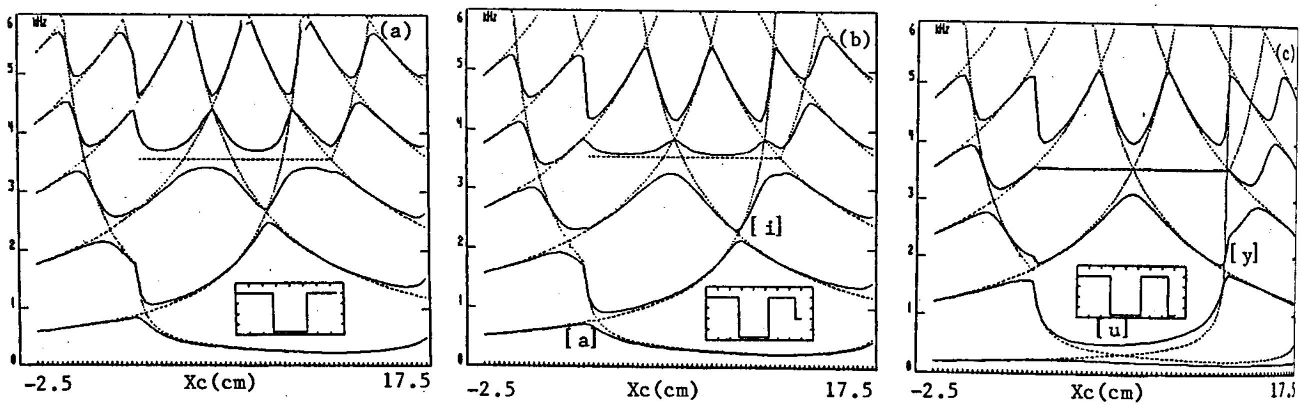


Fig.2 : Resonances of the whole four tubes system vs. constriction location.
(a) maximum lip opening (L_1=0.) ; (b) middle lip opening (A_1=4.cm^2) ; (c) small lip opening (A_1=0.16cm^2).

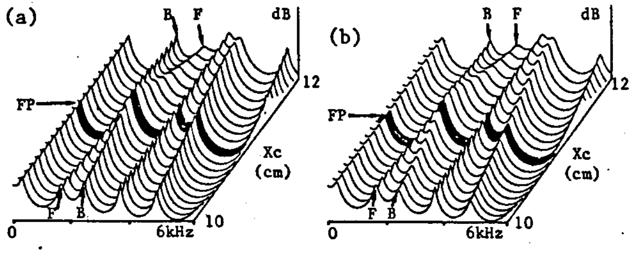


Fig.3 : 3-D Nomogram (transfer functions) around the F2/F3 convergence for |i| (A_1=4.cm^2).
(a) small glottis opening (b) moderate glottis opening.
F : front cavity B : back cavity FP : focal point.

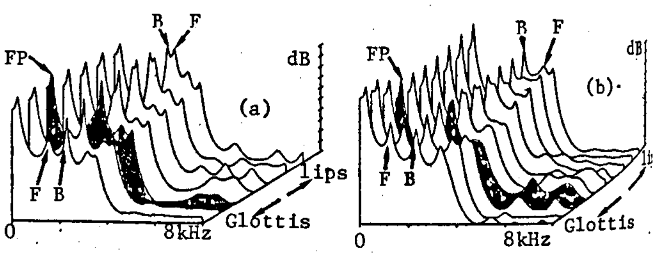


Fig.5 : 3-D spectral representation of focal |i|s
(a) in |tiwit| (b) in |ziwiz|.

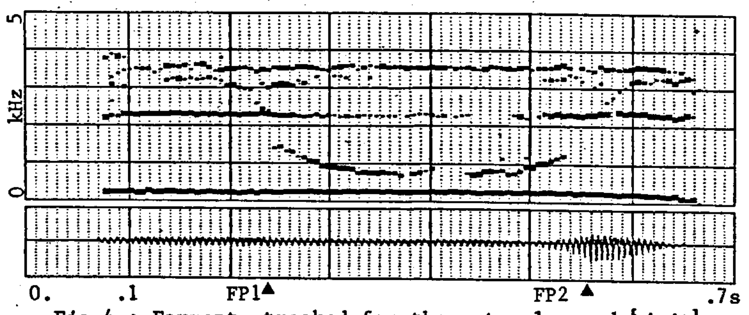


Fig.4 : Formants tracked for the natural sound [tiwit].

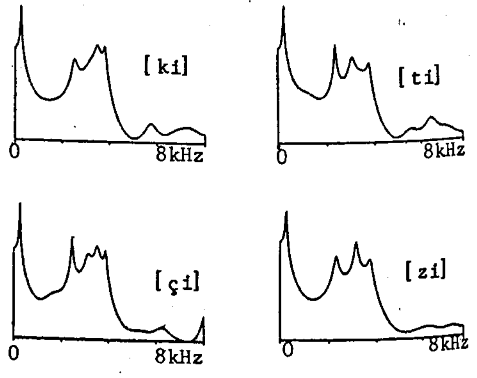


Fig.6 : spectra of |i| in different contexts.

to that of F4 in the previous vowel". By reference to the above simulations, we know that F2 and F3 can merge into a focal point for a given position of the tongue constriction : we believe that an answer to LADEFOGED & BLADON's difficulties is that F2 and F3 are actually merged into a single formant.

To check this hypothesis, we have recorded a series of |CiwiC| sounds, where the transitions |iw| and |wi| correspond roughly to a shift of the tongue constriction, every other parameter being approximately constant (except for lip opening). If C is a dental or post-alveolar consonant, we insure that the tongue constriction shifts from a postfocal |i| (coarticulated with C) to a prefocal |i| (coarticulated with |w|), and thus that the focal point will be gone through. Fig.4 shows the evolution of the formants (tracked from cepstrum) for the sequence |tiwit| : we can easily follow the front cavity resonance going from F3 to F2 and back to F3 through the focal points FP1 and FP2 ; it might be possible to track this resonance even until F4. The spectral representation (obtained by LPC analysis), Fig.5 (a, b), shows a rather striking analogy with the transfer functions from Fig.3 for the behavior of F2 and F3 around the convergence point. This shows that the focal points predicted by our simulations are observable in real speech. It also reconfirms our view on LADEFOGED & BLADON's problem, and provides an explanation for the CHAFCOULOFF & al. (/3/) observations.

This convergence phenomenon may explain a part of the difficulties encountered by phoneticians in measuring F2 and F3 for |i| vowels, and the large dispersion of their data.

2.2 Bandwidth inversion around the F2/F3 focal point

The purpose of this section is to verify on real speech the effect of "bandwidth inversion". Thus we have recorded four sounds with |i| in 4 consonantal contexts, |k|, |ç|, |t| and |z|, corresponding to two articulatory locations (palatal vs. alveolar) and two glottis openings (small vs. large).

According to the transfer functions shown in Fig.3, we could expect the following relations for the bandwidths B2 and B3, and for the amplitudes A2 and A3 of F2 and F3 :

C	glottis opening	articul. location	Bandwidths relation	Amplitudes relation
|k|	small	prefocal	B2 > B3	A2 < A3,
|ç|	large	prefocal	B2 < B3	A2 > A3,
|t|	small	postfocal	B2 > B3	A2 < A3,
|z|	large	postfocal	B2 < B3	A2 > A3.

Fig.6 shows that the spectra of the |i| sounds in the four different coarticulation contexts checks with our expectations. This reconfirms the bandwidth inversion phenomenon around the focal point, and the influence of the glottis losses upon the relative bandwidth values of the formant associated with the back cavity compared to the the one associated with the front cavity.

CONCLUSION

The nomograms have allowed us to interpret the relations between formants and cavities, and to study the influence of the losses. We have illustrated these results with natural speech in a qualitative way, for a focal |i|. In order to obtain quantitative predictions closer to reality, we need to use a more realistic articulatory model : a study is in progress with MAEDA's articulatory model (/6/).

ACKNOWLEGEMENTS

We are very thankful to C. ABRY for his suggestions at the phonetic level, and to M. CONTINI and C. SCHIRRU as "good" speakers.

REFERENCES

/1/ **BADIN P. & FANT G. (1984)**, "Notes on Vocal Tract Computation", STL-QPSR 2-3/1984, 53-108.
/2/ **BOE L.J. & ABRY C. (1986)**, "Nomogrammes et Systèmes Vocaliques", 15èmes JEP GALF, 303-306.
/3/ **CHAFCOULOFF M. CHOLLET G. DURAND P. GUIZOL J. & RODET X. (1980)**, "Observation and Modelling of 'Formant' Transitions using ISAAS", IEEE Int. Conf. ASSP, 146-149.
/4/ **FANT G. (1960)**, "Acoustic Theory of Speech Production", Mouton, ('S-Graven Hague).
/5/ **LADEFOGED P. & BLADON A.(1982)**, "Attempts by Human Speakers to Reproduce FANT's Nomograms", Speech Comm. 1, 185-198.
/6/ **MAEDA S. (1979)**, "An Articulatory Model of the Tongue Based on a Statistical Analysis", J. Acoust. Soc. Am. 65, S1, S22 (A).
/7/ **STEVENS K.N. (1972)**, "The Quantal Nature of Speech : Evidence form·Articulatory-Acoustic Data", in "Human Communication : a Unified View", 51-66, Ed. By E. DAVID & P. DENES, Mac Graw Hill, New York.

DYNAMIC DETERMINATION OF ACOUSTIC VOWEL CONTRAST

FLORIEN J. KOOPMANS - VAN BEINUM   ROB P. DE SAINT AULAIRE

Institute of Phonetic Sciences, University of Amsterdam.

ABSTRACT

A problem in automatic speech recognition as well as in speech synthesis-by-rule is how to cope with the phenomenon of vowel reduction: vowels in connected speech rarely reach their target position (the intended vowel) as defined in isolated-word and isolated-vowel production. This paper describes a semi-automatic dynamic procedure for ongoing vowel analysis, thus providing a dynamically adjustable global measure for acoustic system contrast (ASC). This global ASC-measure, combined with local parameter values in the dynamic vowel analysis, may provide in due time various applications with respect to the description and use of vowel reduction aspects. On the basis of connected speech material (read texts and free conversation) of one Dutch and one Japanese speaker, the present results are compared with similar data earlier derived by hand segmentation and average vowel formant data per vowel segment.

INTRODUCTION

The great variability in the realization of vowel sounds, when produced by the same speaker but in different speech situations, plays an embarrassing role in speech technology. Vowels in connected speech rarely reach their target position (the intended phoneme) as defined in isolated-word and isolated-vowel production. In speech synthesis we badly need a model to describe this variability in order to increase intelligibility as well as naturalness, whereas in automatic speech recognition vowel reduction is an annoying phenomenon that we do not know how to cope with.

It is known that the degree of acoustic contrast between the vowels in a speaker's vowel system is dependent on various factors, partly global and partly local, but it is not clear as to how far all these factors are mutually dependent or independent.

In literature we can find, apart from socio-phonetic and linguistic factors, a large number of acoustic-phonetic factors that are believed to be responsible for the variability and the reduction of acoustic vowel contrast (for a detailed overview see Koopmans-van Beinum, 1980).

As far as the acoustic-phonetic factors are concerned (like speech rate, stress, intonation, and local context), quite a lot of research has been done with respect to the description of vowel contrast in various speech situations.

However, the relations between these factors and more specifically their hierarchical structure have been studied only fragmentarily yet. Lindblom (1963) for instance postulates that duration is the main determinant of vowel reduction, whereas Delattre (1969) claims stress and speech rate to be primary determinants with duration as a product of stress and tempo and therefore a secondary determinant. Gay (1977) and Den Os (1985) both show that an increase of speech rate not necessarily affects the formant frequencies of the vowels. Furthermore Koopmans-van Beinum (1980) indicates a different relation between stress and vowel duration for read texts as compared to texts with a free choice of words (retold story or free conversation). Also from perceptual studies on stress (e.g. Van Katwijk, 1974; Rietveld, 1983; Rietveld and Koopmans-van Beinum, to appear) the relation between loudness, intonation, speech rate, and vowel contrast reduction turns out to be a very complicated one.

In order to reach a better understanding of the relations and the hierarchical structure of the great variability of vowels, it is deemed necessary in our approach of the speech signal to make a distinction between 'global' factors (socio-phonetic aspects such as speaker, speech situation, and sex) influencing this variability, and 'local' factors (acoustic-phonetic and linguistic aspects within the neighbouring context).

We therefore started a project in order to develop and apply strategies to make optimal use of acoustic, socio-phonetic, and if possible also of linguistic information with respect to the variability in the realization of vowel phonemes. This will be done by means of a semi-automatic method for dynamic vowel analysis and cumulative data processing in three phases:

a) Any speech fragment of any speaker may be subjected to a dynamic acoustic-phonetic analysis to provide information on global aspects as mentioned above about the present vowel system (sex of the speaker, overall speech rate, degree of vowel contrast, etc.). Moreover the acoustic parameter values in the dynamic vowel analysis provide the possibility to define the moment when the global measure for acoustic system contrast (ASC) stabilizes. This indicates the duration of the speech sample needed for defining this value (and other global measures), and for dynamically adjusting it, if use is made of a moving window.

b) Subsequently local measures of acoustic vowel contrast or degree of reduction and variability will be developed based on acoustic-phonetic parameters as fundamental frequency, formant frequencies, bandfilter values, vowel duration, amplitude.

c) Finally the results of a) and b) will be used in various applications, as for instance labelling of segments as specific vowel phonemes, merely by using the local acoustic parameter values combined with global contrast measures and general information on the present vowel system, and defining the hierarchical structure of factors influencing the variability in vowel phonemes.

This paper reports on our first steps within this project, viz. the development of a method for the dynamic determination of the global measure for acoustic system contrast and its application to two quite distinct languages, Dutch and Japanese.

So three main questions have to be answered: 1) what differences yields the dynamic cumulative analysis method compared to the traditional static one; 2) what differences yields the (semi-)automatic procedure compared to the traditional manual one; 3) is the dynamic (semi-)automatic procedure applicable to two phonetically quite distinct languages.

DESIGN OF A DYNAMIC ANALYSIS AND DATA PROCESSING METHOD

As the aim of the present subproject was to develop a (semi-)automatic dynamic procedure of data processing, two parallel methods had to be compared: a) the traditional method making use of manual segmentation of vowels in the digitized speech fragment by means of a speech editor, followed by a dynamic acoustic-phonetic vowel analysis, and b) a (semi-)automatic method by carrying out a dynamic acoustic-phonetic analysis firstly on all speech frames, followed by an automatic vowel segmentation. Subsequently both methods are followed by a data processing program (based on formant frequencies or based on bandfilter values) which calculates in a cumulative way the acoustic system contrast measure ASC (Koopmans-van Beinum, 1980; De Saint Aulaire, 1986). This ASC measure is defined by the total variance of all vowels in the present vowel system, based on frequencies of the first (F1) and second formant (F2), (transformed in 100 * 10log Hz), using the formula:

$$ASC = \frac{1}{N} \sum_{j=1}^{N} (\vec{V}_j - \vec{C})^2 \qquad \text{in which}$$

$\vec{V}_j$ = the 2-dimensional vector of vowel j in the F1/F2-plane,
$\vec{C}$ = the 2-dimensional vector of the centroid C,
N = the number of vowels in the vowel system.

Apart from our formant-based ASC-measure we also developed a similar ASC-measure based on bandfilter variance, which turned out to be a good alternative for the formant-based one, but in this paper we left it out of consideration (Koopmans-van Beinum and De Saint Aulaire, 1986).

SPEECH MATERIAL

It is claimed that in so-called syllable-timed languages, like e.g. Spanish, Italian, and Japanese, the degree of spectral reduction is much less, if present at all, than in so-called stress-timed languages like e.g. English, Russian, and Dutch. In previous work, however, we met for Dutch and for Japanese a similar degree of vowel reduction expressed in comparable values of acoustic system contrast (De Graaf & Koopmans-van Beinum, 1982/83). Therefore we decided to test our dynamic analysis and data processing method on Dutch as well as on Japanese speech material.

The Dutch vowel system consists of twelve more or less monophthongal vowels and three diphthongs. All vowels and diphthongs may occur in stressed as well as in unstressed position. Furthermore about 30% of all vowel phonemes in Dutch consists of schwa sounds apart from reduced vowel sounds. The schwa phoneme occurs only in unstressed position. The diphthongs are longest in duration, then there are four long monophthongs, and the remaining vowels including schwa are short, at least in connected speech. Spectrally the Dutch short vowels are not more centralized than the long vowels (for more details see Koopmans-van Beinum, 1980).

The Japanese vowel system is a rather simple one consisting of only five vowels. According to Takebayashi (1975) the vowels /i/ and /u/ are often devoiced when they occur between voiceless consonants and in word-final position. Stress does not seem to play any phonological role in Japanese and it is claimed that all vowels always are pronounced without serious qualitative distortion. However, the incorrectness of the latter claim is proved by De Graaf & Koopmans-van Beinum (1982/83; 1984) who demonstrated a similar degree of reduction in connected speech for a number of languages including Japanese.

As for Dutch we used recorded speech material of the same trained male speaker as in Koopmans-van Beinum (1980). This provided us with the possibility to compare the results of the present procedures with previous results. Nevertheless an important difference remained: in the present speech material we used all segments automatically labelled as being vowel-like in the chosen speech fragment, and also in the order in which they occurred. Moreover measurements were carried out dynamically with ten millisecond steps. This means that frequency of occurrence of all vowels in normal running speech now got the attention it deserves, and that the duration of each occurring vowel weighs proportionally in the calculation of the acoustic system contrast. In the former study ten items of each vowel were used and were measured only at one point more or less in the middle of the vowel. An accidental advantage of the present 'weighing' procedure is the fact that it is no longer necessary to 'label' the vowel segments, i.e. we no longer need to know which vowels the speaker intended to say. The acoustic system contrast ASC of a speech fragment of a specific speaker is defined now by the total variance of all vowels, i.e. of all analysed 10 ms vowel frames, just as they occur in the speech fragment. The moment at which this ASC stabilizes actually defines the length of the speech fragment needed for the determination of the ASC for that specific speaker in that specific speech situation. As to how far length of fragment depends on speaker, on speach situation, and on language is one of the research questions of the project as a whole.

As for the Dutch speaker we made use of two speech situations: free conversation (a 30 sec fragment) and read text (a 10 sec fragment). The speech fragments were selected from existing recordings, which provides us with the possibility to compare the static and dynamic analysis method using fragments of the same recording (not exactly the same fragment). Our decision to confine ourselves to a 30 sec fragment is based on literature indicating that variables concerning the distribution of spectral energy stabilize within that period of time (Li, Hughes, and House, 1969; Zahorian and Rothenberg, 1981). Our choice of only a 10 sec fragment of read text is defined by the results obtained from the free conversation fragment and the need of confining the material.

As for Japanese the speech material of one male speaker (free conversation and read text) has been recorded in Japan recently. Since this speaker was not involved in the earlier studies on the Japanese vowel system, comparison of the earlier analysis results with the results of the dynamic analysis did not make much sense. Therefore in order to answer our questions and at the same time to limit the analysis material we confined ourselves to Dutch conversation (30 sec, only manual analysis), Dutch read text (10 sec, manual and automatic), Japanese conversation (10 sec, only automatic) and Japanese read text (10 sec, manual and automatic).

MEASUREMENTS

By means of a speech editing program (Buiting, 1981) all vowel items in the digitized speech fragments were isolated in such a way that the starting-point of the vowel was considered to be the place were the formant pattern of the vowel was clearly visible in the oscillogram for the first time, and the end was taken to be the point were the specific formant pattern disappeared. In case of adjacent voiced consonants only those successive samples were segmented that did not display any auditorily nor visually observable consonant information. Once the vowel segments were isolated, their durations were of course known as well. From the 30 sec fragment of free conversation 121 vowels could be selected with an average duration of 68.66 ms. From the 10 sec fragment of read text 57 vowels were segmented with an average duration of 71.12 ms. Each vowel segment has been analysed dynamically in 10 ms steps (window size 25.6 ms) by means of a spectral analysis program called QQ (Weenink, 1986) using an LPC order of 12 as a standard.

Apart from a number of other data, not relevant for this study, the program QQ provides us with:

- fundamental frequency (F0) using the sieve algorithm (Duifhuis et al., 1982);
- formant frequencies determined by some optional methods; in our case we used Prony's method for LPC-analysis;
- bandfilter values: a bandfilter analysis based on the FFT amplitude spectrum is carried out with filter specifications given by Sekey and Hanson (1984).

The resulting data are stored in analysis files consisting of successive records, each of them containing the analysis results of one 10-ms vowel frame. In this way all kind of selections and calculations can be carried out in subsequent data processing programs.

With respect to the development of an automatic procedure of data processing, one of the main problems to overcome is the segmentation of vowels from the speech fragment (cf. Kasuya and Wakita, 1979). We therefore designed a procedure in which the spectral analysis precedes the vowel segmentation. The output records are selected as 'vowel' on the basis of three criteria:

- F0-criterium: each data record with F0=0 was rejected (unvoiced);
- high/low ratio (H/L): the definition of low and high frequency areas in literature is not uniform: Weinstein et al. (1975) use L=0-900 Hz and H=3700-5000 Hz; Kasuya & Wakita (1979) use L=0-500 Hz and H=3800-5000 Hz, whereas for Dutch speech material Rietveld (1983) defines L=262-2230 Hz and H=5575-11150 Hz. In the present study we used the filters 1-6 for the low frequency area (92-856 Hz) and the filters 13, 14, and 15 for the high frequency area (2549-4239 Hz), since filter 16 turned out not to be reliable in all cases. So if the ratio H/L>1 then the data record is rejected as a vowel record.
- vocal tract length VTL: based on the analysis results of QQ this program calculates also the VTL per record (Wakita, 1977). Considering the formant frequencies and VTL together revealed that in case of low (nasal) F1 the VTL showed very unreal values (0.0 or -1.0 cm), whereas for records with an extreme high F1 value (e.g. F1>1500 Hz) the calculated VLT attained to about 10 cm. All other records display a more or less normal distribution of VTL values. Although this VTL criterium needs some more refinement, we obtained satisfactory results in this study by using the criterium that each vowel record had to attain a VTL value of:

  $$\overline{\mathrm{VTL}} - 0.5{*}\mathrm{s.d.} \leq \mathrm{VTLx} \leq \overline{\mathrm{VTL}} + 1.0{*}\mathrm{s.d.}$$

  in which VTLx = the VTL of data record x.

Within the automatic vowel segmentation program the following hierarchy of criteria is used: 1) a first selection is done based on the F0- and the high/low ratio criterium; 2) a second selection is done based on the VTL-criterium, applied to the remaining records.

Both procedures (manual and automatic) end up in a set of data processing programs calculating cumulatively (i.e. record after record) the mean values and the variance of the fundamental frequency, of the first four LP-formants, and of the 16 bandfilter values. During processing the mean F1, F2, mean bandwidths of each formant, mean level of each bandfilter, F0, and the ASC are stored in an output file, together with the deviation of the new ASC compared to the preceding ASC value, each time when a record is closed. At the end of the processing the output consists of the final mean values with variance of the parameters mentioned above, and the total number of processed records (= the number of 10 ms vowel frames).

The program provides the possibility of cumulatively processing the acoustic system contrast ASC, and of defining the moment when the ASC value stabilizes.

RESULTS AND CONCLUSIONS

The results of the manual and the automatic procedure, compared with the results from our preceding studies, are described in detail in Koopmans-van Beinum and De Saint Aulaire, 1986 and will be presented at the Congress. Summarizing the results we can state that

1) the measure for acoustic system contrast ASC, cumulatively processed on the output data of a dynamic acoustic-phonetic vowel analysis, compares favourably with the ASC-values as processed on output data from static vowel analysis; it should be kept in mind, however, that in the dynamic method all vowel segments (including diphthongs and schwa) are processed in their total duration;

2) the here presented automatic procedure for processing running speech provides us for the time being with a satisfying possibility to define quickly and for extended speech material, global reduction data in terms of acoustic system contrast. Moreover our methods provide tools to recognize where discontinuities in the global measures occur in the processed speech material, possibly caused by accentuation and indicating important events in running speech.

For the analysed Japanese speech material the resulting ASC-values fit well with the previous results of the static analysis method on speech material of three other speakers. Nevertheless the manual and automatic procedure, applied on the Japanese read text, display slight differences in the results, possibly caused by our poor knowledge of the Japanese language necessary for proper manual segmentation.

For Dutch as well as for Japanese free conversation the measure for acoustic system contrast stabilizes within 2 sec. of vowel material, which means for both languages that about 6 sec. of free conversation should be enough. The high number of schwa sounds in Dutch will cause a lower ASC-value than in languages without schwa phonemes. In free conversation this is confirmed in our data, but more speakers are needed to prove the reliability, since vowel reduction turns out to be greatly speaker dependent. In the read texts of both languages, the fluctuations in ASC-value are much more persistent, mainly caused by F2 fluctuations which are more violent for Japanese than for Dutch. Here again we can possibly trace the influence of the frequently occurring schwa sounds in Dutch.

In the near future our research will concentrate on combining the dynamically processed global measure for acoustic system contrast with local parameter values in order to be able to control the influence of vowel reduction aspects on speech recognition and speech synthesis.

REFERENCES

Buiting, H.J.A.G. (1981). SESAM, Speech Editing System Amsterdam, IFA-report 70, Amsterdam.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. IRAL 7, 295-325.

Duifhuis, H., Willems, L.F. & Sluyter, R.J. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. J. Acoust. Soc. Am. 71, 1568-1580.

Gay, T. (1977). Effect of speaking rate on vowel formant movements. Haskins Lab. Status Report on Speech Research SR-51/52, 101-117.

Graaf, T. de & Koopmans-van Beinum, F.J. (1982/83). Vowel contrast reduction in Japanese compared to Dutch. IFA-Proceedings 7, 27-38.

Graaf, T. de & Koopmans-van Beinum, F.J. (1984). Vowel contrast reduction in terms of acoustic system contrast in various languages. IFA-Proceedings 8, 41-53.

Kasuya, H. & Wakita, H. (1979). An approach to segmenting speech into vowel-like and nonvowel-like intervals. IEEE Trans. ASSP-27, 319-327.

Katwijk, A.F.V. van (1974). Accentuation in Dutch. Diss. RU Utrecht.

Koopmans-van Beinum, F.J. (1980). Vowel Contrast Reduction. Diss. Univ. of Amsterdam.

Koopmans-van Beinum, F.J. & Saint Aulaire, R.P. de (1986). A method for the dynamic determination of acoustic vowel contrast. IFA-Proceedings 10, 1-17.

Li, K.P., Hughes, G.W. & House, A.S. (1969). Correlation characteristics and dimensionality of speech spectra. J. Acoust. Soc. Am. 46, 1019-1025.

Lindblom, B.E.F. (1963). Spectrographic study of vowel reduction. J. Acoust. Soc. Am. 35, 1773-1781.

Os, E.A. den (1985). Vowel reduction in Italian and Dutch. PRIPU 10-2, 3-12.

Rietveld, A.C.M. (1983). Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands. Diss. KU Nijmegen.

Rietveld, A.C.M. en Koopmans-van Beinum, F.J. (to appear). Vowel reduction and stress. Speech Communication.

Saint Aulaire, R.P. de (1986). Klinkerreductie en taalstructuur: een verwerkingsmethode. IFA-report 85, Amsterdam.

Sekey, A. & Hanson, B.A. (1984). Improved 1-Bark bandwith auditory filter. J. Acoust. Soc. Am. 75, 1902-1904.

Takebayashi, S. (1975). The vowels of Japanese and English. Lexicon 4, 49-67.

Wakita, H. (1977). Normalisation of vowels by vocal tract length and its application to vowel identification. IEEE Trans. ASSP-25, 183-192.

Weenink, D.J.M. (1986). QQ: een programma voor analyse, resynthese en herkenning van klinkersegmenten. IFA-report 82, Amsterdam.

Weinstein, C.J., Mc Candless, S.S., Mondshein, L.F. & Zue, V.W. (1975). A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. ASSP-23, 54-67.

Zahorian, S.A. & Rothenberg, W. (1981). Principal-component analysis for low-redundancy encoding of speech spectra. J. Acoust. Soc. Am. 69, 832-845.

THE ROLE OF SYLLABLE STRUCTURE IN THE ACOUSTIC REALIZATIONS OF STOPS*

Mark A. Randolph and Victor W. Zue

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

ABSTRACT

This paper examines the role of the syllable in the description of systematic acoustic-phonetic variations. We present results of an acoustic study based on over 5,000 stops collected from 1,000 sentences spoken by 100 talkers. Our results indicate that the acoustic properties of stops depend on the syllable locations in which they appear. On the basis of these results we propose a syllable-based rule framework in order to describe acoustic-phonetic variations in categorical as well as continuous terms. Implications to linguistic and speech recognition research are discussed.

INTRODUCTION

It is well known that the acoustic characteristics of speech sounds vary according to the context in which they appear. Traditionally, *systematic* acoustic variation has been described using context-sensitive *rewrite rules* of the form: $A \rightarrow B \ / \ C \_ D$, where elements A, C, and D correspond either to individual phonemes or classes of phonemes and element B corresponds to a specific phonetic realization [2]. As an example, rule (1) states that voiceless stop consonants are aspirated when followed by vowels.

$$\left\{\begin{matrix} p \\ t \\ k \end{matrix}\right\} \rightarrow \left\{\begin{matrix} p^h \\ t^h \\ k^h \end{matrix}\right\} / \_ V \qquad (1)$$

There are at least two disadvantages associated with such a rule description. First, it is awkward to describe the important role played by larger phonological units such as syllables or metrical feet. Second, it implicitly assumes that variations can be described in categorical terms, despite the fact that many acoustic changes are inherently continuous.

This paper proposes an alternative framework for describing acoustic-phonetic modifications. Central to this description is the notion of the syllable. We show how a rule framework based on the syllable may be augmented so as to describe contextual variations both concisely and accurately. We describe a set of acoustic studies focusing on the stop consonants in American English, and show that the proposed framework is well suited for interpreting the results. Finally, we describe the implications of the proposed framework for linguistic and speech recognition research.

THE SYLLABLE FRAMEWORK

The notion that phonological rules may be sensitive to syllable structure has been suggested by many linguists. Kahn [6], for example, argues that allophonic variation and phonotactic constraints can be described more effectively using a syllable-based phonological framework. Fujimura and Lovins [4] have provided articulatory data along with a summary of a number of acoustic-phonetic studies which provide concrete support for the syllable. Nakatani and Dukes [8] provide evidence from the perceptual domain. Their experiments indicate that the syllable-initial and syllable-final allophones of phonemes provide important perceptual cues for word juncture and that humans may rely on this kind of information for parsing phonetic sequences into words. While these studies provide compelling evidence in support of a syllable-based phonological representation, we are still in need of considerably more acoustic-phonetic data: quantitative results, derived from a large body of speech, showing that the surface acoustic realizations of phonetic units are governed by their positions within this unit.

In the next section, we show that if structured in the proper way, these results could be particularly relevant to the notion of a syllable hierarchy [3] [10], a structural description of the syllable in terms of an immediate constituent grammar. Linguists have found this hierarchical description important for the concise statement of phonotactic restrictions. As we will discuss later in this paper, this hierarchical representation also provides an effective means of incorporating the syllable into a description of acoustic-phonetic modifications.

THE CURRENT INVESTIGATION

We begin by describing the syllable template shown in Figure 1. We have used this template to label our experimental database and for the subsequent interpretation of our results. The form of this template closely resembles the syllable hierarchy proposed by Fudge [3]. We have modified his template by positing three affix positions and by providing labels for the *outer-onset*, *inner-onset*, *inner-coda*, and *outer-coda* positions. In addition, we have added an additional slot to the onset for the phoneme, /s/, which forms syllable-initial clusters with nasals, stops, and stop-semivowel sequences. The other terminal elements of this hierarchy are manner of articulation classes.

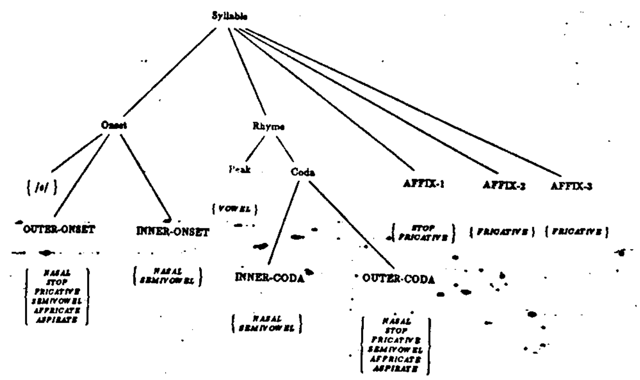


Figure 1: Syllable-constituent structure described in terms of br[illegible] phonetic categories

The Acoustic Study

Data for our acoustic study has been obtained from 1,000 sentences spoken by 100 talkers (50 male and 50 female). The corpus was the first five hundred of the well-known Harvard list of phonetically-balanced sentences. During recording, lists of ten sentences were read by one male and one female talker. For all the collected data, both phonemic and phonetic transcriptions were provided and aligned with the waveforms. In addition, syllable boundaries and lexical stress markers were inserted in the transcriptions. From this database, a sample of approximately 5,200 stops was extracted for the present set of experiments.

For each stop, we measured the closure duration and the release duration (VOT) separately. We also measured the durations of adjacent phonemes. From these measurements and transcriptions, we were able to determine whether a stop was released, unreleased, or deleted. We marked a stop as released if its release duration was greater than zero, unreleased if the release duration equalled zero, and deleted if the stop was present in the phonemic transcription, but absent in the phonetic. We should note that a stop was transcribed as unreleased if it could not be heard, and if a noticeable burst could not be observed from either the waveform or the spectrogram by the transcriber. In addition to duration measurements, we also computed several energy related parameters in order to infer the relative strength of a stop's release.

We are primarily interested in quantifying the effects of a stop's syllable position on these acoustic properties. However, we are also interested in understanding any possible influence of local phonetic context. In order to reduce the number of categories of local phonetic context to a reasonable size, we grouped the phonemes forming each stop's left and right context into seven equivalence categories corresponding roughly to manner of articulation. These categories are: Vowel (V), Semivowel (G), Nasal (N), Fricative (F), Stop (S), Affricate (A), and Aspirate (H).

Stops were categorized according to both local phonetic context and syllable position. Space limitations prohibit us from presenting data for all combinations of these two factors. Instead, we will present three examples from this larger pool of results. We will examine stops in two local phonetic environments, for each, we will examine the effect of syllable position on a stop's acoustic properties. In a third example, we examine the effect of post-vocalic voicing on vowel duration, also as a function of the stop's syllable position.

Results

Our first set of results compares intervocalic singleton stops in the outer-onset verses outer-coda positions. There were 668 outer-onset stops in this local phonetic environment, of which, 96% were released. In contrast, only 65% of 168 outer-coda stops were released. For singleton stops in the outer-onset, VOT is a reliable measure for voicing contrast. This can be seen from the histograms for voiced and voiceless stops shown in Figure 2. For syllable-final voiceless stops that were released (also shown in Figure 2), VOT is substantially reduced, such that there is considerable overlap of the distributions for outer-onset voiced stops and outer-coda voiceless stops.

The second example involves stop-semivowel sequences appearing between two vowels, i.e. the $V \_ GV$ context, where the stop is voiceless. In the outer-onset position (e.g., in the word sequence "gray train"), about 98% of the stops were released. On the other hand, only about 45% of the stops were released when they appeared in the outer-coda position (e.g. "great rain"). In Figure 3 we have plotted VOT versus the averaged total energy within the release for voiceless stops in both the outer-onset and outer-coda positions. We see that syllable-initial stops generally have releases that are both longer and stronger than their syllable-final counterparts.

Our final example concerns the effect of voicing of a stop on the duration of a preceding vowel. It is well known that the duration of the vowel is influenced by the voicing characteristic of the following consonant (e.g, the vowel in "bag" is longer than the vowel in "back") [9]. However, there seems to be evidence from our study that such influence is conditioned upon whether the vowel and stop belong to the same syllable. When the stop is in the outer-coda position, the preceding vowel is lengthened when the stop is voiced. However, the trend is reversed when the stop is in the onset of the following syllable. These results are summarized in Figure 4.

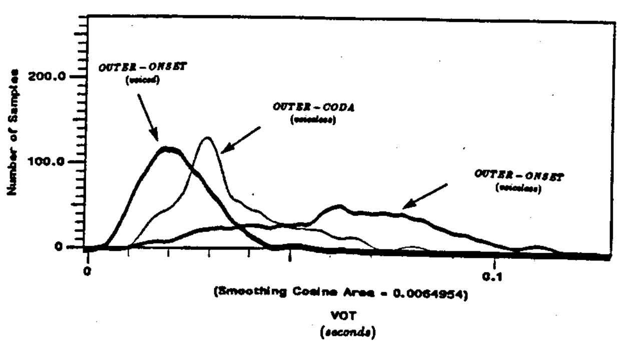


Figure 2: Influence of syllable position on the VOT of intervocalic, singleton stops

*This research was supported by ONR under contract N00014-82-K-0727, monitored through Naval Electronic Systems Command and AT&T Bell Laboratories Cooperative Research Fellowship Program

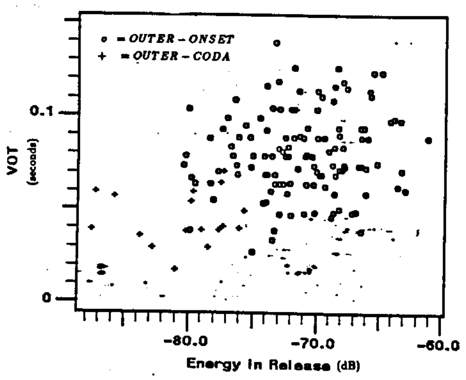


Figure 3: Influence of syllable position on voiceless stops in the *V* _ *GV* context

DISCUSSION

From the results of our experiments, we may conclude that the acoustic characteristics of stop consonants depend on their positions within the syllable. However, our results also indicate that a more accurate description of these acoustic modifications may require an alternative rule framework in which acoustic information in the form of parameter values can be accommodated.

The Proposed Framework

The first aspect of our proposal is inspired by the work Church [1] and is motivated by principles of *information factoring*. The idea is to encode the description of a phoneme's contextual environment in terms of the syllable hierarchy. As a result, it becomes possible to replace a phonological grammar consisting of context sensitive rules by one which is context free. In general, context free grammars describe languages that are easier to parse, and in many cases, provide a more concise statement of phonological rules. For example, rather than inserting syllable boundary markers into a rule to describe the syllable positions for which stops are aspirated, one may describe these contextual environments more succinctly by restricting aspirated stops to particular slots within the template shown in Figure 1.

These new rules, however, since they ignore quantitative acoustic differences between phonemes appearing in the various syllable positions, still do not provide an adequate description of the facts. For example, aspirated stops can appear both in the outer-onset and outer-coda positions, but with differences in VOT that turn out to be important for determining the syllable structure of an utterance. The second aspect of our proposal is to augment this categorical representation with an acoustic description.

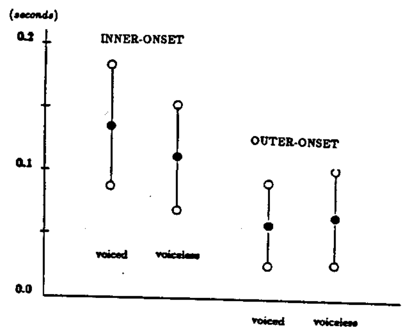


Figure 4: Influence of voicing and syllable position on preceding vowel duration

A more accurate mechanism would be to state these rules in the form of a conditional probability function such as the one shown in Equation (2).

$$p\left[\vec{A}|S,\sigma,\alpha,\beta\right] \tag{2}$$

The vector quantity $\vec{A}$ in this "rule" is a set of acoustic properties, some of which may be discrete (e.g., released, deleted, etc.), others may be continuous (e.g., VOT, the measured energy in release, etc.). The conditioning variables in this rule, or explanatory factors, are phonological in nature and reflect the phonemic identity of a segment and its phonological context. For example, the factor S in this rule may denote a particular phoneme (e.g., /p/, /t/, /k/, etc.) or a phoneme class (e.g., *STOP*, *FRICATIVE*, etc.), σ denotes S's syllable position (e.g., *outer-onset*, *inner-onset*, *peak*, etc.), and α and β specify the left and right context, respectively.

Since it attempts to describe the acoustic properties of phonemes directly, this rule framework bypasses an allophonic description of the speech waveform and therefore suggests a paradigm for research that is a hybrid of traditional phonetics and phonological methodologies [7]. The task involved in rule discovery is to seek a parsimonious combination of explanatory factors that best account for the acoustic-phonetic data. These steps would be carried out within the context of an acoustic study like the one described above.

Implications for Automatic Speech Recognition

The applicability of this probabilistic rule framework for automatic speech recognition may be readily seen by straightforward manipulations of the quantity shown in Equation (2). For example, given a particular syllable hypothesis, and a hypothesized local context, the *a posteriori* probability of a particular segment hypothesis is $p\left[S|\vec{A},\sigma,\alpha,\beta\right]$, and may be obtained using Bayes rule. In this function, the vector $\vec{A}$ denotes some appropriate set of acoustic parameters designed to identify S.

The quantity given in Equation (2) may also be useful for lexical retrieval. Church proposed a speech recognition framework in which a *narrow* phonetic transcription is parsed into syllables prior to lexical retrieval, using extrinsic allophonic variation as a means of constraint. The practical limitation of Church's approach is that it may not be possible to obtain such a detailed phonetic transcription from an acoustic front-end. However, a partial phonetic description of the speech signal in the form of a broad phonetic transcription consisting of a sequence manner categories, may be a more realistic alternative. This approach has been suggested by Huttenlocher and Zue [5] for the task of large vocabulary isolated word recognition.

Church's grammar would have to be rewritten, more along the lines of the syllable template shown in Figure 1. The direct consequence is a grammar which has a higher degree of ambiguity. Figure 5 shows the result of parsing the broad phonetic transcription of the phrase, "black lead." The output is provided in the form of a *syllable lattice*: a set of arcs (shown as rectangular boxes) spanning the input string. The arcs are labelled with the names of syllable constituents corresponding to what the parser has hypothesized: For this example, we see that the phoneme /k/ can be parsed as either the outer-coda of the first syllable or the outer-onset of the second. Such ambiguity arises because detailed phonetic information is no longer available. From Figure 3, however, we note that a voiceless stop in the outer-coda position will have reduced VOT and energy compared to its outer-onset counterparts. For this example, these attributes can be confirmed from the spectrogram in Figure 5.

Our approach to reducing the number of competing syllable hypothesis is to select a set of appropriately chosen acoustic attributes (e.g. VOT for stops) and to use the *a posteriori*, probability $p\left[\sigma|\vec{A},S,\alpha,\beta\right]$, to aid in disambiguating a parse. We believe that such a strategy offers the advantage of not requiring a detailed transcription to be available, while directly making use of acoustic measurements that are potentially more accurate. Efforts in implementing such a recognition strategy is currently under way.

SUMMARY

We have examined the role of syllable structure in the acoustic realizations of stop consonants in American English. The results of our acoustic study indicate that much of the apparent variability that a stop is subject to, may be explained in terms of its position within the syllabic unit. We have proposed a rule framework that is intended to capture this variability both concisely and accurately. Each rule in our framework is stated in the form of a conditional probability function. The conditioning variables (i.e., each rule's input) represent both the underlying phonemic identity of a segment and its phonological context. The rule's output is a description of its acoustic consequences. Finally, the relevancy of our proposal to linguistic and automatic speech recognition research was discussed.

REFERENCES

[1] Church, K. W., "Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints," Ph.D. Thesis, Massachusetts Institute of Technology, January 1983.

[2] Cohen P.S., and Mercer, P.L., "The phonological component of an Automatic Speech Recognition System," in *Speech Recognition*, R. Reddy, Ed., Academic Press, New York, pp. 275-320.

[3] Fudge, E.C., "Syllables, " *Journal of Linguistics*, Vol. 5, pp. 253–286.

[4] Fujimura, O. and Lovins, J., "Syllables as Concatenative Units," Indiana University Linguistics Club, 1982.

[5] Huttenlocher, D.P. and Zue, V.W., "A model of Lexical Access from Partial Phonetic Information," Proc. ICASSP, 1984

[6] Kahn, D., "Syllable-based Generalizations in English Phonology," Ph.D. Thesis, Department of Linguistics, Massachusetts Institute of Technology, September 1977.

[7] Liberman, M.Y., "In Favor of Some Uncommon Approaches to the Study of Speech," in *The Production of Speech*, MacNeilage, P.F., Ed., Springer-Verlag, New York, 1983.

[8] Nakatani, L., and Dukes, K.D., "Locus of Segmental Cues for Word Juncture," *J. Acoust. Soc. Am.*, Vol. 62, no. 3, pp. 714–719.

[9] House, A.S. and Fairbanks, G., "The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels," *J. Acoust. Soc. Am.*, Vol. 25, pp. 105–113.

[10] Selkirk, L.O, "The Syllable," in *The Structural of Phonological Representations*, Part II, Foris Publications, Dordrecht, Holland, pp. 337–383.

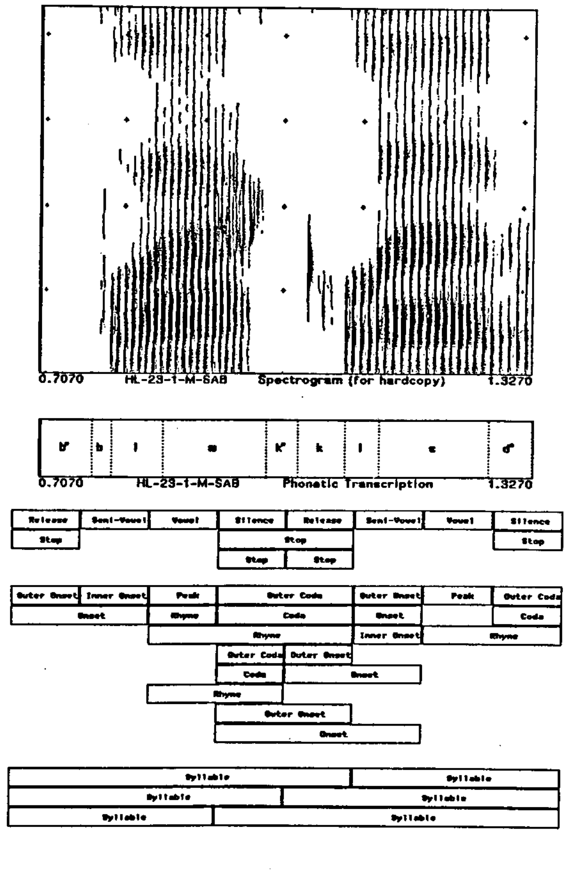


Figure 5: Syllable lattice generated from the broad syllable parser.

THE USE OF TEMPORAL FREQUENCY IN SPEECH SIGNAL ANALYSIS

DAVID A. SEGGIE

Department of Phonetics and Linguistics, University College London,
London NW1 2HE, U.K.

ABSTRACT

An analytic signal representation enables the estimation of speech signal temporal frequency. The use of this time-domain attribute in speech signal analysis is illustrated. In addition, the relationship between a signal's temporal frequency and its spectral composition is elucidated.

INTRODUCTION

In speech signal analysis, a basic goal is to extract from the signal those acoustic attributes useful in signifying phonetic contrasts. Given the fact that these attributes appear to be encoded in the speech signal in a highly complicated manner, attempts at achieving this goal often involve the transformation of the data into what is thought to be a more appropriate representation - appropriate in the sense that salient acoustic characteristics are brought to the fore. For example, models of the acoustics of speech production [1] and studies of the frequency selectivity of hearing [2], indicate that one such appropriate representation of the speech waveform is in terms of its short-term amplitude spectra. Salient acoustic features, (e.g. temporal variations in formant frequencies), can then be readily estimated from these spectra.

However, the demonstrated utility of established speech signal representations should not prohibit the assessment of novel ways of viewing speech pressure waveforms. Indeed, given that a phonetic contrast is usually signalled by many different acoustic parameters, it would seem eminently sensible to view the waveform in several different ways in order to uncover the overall acoustic pattern .

Recent work in both seismic signal processing [3,4] and ultrasonic imaging [5,6], indicates that the temporal frequency characteristics of acoustic signals encode useful information. In speech signal processing, preliminary studies [7,8] point to the possibility of extracting phonetically relevant information from this particular time-domain signal attribute. Temporal frequency, (sometimes referred to as instantaneous frequency), is defined via an analytic signal representation [9]. The analytic signal is a complex-valued function of time defined as,

$$a(t) = s(t) + j\tilde{s}(t)$$

where $j = \sqrt{-1}$, s denotes the real speech pressure waveform, and $\tilde{s}$ is the Hilbert transform of s. Manipulation of a(t) allows the unique separation of the speech signal into time-domain envelope and phase. Instantaneous envelope, e(t), is defined as,

$$e(t) = \mathrm{mod}\left[a(t)\right] = \sqrt{(s^2(t) + \tilde{s}^2(t))}$$

Instantaneous phase, $\phi(t)$, is given by,

$$\phi(t) = \arg\left[a(t)\right] = \arctan\left[s(t)/\tilde{s}(t)\right]$$

Temporal frequency (in radians) is simply the time derivative of instantaneous phase, i.e.,

$$\omega(t) = d\phi(t)/dt$$

Note that instantaneous phase as defined above is modulo 2π , and shows discontinuities whenever it extends beyond $\pm\pi$. Therefore, prior to differentiation, a standard "unwrapping" algorithm was applied in order to extract the desired continuous phase function [10].

The analytic signal and the time-domain signal attributes derived from it can be understood in the following way. a(t) can be thought of as the path traced out in complex space by a vector whose length and rate of rotation vary as a function of time. e(t) decribes the temporal variations in the length of the vector, and can be regarded as a measure of the instantaneous strength of the speech signal. $\omega(t)$ describes the temporal variations in the vector's rate of rotation. This time-domain function can be used as a measure of speech signal continuity.

The temporal frequency characteristics of speech signals are illustrated in Figs. 1 - 4. Figure 1 shows a speech pressure waveform for the simple VCV token [ɑ:dɑ:]. The waveform was low-pass filtered, (cut-off frequency = 8.4 kHz), and digitized at a sampling frequency of 20 kHz, to a maximum amplitude resolution of 12 bits. Figure 2 shows the temporal frequency function of the signal depicted in Fig. 1. Figure 3 shows a speech pressure for the utterance [tu:zɪərəʊ] ("two zero"); same speaker and data acquistion conditions as in Fig. 1. Figure 4 is the temporal frequency function for the signal shown in Fig. 3. These figures show that the quasi-periodic and noisy regions of the speech waveforms associated with sonorant and non-sonorant segments respectively, are clearly delineated by marked changes in both the structure and mean value of the temporal frequency.

MEAN TEMPORAL FREQUENCY

Although there is no one-to-one correspondence between time-domain and Fourier-domain frequencies, mean temporal frequency can be related to the spectrum of the speech signal This relationship, outlined by Vile [11] (see also [12]), can be made more general in order to apply to speech data segments of arbitrary duration.

Without loss of generality, a speech signal segment of duration T, centred at $t = \tau$, can be modeled as,

$$s(\tau;t) = \mathrm{Re}[e(t)\exp(j\phi(t))]w(T,\tau;t) \qquad (1)$$

where e(t) is a non-negative envelope function, $\phi(t)$ is a phase function, and ,

$$w(T,\tau;t) = \begin{cases} 1 & \tau - T/2 < t > \tau + T/2 \\ 0 & \text{otherwise} \end{cases}$$

The mean Fourier-domain frequency of the data segment, f_τ, can be defined as,

$$f_\tau = \frac{\int_0^\infty f|S(\tau;f)|^2 df}{\int_0^\infty |S(\tau;f)|^2 df} \qquad (2)$$

where $S(\tau;f)$ is the Fourier transform of $s(\tau;t)$. Since $s(\tau;t)$ is real, $|S(\tau;f)|$ is an even function. Consequently, the integration in equ. (2) ranges over the positive frequencies only, to ensure a non-zero f_τ value. Given that the analytic signal can be written as,

$$a(\tau;t) = s(\tau;t) + j(1/\pi t)\bigstar s(\tau;t)$$

where $\bigstar$ denotes the convolution operator, it follows that,

$$a(\tau;t) = 2[\delta(t)/2 + j/2\pi t]\bigstar s(\tau;t)$$

i.e.,

$$A(\tau;f) = 2H(f)S(\tau;f)$$

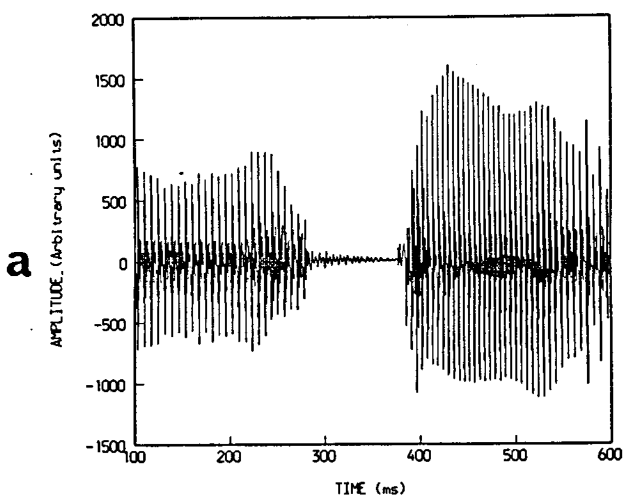
where δ is the Dirac-delta function, H is the Heaviside unit step function, and $A(\tau;f)$ is the Fourier transform of $a(\tau;t)$. Using this one-sided property of $A(\tau;f)$, equ. (2) can be re-written as,

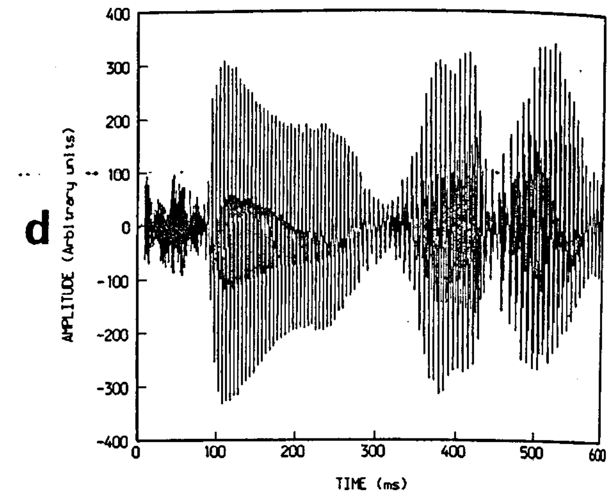
$$f_\tau = \frac{\int_{-\infty}^{\infty} f|A(\tau;f)|^2 df}{\int_{-\infty}^{\infty} |A(\tau;f)|^2 df}$$

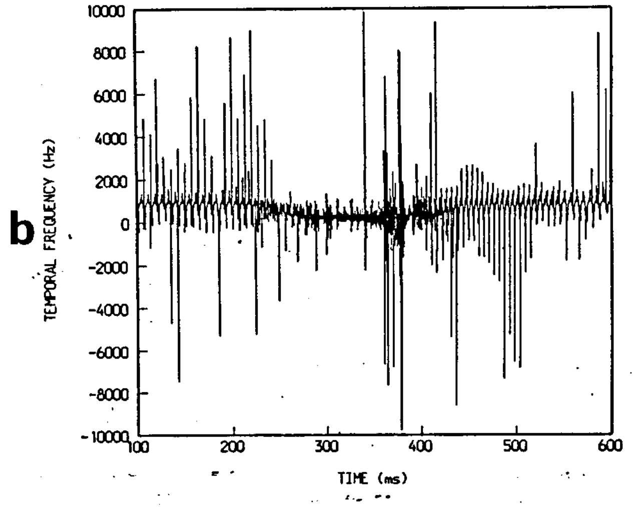
$$= \frac{\int_{-\infty}^{\infty} fA(\tau;f)A^*(\tau;f)df}{\int_{-\infty}^{\infty} |A(\tau;f)|^2 df}$$

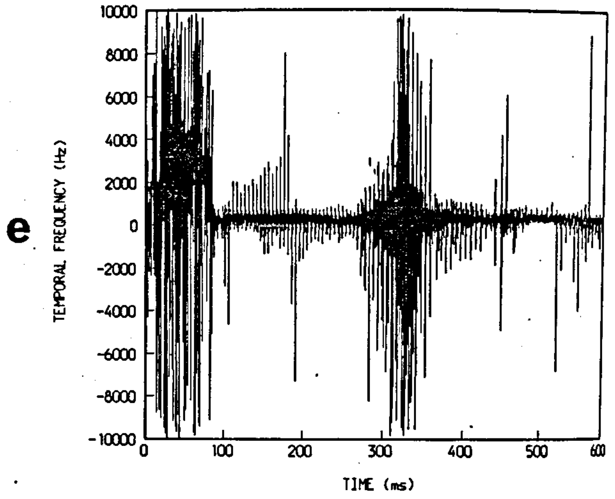
where * denotes complex conjugate. Using the derivative theorem, and expressing $A(\tau;f)$ as a Fourier integral gives,

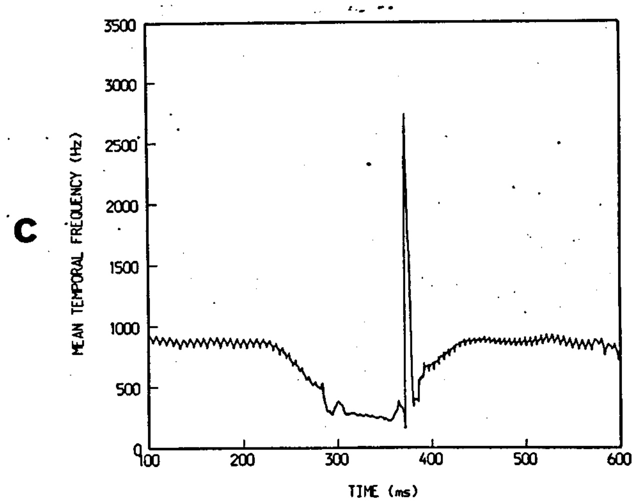
$$f_\tau = \frac{\int_{-\infty}^{\infty} df \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} dt' \dot{a}(\tau;t') a^*(\tau;t) \exp[2\pi jf(t-t')]}{2\pi j \int_{-\infty}^{\infty} |A(\tau;f)|^2 df}$$

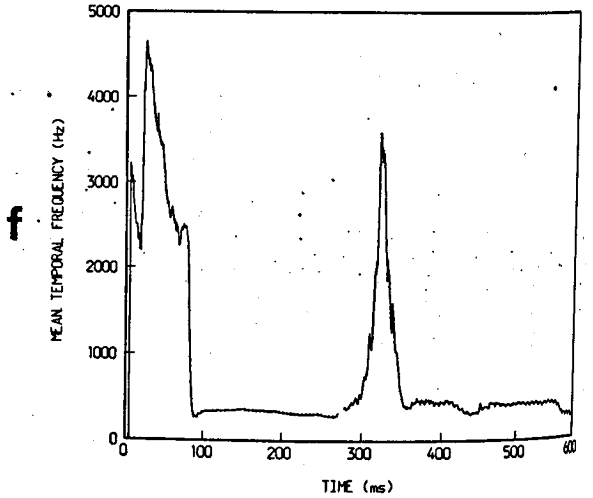












a – Fig. 1 Speech signal for [ɑ:dɑ:]
b – Fig. 2 Temporal frequency function for signal in Fig. 1
c – Fig. 5 Mean temporal frequency function for signal in Fig. 1

d – Fig. 3 Speech signal for [tu:zɪərəʊ] *("two zero")*
e – Fig. 4 Temporal frequency function for signal in Fig. 3
f – Fig. 6 Mean temporal frequency function for signal in Fig. 3

where ˙ denotes time derivative. From above it follows that,

$$f_\tau = \frac{\int_{-\infty}^{\infty} dt\dot{a}(\tau;t')a^*(\tau;t)}{2\pi j \int_{-\infty}^{\infty} |A(\tau;f)|^2 df}$$

Using the signal representation given in equ. (1),

$$f_\tau = \frac{\int_{\tau-T/2}^{\tau+T/2} [e(t)\dot{e}(t) + e^2(t)\dot{w}(t) + j\dot{\phi}(t)e^2(t)]dt}{2\pi j \int_{\tau-T/2}^{\tau+T/2} |A(\tau;f)|^2 df}$$

Assuming $e(\tau+T/2) \simeq e(\tau-T/2)$, the above expression reduces to,

$$f_\tau = \frac{\int_{\tau-T/2}^{\tau+T/2} \dot{\phi}(t)e^2(t)dt}{2\pi \int_{\tau-T/2}^{\tau+T/2} |A(\tau;f)|^2 df}$$

Using Rayleigh's theorem,

$$\overline{f_\tau} = \frac{\int_{\tau-T/2}^{\tau+T/2} \dot{\phi}(t)e^2(t)dt}{2\pi \int_{\tau-T/2}^{\tau+T/2} |a(\tau;t)|^2 dt}$$

$$= \overline{\dot{\phi}}/2\pi = \overline{\omega_\tau}/2\pi$$

That is, for a speech signal segment of arbitrary duration, the centre of gravity of the power spectrum is equal to the envelope. squared-weighted temporal frequency. Using the above expression, the time evolution of the mean temporal frequency for the signals shown in Figs. 1 and 3 was computed; the results are shown in Figs. 5 and 6 respectively. In both cases the window duration was 10 ms. Figures 5 and 6 show that plots of $\overline{\omega}(t)$ highlight the differences in the spectral composition of speech signal segments associated with sonorant and non-sonorant sounds. Note particularly the very clear delineation of the plosive release in Fig. 5.

DISCUSSION

Appropriate manipulation of speech signal temporal frequency enables the estimation of the time evolution of the centre of gravity of the signal's short-term power spectra, without the computational effort involved in moment calculation from the Fourier transforms of many short data segments. Initial results indicate that plots of $\overline{\omega}(t)$ may be useful in automatically segmenting speech waveforms; particularly in determining the presence of plosives. One other interesting aspect of this study is the presence of large, time-localized fluctuations in speech temporal frequency functions, (see Fig. 2 & 4). Work in ultrasonic signal processing has shown that an analysis of such features yields information which is of use both in imaging and in signal parameter estimation [6]. The possibility that speech signal temporal frequency structure encodes similarly useful information is being investigated.

REFERENCES

[1] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signals", Prentice-Hall, 1978.
[2] B. C. J. Moore, "Frequency selectivity in hearing", Academic Press, 1986.
[3] M. T. Taner et al., "Complex trace analysis", Geophysics, vol. 44, pp. 1041-1066, 1979.
[4] R. L. Kirlin er al., "Enhancement of seismogram parameters using image processing techniques", Geoexploration, vol. 23, pp. 41-76, 1984.
[5] D. A. Seggie, S. Leeman, "Deterministic approach towards ultrasound speckle reduction", IEE Proc., vol. 134, Pt. A, no. 2, pp. 188-192, 1987.
[6] D. A. Seggie, S. Leeman, G.M. Doherty, "Time domain phase: a new tool in ultrasound imaging", Mathematics and Computer Science in Medical Imaging, Springer-Verlag, in press.
[7] C. Berthomier, "Instantaneous frequency and energy distribution of a signal Sig. Proc. vol. 5, pp. 32-45, 1983.
[8] D. A. Seggie, "The application of analytic signal analysis in speech processing", Proc. IOA, vol. 8, Pt. 7, pp. 85-92, 1986.
[9] D. Gabor, "Theory of communication", J. Inst. Elect. Eng., vol. 93, Pt. 1, pp. 429-441, 1946.
[10] A. V. Oppenheim, R. W. Schafer, "Digital signal processing", Prentice-Hall, pp. 507-509, 1975.
[11] J. Vile, "Theories et applications de la notion de signal analytique", Cables et Transmissions, vol. 1, pp. 61-74, 1948.
[12] L. Mandel, "Interpretation of instantaneous frequencies", Am. J. Phys., vol. 42, pp. 840-846, 1974.

A PRIMARY EXPERIENCE:
THE VECTOR QUANTIZATION TECHNIQUE IS AN EFFECTIVE TOOL FOR PHONETIC RESEARCH

ZHAO GUOTIAN

Speech Technology Lab.
Dept. of Radioengineering
Harbin Institute of Technology
Harbin, P.R. China

ABSTRACT

The effective phonetic symbol system representing the phonetic feature of speech exactly is important tool for speech processing technique. But the present spelling symbol system of Chinese is suitable for teaching only, but not for speech technology. In this paper, as a beginning, we have investigated the spelling symbol [i] in different environments by the method of vector quantization, which follows consonants [j, q, x, z, c, s, zh, ch, sh, t, b, p, l, m, n, r, y]. The results show that Chinese symbol [i] can be represented by International Phonetic Symbols [i, ɿ, ʅ] in detail.

INTRODUCTION

In phonetics, phonetic symbols are used to indicate phoneme, for instance, International phonetic symbols. The phonetic symbols system which is established for one language, in teaching language and studying pronunciation and correctly expressing language contents with speech, is very sufficient and clear for hearing with which people are satisfied. But, in the field of speech signal processing which is related to phonetics closely, the phonetic symbol system for teaching speech and studying pronunciation becomes too rough to represent the similarities and differences between symbols. Specially the Chinese spelling symbol system was established for Chinese teaching and it is simplitic. The requirements for speech signal processing can not be content with it. Therefore, the meticulous classification for phonetic symbols is required in the field of speech signal processing. For example, in Japanese, for phonetic symbol [n], only one is applied to teaching speech and studying pronunciation. But in speech signal processing, for instance, in digits recognition, it is divided into two [2].

There are 32 phonetic symbols in modern Chinese [1]. For some symbols, the problems similar to above exist, and is even more serious as compared with other languages, because the present symbols for Chinese are more sketchy than international symbol, probably. So it is necessary to research Chinese phonetic symbols deeply and meticulously to meet the needs of speech signal processing technique. In this paper, [i] of Chinese phonetic symbol is analysized and reseached. The method used is vector quantization.

The similarities and differences of [i] are obtained under the various possible environments. Finally, [i] is divided into three [i], [ɿ] and [ʅ] in detail. They are suitable for speech processing, in present stage.

THE METHOD FOR ANALYSIS

As feature parameters, the linear prediction coefficient cepstrum are applied to our research. The experience demonstrated that they are more sufficient feature parameters for speech recognition.

The vector quantization technique has been developed in the field of speech compression for communication. But recently this technique has also been introduced to the field of speech recognition. The basic concept of vector quantization applied to speech compression is schematically in Fig. 1. Both training and input vectors are the same kind of speech feature vectors. In our research, the capacity of codebook is 64, and convergence threshold is 0.01 [3]. Each input vector is compared with the codewords in codebook and then is endowed with the code of the most similar one, so the code substitutes for the input. The further analysis for input vector will be simplified.

In speech analysis with above method, what is analysized statistically is the code quantized of speech festure vector, not the speech vector itself. According to the code, the similarities and differences for some symbol in different environments are found out. With this method, investigation for phonetic symbols gets more convenient and sufficient.

EXPERIMENT AND RESULTS ANALYSIS

In our experiment, the similarities and differences of symbol [i] has been investigated in different environments, namely in combination with different consonants. The consonants followed by [i] are [j, q, x, z, c, s, zh, ch, sh, d, t, b, p, l, m, n, r, y].

Speech samples for research are from seven people, including five men and two women, and everyone has 20 samples for each of 18 monosyllable. After being quantized, the code sequence of stationary part of [i] following [j, q, x, d, t, b, p, l, m, n, y, z, c, s, zh, ch, sh, r] are obtained, as shown, for examples, in Tab. 1 which belongs to two people respectively.

According to the tables, the codes of [i] following [j, q, x, d, t, b, p, l, n, m, y] are more similar to each other, and that following [z, c, s] and [zh, ch, sh, r] come to the same thing.

So we can divide Chinese symbol [i] for teaching into three minimum: [i] following [j, q, x, d, b, t, p, j, m, n, y] is expressed as [i], and that following [z, c, s] and [zh, ch, sh, r] as [ɿ] and [ʅ] respectively.

CONCLUSION

According to our experimental investigation, the results show that representing by International Phonetic Symbol [i, ɿ, ʅ] is more proximal and representing by Chinese symbols used now for teaching is away from phonetic practice in principle, except depending upon environment. What we have done is as a part of investigation of phoneme representation for research and it is being continued.

Also, the experimentation shows the VQ technique is an effective tool for phonetic research, it can show similarity between phonemes evidently.

Tab. 1

Speaker: Mr. Zhao

Frame	11	12	13	14	15	16	17	18	19	20	21
Ji	51	52	52	63	63	62	63	52	56	56	64
Qi	59	60	60	60	60	60	60	60	57	57	57
Xi	50	50	51	51	51	51	50	63	64	64	64
Zi	21	17	17	17	19	19	20	20	20	20	20
Ci	20	20	20	20	20	20	20	20	20	20	20
Si	18	18	18	17	17	17	17	17	17	19	19
ZHi	1	1	4	4	1	1	1	1	1	2	3
CHi	27	13	4	1	3	3	3	3	1	1	1
SHi	12	11	11	11	11	11	11	11	11	11	11
Di	50	64	64	62	64	64	64	62	62	62	62
Ti	63	63	63	63	64	64	64	55	55	55	60
Bi	63	52	62	63	52	64	64	64	64	52	62
Pi	63	63	63	63	64	63	52	64	64	52	54
Li	58	62	58	63	62	62	64	62	64	58	64
Mi	56	56	53	53	53	53	53	53	53	53	53
Ni	55	56	56	54	56	54	54	54	54	54	51
Ri	3	3	3	3	3	3	3	3	3	3	3
Yi	62	62	62	62	64	56	56	56	56	54	54

Speaker: Mrs. Luo

Ji	56	56	56	56	56	56	56	56	56	56	56
Qi	41	44	42	42	42	42	56	56	54	54	56
Xi	42	42	47	47	47	47	54	47	54	45	54
Zi	3	3	4	4	4	4	4	4	4	4	4
Ci	14	10	3	3	3	3	3	3	3	3	3
Si	10	9	9	3	3	3	3	3	3	3	3
ZHi	23	23	23	23	23	23	23	23	23	23	32
CHi	41	44	17	22	27	22	27	22	22	22	22
SHi	43	15	15	15	15	15	22	21	27	27	27
Di	56	56	56	56	50	50	50	42	42	42	42
Ti	46	46	60	60	45	45	46	48	46		
Bi	55	55	55	55	55	57	57	50	50	42	42
Pi	45	45	45	45	45	45	58	58	58	58	58
Li	56	56	56	56	56	54	54	55	55	56	50
Mi	63	63	63	63	63	46	58	58	46	46	58
Ni	48	48	63	63	64	64	64	64	64	64	64
Ri	30	30	30	29	30	30	32	32	32	32	32
Yi	50	50	50	55	56	57	57	59	58	57	59

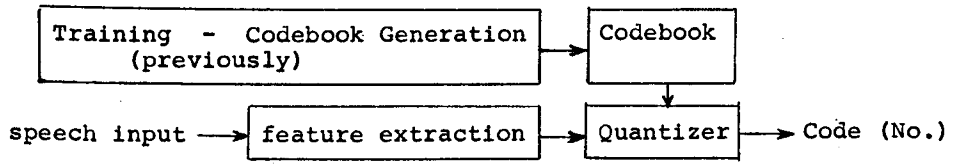


Fig. 1. Vector Quantization Process

REFERENCES

[1] Xu Shirong: On Putonghua Phonetics. 1980, Beijing.

[2] Masaki KOHDA et "Spoken Digit Mechanical Recognition System" Trans. of IEC, vol. 55-D No. 3 (Japanese).

[3] Anderes Buzo, Augustine H. Gray, Robert M. Gray and John D. Markel: Speech Coding Based Upon Vector Quantization. IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-28, No. 5, October 1980, pp. 562-574.

A SPECTRAL WARPING MODEL
A study of French Nasal vowels

P.F. MARTEAU, J. CAELEN & M.T. JANOT-GIORGETTI

Laboratoire de la Communication Parlée
I.C.P. Unité associée au CNRS (UA 368)
INPG/ENSERG - 46 avenue Félix Viallet - 38031 GRENOBLE CEDEX FRANCE

ABSTRACT

Most of the speech signal analysis methods deal with the spectral properties of homogeneous segments, assuming the properties are invariant on these segments. These methods do not take into account the dynamic aspects and the spectral warping between two contiguous segments. We propose a descriptive model which focuses on what does in fact vary in a set of successive spectra. This model is based on a time-domain statistic analysis of the frequency band energies of the spectra. We have experimented this approach by analysing the behavior of some French nasal vowels in a continuous speech corpus. We think that this sort of model is able to follow temporal transitions of different masses of energy. We show that these masses and the phases of their motion could be used as acoustic correlates.

INTRODUCTION

The speech signal is a time-domain varying signal. Whatever the kind of processing used, we always come back to a set of parameters which vary in time. (This set of parameters at the instant t represents the signal on a relatively short segment centered on t). If we suppose that the signal is represented by a m-parameters spectral vector (S), the evolution of this signal will be materialized by the motion of the point P representing the vector extremity ; with time, P follows a trajectory which reflects articulatory gestures in the R space. We can then notice on this trajectory relevant properties such as accumulation or turn-back points, rectilinear or curve segments (CAELEN 1986). If an accumulation point (a reached target) expresses a steady state easy to analyse, it is not the same for turn-back points (not reached targets) or transitions.

Some questions then arise : can we extract some laws relative to the spectral warpings in the vicinity of these transcient segments ? If these laws exist, can we interprete them as the trace in the acoustic space of the movement of the articulators.

From a perceptual level too, the role of the spectral change seems to be of great importance, even for the perception of vowels considered usually as monophtongs vowels. The modeling of these spectral changes are mainly based on formant transitions (LINDBLOM & STUDDERT-KENEDY 1967, GAY 1968, NEAREY & ASSMAN 1986, BROAD & CLERMONT 1987). The model that we suggest in order toanalyse acousticalevents is a mid term model of spectral warpings (its validity being reduced to a length of time about 100 ms).

We shall see that it is possible to extract information relative to the displacement of masses of energy which characterize the temporal signal variations. Finally, we shall evaluate this kind of model within the framework of a study on french nasal vowels in a continuous speech corpus.

1. THE MODEL

The acoustic signal is transposed in the frequency space by computingevery (t=5ms) a LPC spectrum (14 coefficients), defined on (m = 22) frequency bands, one Bark wide. The integration to one Bark brings us closer to the auditive transformation observed in man. Taking a temporal window W, correponding to the instants (1,...,q), we may note $J=\{1..q\}$ and $I=\{1..m\}$. On this window, the speech signal is thus represented by a rectangular array of real numbers: $\{S_i^j \mid i\in I;\ j\in J\}$.

We say that (S_i^j) is the energy of the i-th band for the j-th sample.

we may note $E=R^I$ the vectorial space of the real functions (f) defined by : $E=\{f \mid f^i\in R; i\in I\}$ and $G=R^J$ the vectorial space of the real functions (g) defined by : $G=\{g \mid g^j\in R; j\in J\}$. There are then two different ways to analyse the array (S_i^j) wether we consider the cloud C(I) of the functions S^i in the space G or the cloud C(J) of the functions S^j in the space E.

C(J) can be interpreted as a time-domain trajectory in the space E and C(I) as a frequency-domain trajectory in the space G. We find again the time/frequency duality. The two clouds can be then analyzed in terms of inertia, in order to extract the main tendency of the temporal or frequency evolutions. This is the purpose of the **factorial analysis (BENZECRI 1973).** This type of analysis applied to speech processing is not new (CARTIER & GRAILLOT 1974, CAELEN & VIGOUROUX 1983). We insist on the fact that here the objects are not taken from a set of independant observations : the objects are linked by a time relation. The theory of the factorial analysis tells us that the two previous points of view are equivalent (or dual). This means that the inertia characteristics of the two clouds are linked by a bijective relationship. **This is why, superposing the two spaces E and G, we may interpret simultaneously the transitions of the signal as the dislocation of some masses of energy in the time domain and in the frequency domain.**

We consider then the cloud C(J) in the space E. We may note $\{\nu_j \mid j\in J\}$ a set of positive numbers : ν_j is the weights assigned to the j-th sample (spectrum). (In practice we use the Hamming window or the rectangular window). Let's note $\bar{\nu}$ the sum of the weights. If we fit the space E with the euclidian distance (d), (E,d) is a topologic space. The distance between two objects j and j' is given by :

$d(j,j')^2 = \Sigma\{(S_i^j - S_i^{j'});\ i \in I\}$

We call G the gravity center of the functions S :

$G_i = \Sigma\{(\nu_j/\bar{\nu})S_i^j;\ |\ j \in J\}$ and $\{\sigma_{ii'} | i \in I;\ i' \in I\}$ the symetric square array representing the quadratic inertia function of the cloud C(J).

$\sigma_{ii'} = \Sigma\{\nu_j S_i^j S_{i'}^j\ |\ j \in J\} - G_i G_{i'} \bar{\nu}$.

$U = \{u^i\ |\ i \in I\}$, the set of the orthonormal eigen vectors for the application σ ; $\sigma(u)^i = \lambda u^i$

$V = \{v^j\ |\ j \in J;$ u U . G , the corresponding main components $v^j = \Sigma\{u^i(S_i^j - G_i)\ |\ i \in I\}$

Then, if we consider the sub-system U(p) of the p-first eigen vectors, assigned to the p-highest eigen values , we obtain a p-order estimate of the array S which is given by :

$$\hat{S}_i^j = G_i + \Sigma\{(\sqrt{\lambda(h)}\,v^j(h)\,u_i(h)\ );\ \ h=1..p\}\ \ (1).$$

S is the p-order array which minimize the criterion

$$\Sigma\{(\hat{S}_i^j - S_i^j)^2\ |\ i \in I;\ j \in J\}$$

What is the meaning of the model described by the equation (1) ? For the speech signal analyzed on a temporal window W having an order of magnitude of 100 ms, the model is reliable when n equals 2 or 3. This means that on this analysis window, the trajectory of the representative spectral evolution's point is approximatively contained in a 2 or 3 dimensional sub-space. The (u(h)) forms can be interpreted as frequency mask functions, balanced by the $(\sqrt{\lambda(h)}v(h))$ coefficients. Thus, we can compare them to the spectral cues defined by ROSSI or CAELEN (ROSSI & al 1983, CAELEN & CAELEN-HAUMONT 1981). Such a model is entirely related to the choice of the analysis window. Nevertheless, we can test its reliability to small time translations, i.e. its predictive power. Let the temporal analysis window be W = $\{1..q\}$ and $\{u(h), v(h), \lambda(h);\ \ h=1..p.\}$ the p-order estimate model. Let the spectrum S^{q+1} calculated at instant (q+1). Its projection in the sub-space E gives for this spectrum the values of the function: $\sqrt{\lambda(h)}\,v^{q+1}(h)$ The euclidian distance between the spectrum and its reconstruction through the model is a rupture criterion, which is expressed by

$$\delta^{q+1} = \Sigma\{(S_i^{q+1} - \hat{S}_i^{q+1})^2\ |\ i \in I\} \qquad (2)$$

assuming that the window W includes a part of a steady state, δ^{q+1} will stay small and the same will happen with δ^{q+i} and i ⩾ 1. The model will be reliable with regard to a small future insight. Nevertheless, for a larger temporal insight, we expect the criterion quickly to deteriorate.

This means that the model is not reliable enough, that the past is not sufficient to explain the future. Hence, the analysis window parameters : position and length of time, are of great importance when considering the convergence of the model and its power for describing spectral warpings with coherence. The choice of the appropriate window is a complex problem. In our opinion, this choice may be conditioned by the search for instabilities, since in the way it is built the model focuses on these instabilities.

the rupture criterion exposed before (relation N°2) can produce an automatic segmentation of the speech signal.

The results of such a segmentation will be presented later.

exposed before (relation N° 2) can then produce an automatic seg- mentation of the speech signal. The results of such a segmentation will be presented later. Let us say simply that a signal fragment may be split up into smaller fragments by finer and finer downward analysis. This is performed by eliminating the successive instabilities detected on topologic criteria. We thus realize a tree of sub-models representing the speech signal on a given time interval. The leaves of this arborescence are close to homogeneous phones (CAELEN 1981) and describe acoustical events.

Following this, we try to show that in a suitable window, i.e. a window with few instabilities, the basis $\{u(h)\ |\ h=1..p\}$ and the evolution of their associated components $\{(\sqrt{\lambda}v(h)\ |\ h=1..p\}$ express energy displacement interpreted as motion towards energy targets. We will notice that these masses of energy have not necessary relationship with formants. We have chosen for this study the French nasal vowels (/ã/, /ɔ̃/, /ɛ̃/, /œ̃/) because of their complex structures which have given rise to many studies. We will then try to classify different observation cases, in order to extract laws and descriptive rules for the phonetic feature of nasality /+ NASAL/.

2. THE NASAL VOWELS

From an articulatory level, the nasal vowels are the result of an acoustic coupling between oral and nasal cavities. The coupling occurs approximatively at the middle of the vocal cavity, between the lips and the glottis. The consequence of this coupling is the displacement of the first natural oral formant and the appearance of a pole-zero pair in the transfer function (FUJIMURA & LINDQUIST 1971, MRAYATI 1976, MAEDA 1982, HAWKINS & STEVENS 1985...). According to the authors, the resulting acoustical correlates generally affect the low frequency part of the spectrum ; these correlates are in fact the frequency displacement of the first formant and the widening of the bandwith. A formant with high damping also appears at the viciniy of 300 Hz. Some studies try to introduce more accurate dynamic information.

MERMELSTEIN set up four acoustical parameters which are relative variations of the energy in the frequency bands [0-1], [1-2], [2-5 kHz] and the centroïd of the frequency band [0 - 500 Hz].

FENG proposed the nasopharyngal target concept which is characterized by the appearance of two formants : one is located near 300 Hz, the other located near 1000 Hz. This concept arises from articulatory simulation (FENG & al 1985).

CHENG showed from psychoacoustical data the possibility of a balance between two masses of energy centered in the vicinity of 300 and 1000 Hz (DELATTRE 1969, CHENG 1987).

Hence, nasality should be characterized by spectral deformations from an initial spectrum corresponding to an oral or quasi-oral vowel, towards a final spectrum with greater energy in the neighbourhood of 300 and 1000 Hz. Since we observed a sort of analogy between the articulatory and perceptive concepts, can we encounter similar properties in the acoustical space ? We shall see that the spectral warping model brings some significant elements.

3. THE STUDY

3.1. corpus

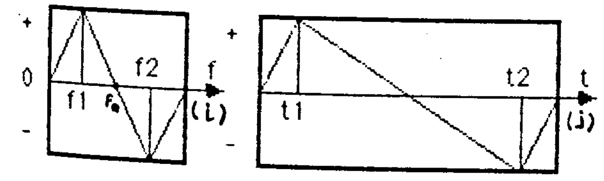
The corpus includes a 45 words text, the average time duration of which is 29 s. The text is an abstract of a paper published in **"Science et vie"**. The records of ten speakers (5 women and 5 men) are used. These speakers were requested to speak naturally and comprehensibly. The records were made in a quiet room. A N 4420 Radiola recorder was used.

"D'éminents biologistes et d'éminents zoologistes américains ont créé pour des vers géants, un nouveau phyllum dans l'actuelle classification des nombreuses espèces vivantes. Ces longs vers prospèrent sur le plancher marin des zones sous-marines profondes. Des sources thermales chaudes y maintiennent une température moyenne élevée."

The corpus was manually labelled by an expert in phonetics, according to labelling principles based upon a spectral analysis (VIGOUROUX & CAELEN 1985). Selection of the nasal vowels is then performed automatically from a label file. The corpus includes 160 nasal vowels (7/ã/ 5/ɔ̃/, 4/ɛ̃/) for ten different speakers. In this study, we do not take into account the differences between and

3.2. Interpretation procedure

The interpretation is carried out through qualitative analysis of the pair $(u(h), \sqrt{\lambda(h)}v(h)\ |\ h=1..p)$. For instance let the pair $(u, \sqrt{\lambda}v)$ be expressed as follow :



from t_1 to t_2, energy decreases in the vicinity of frequency f1 and increases in the vicinity of frequency f2. Energy in the vicinity of fa is constant for this transition. The extreme values of the $(\sqrt{\lambda(h)}v(h))$ functions (for t1 and t2) can be interpreted as targets.

- t1 is the target characterized by spectra the energy of which is maximal in the vicinity of f1 and minimal in the vicinity of f2.
- on the contrary, t2 is the target featured by spectra the energy of which is minimal in the vicinity of f1 and maximal in the vicinity of f2.

Thus, analysis of the segment /a/ (Fig. 1.1 and 1.2) shows that two targets exist :

- The first one corresponds to the energy masses centered around the following frequencies (450-540 Hz/1550-1790 Hz). These two masses are connected to the two first formants of the /a/. The stability of this target implies an accumulation point on the trajectory.
- The second one corresponds to the energy masses centered around the following frequencies (450-540 Hz/2060-2350 Hz) and deals with the increase of energy in the vicinity of the first formant and a dislocation of the second formant towards higher frequencies (influence of /ki/ context). This second target, not reached, creates a turn back point on the trajectory.

The analysis of segment /ã/ (fig. 2.1 and 2.2.) shows as well two instable targets.

- One near the frequencies (540-650 Hz, 1130-1340 Hz) corresponds to an oral vowel.
- The other near the frequencies (260-320 Hz, 880-1000 Hz) corresponds to a nasal vowel:

These two targets are represented by two turn-back points on the trajectory on either side of u axis. The trajectory is nearly rectilinear between the two targets.

3.3. Results

For the whole corpus results are plotted on figures similar to figures (1.2) and (2.2).

The interpretation is manual.

- We may note by (CN) the presence of a target characterized by a simultaneous increase of energy in the frequency bands [200-500 Hz] and [760-1340 Hz]
- We may note by (CH) the presence of a target characterized by an increase of energy in the frequency band [760-1340].
- We may note by (CB) the target characterized by an increase of energy in the frequency band 200-500 Hz
- We may note by (ER) the presence of an error, when no interpretation is possible.

The results are reported in fig. 3 at the end of this paper.

On the whole nasal vowels of the corpus, all context taking into account, the following occurences appear : CN = 65 %, CH = 29,5 %, CB = 0 %, ERR = 5,5 %.

Based on this data, the energy increase in the band [760-1340 Hz] appears to be a significant acoustical correlate since it is present in 94,5 % of the cases. In 35 % of the cases, the model is not sensitive to any significant variation in the band [200-450 Hz]. This means that the low frequency formant variation is not always necessary to perform a French nasal vowel. Analysis-synthesis techniques, using a formant synthesizer, leads to similar conclusions for nasal vowels of French spoken in Montreal (LAFFERIERE & O'SHAUGHNESSY 1986).

Nevertheless it is interesting to observe that results obtained from a dynamic articulatory model (FENG 1986) and pyschoacoustical results featuring the sensitivity of controlled parameters of synthesizer with respect to the perception associated to nasality (FENG 1987), are partially consistent to the results given by the model of spectral deformations : the existence of two dynamic energy masses (64% of the cases). The first one near 300 Hz, the other one, dominant, near 1000 Hz. The target concept is represented by the evolution of the $(\sqrt{\lambda}v)$ function and the notion of enèrgetic balance could be interpreted by the structures of the (u) forms.

CONCLUSION

We have proposed a model which takes into account local transitions of energy masses. These masses are computed. Thus we do not have to face the problem of detection. These energy masses are usuallyassociated with formants when these are dynamic either by frequency translations or by widening the bandwidth.

But they also take into account phenomena such as spectral flattening. This spectral warping model has allowed us to relocate roughly the results of articulatory and psychoacoustic analysis within the framework of the study of the French nasal vowels.

We think that the procedure of interpretation may be automated by the use of expert system. The experience and the knowledge acquired from spectogram reading systems will be extremely usefull. Finally the use of such a model is not limited to the field of acoustic-phonetic decoding. In our opinion it could also be used in low rate coding.

BIBLIOGRAPHY

BENZECRI J.P. (1973)
L'analyse des données
Dunod (1973).
BROAD D.J., CLERMONT F. (1987)
A methodology for modeling vowel formant contours in CVC context
J. Acoust. Soc. Am., 81 (1), 155-665.
CAELEN J., CAELEN-HAUMONT G. (1981)
Indices et propriétés dans le projet ARIAL II
Proceedings GALF-CNRS "Processus d'encodage et de décodage phonétique".
C.Abry, J.Caelen, J.S.Lienard, G.Perennou & M.Rossi.
CAELEN G., VIGOUROUX N. (1983)
Les indices de distribution spectrale. Etude comparative au travers de 2 analyses discriminantes monolocuteur et interlocuteur
Speech Communication, 2, 133-136.
CAELEN J., VIGOUROUX N., (1985)
Une base acoustique et phonétique hiérarchisée : des faits aux connaissances
Actes du Symposium Franco-Suédois sur la parole, GRENOBLE.
CHENG Y.M., GUERIN B. (1987)
Nasal vowel study : formant structure, perceptual evaluation and neural representation in a model of the peripheral auditory system
Institut de la Communication Parlée, Bulletin n° 0.
DELATTRE P. (1969)
The General Phonetic Charactéristic : Final report US Dept. of Health, Education & Welfare. Office of Education Institute of International Studie.
FENG G., ABRY C., GUERIN B. (1985)
How to cope with nasal vowels ? Some acoustic boundary poles
Actes du Symposium Franco-Suédois sur la parole, GRENOBLE.
FUJIMURA O., LINDQUIST J. (1971)
Sweep-tone measurement of vocal-tract characteristics
J. Acoust. Soc. Am., 49 (2), 541-558.
GAY T. (1968)
Effects of speaking rate on diphthong formant movements
J. Acoust. Soc. Am., 44, 1570-1573.
GAY T. (1978)
Effects of speaking rate on vowel formant movements
J. Acoust. Soc. Am., 63, 223-230.
HAWKINS S., STEVENS K.N. (1985)
Acoustic and perceptual correlates of Non-Nasal/Nasal Distinction for vowels
J. Acoust. Soc. Am., 77, 1560-1575.
LAFERRIERE F., O'SHAUGHNESSY D. (1986)
Analyse-synthèse et études de règles acoustiques de production avec un synthétiseur à formant
15ème J.E.P., 11-14, AIX EN PROVENCE.
LINDBLOM B., STUDDERT-KENNEDY M. (1967)
On the role of formant transitions in vowel recognition
J. Acoust. Soc. Am., 42, 830-843.
MAEDA S. (1984)
Une paire de PICS spectraux comme corrélat acoustique de la nasalisation des voyelles
13ème J.E.P., 223-224, BRUXELLES.
MERMELSTEIN P. (1977)
On detecting nasals in continuous speech
J. Acoust. Soc. Am., 16 (2), 581-587.
MRAYATI M. (1976)
Contribution aux études sur la production de lao parole
Thèse Doct. d'Etat, I.N.P. GRENOBLE.
ROSSI M., NISHINUMA Y., MERCIER G. (1983)
Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole
Speech Communication, 2, 215-217.

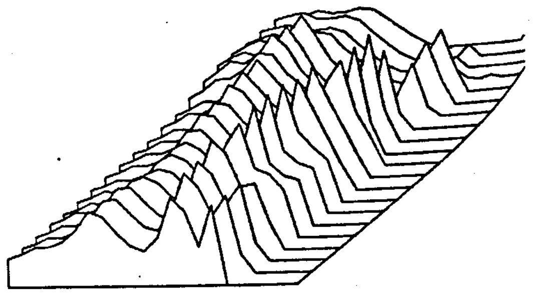


FIG N° 1.1 : The set of successive spectra S_i^t on the /a/ segment.

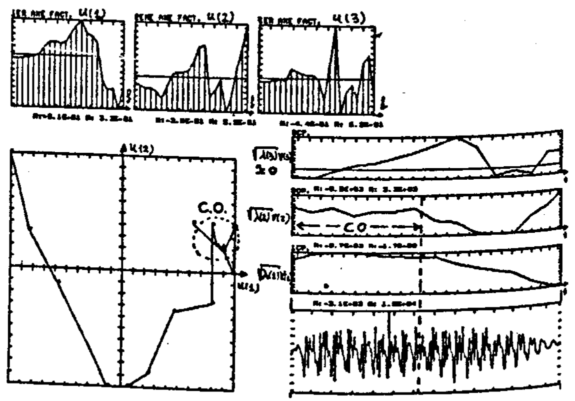


FIG N° 1.2 : The model on the /a/ segment; The eigen vectors (u) and the associated main components (Qv)
The trajectory in the space (u(1),u(2))

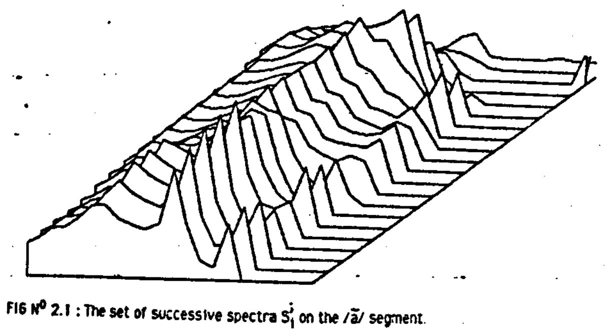


FIG N° 2.1 : The set of successive spectra S_i^t on the /ã/ segment.

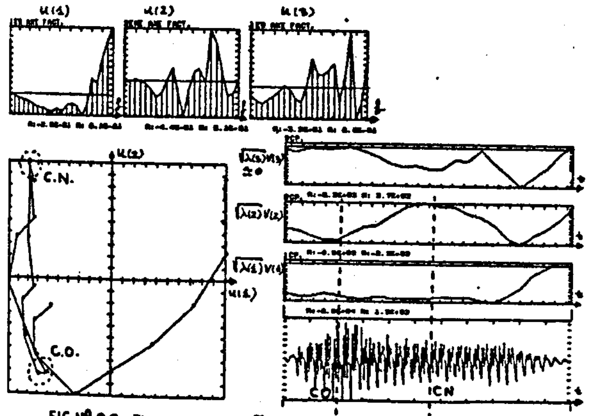


FIG N° 2.2 : The model on the /ã/ segment, The eigen vectors (u) and the associated main components (Qv)
The trajectory in the space (u(1),u(2))

THE RESULTS

Vowel	Context	CN %	CH %	CB %
/ã/	/nãb/	70	30	0
	/nãz/	50	50	0
	/eãm/	90	10	0
	/dãl/	60	30	0
	/vãt/	60	30	0
	/nãS/	80	10	0
	/tãp/	80	10	0
/õ/	/kõk/	30	70	0
	/jõd/	60	30	0.
	/nõb/	80	10	0
	/nõv/	60	40	0
	/tõd/	80	20	0
/ɛ̃/	/kɛ̃ /	80	20	0
	/rɛ̃g/	60	30	0
	/ ɛ̃n/	60	30	0
	/mɛ̃t/	50	40	0

FIG. No 3

DETECTION AND IDENTIFICATION OF PLOSIVE SOUNDS IN WORDS

MASUZO YANAGIDA　　YOUICHI YAMASHITA　　OSAMU KAKUSHO

The Institute of Scientific and Industrial Research
Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan

ABSTRACT

A system for detecting and identifying plosive sounds in Japanese words are presented. The fundamental parameter employed here to detect plosive sounds is the cepstral distance between the analysis results of an odd pair of frames with their starting positions coincided. Introduced in this report for elimination of removable candidates are short time power, pole positions obtained by low-order analyses, and a measure for representing relative disposition in the discrimination space. Phoneme identification or mutual discrimination among the detected candidates for plosive sounds is carried out by using a following-vowel dependent discrimination algorithm formerly developed by the authors.

INTRODUCTION

For continuous speech recognition of large vocabulary for unspecified speakers, discrimination by phoneme or phoneme group is required, but no reliable automatic detection/identification method has been developed for most of the phonemes, particularly for plosive sounds, whose intra-group discrimination is most difficult among phoneme groups.

For discrimination of Japanese voiceless plosives Kitazawa and Doshita[1] proposes a method using spectral information at the burst. While Tominaga et al[2] proposes discrimination of voiced plosives using transition properties form the burst to the following vowel, since Japanese plosive sounds are always followed by one of the Japanese five vowels. Ide et al[3] proposes a discrimination of Japanese voiceless plosives introducing time-spectrum pattern.

The authors[4,5] have proposed a discrimination method for CV syllables uttered isolatedly with plosive sounds as C followed by one of the Japanese vowels as V employing both instantaneous and dynamic properties of acoustic parameters at the burst and during the transition parts, respectively, where LPC cepstral coefficients and short time power are used as the acoustic parameters and their regression lines are introduced to represent their dynamics. All these experiments were conducted under condition that the given speech sample contains one of the plosive sounds.

This paper proposes an approach to detect plosive sounds in words. The proposed method employs LPC cepstral distance as the primary parameter to find out possible candidates for plosives. Also used as the auxiliary parameters for eliminating other phonemes from the candidate list are short time power, the pole positions obtained by low-order analyses, and a measure for representing relative disposition in the discrimination space.

This paper mainly discusses the process to eliminate excess errors of preliminary detection procedure. Phoneme identification or mutual discrimination among the detected plosive sounds is carried out by a following-vowel dependent discrimination algorithm formerly developed by the authors.

DETECTION OF PLOSIVE SOUNDS

Fig.1 shows the flow chart of the process from speech input to phoneme identification. The whole process is divided into two parts: one is detection of plosive sounds, and the other is mutual discrimination among them. In the detection part, possible candidates are first searched based on cepstral distance between an odd pair of frames with their starting points coincided. Then removable candidates are discarded from the candidate list in the elimination procedure. The rest of the process is related to discrimination among plosives. Followings are the detail of the processing blocks.

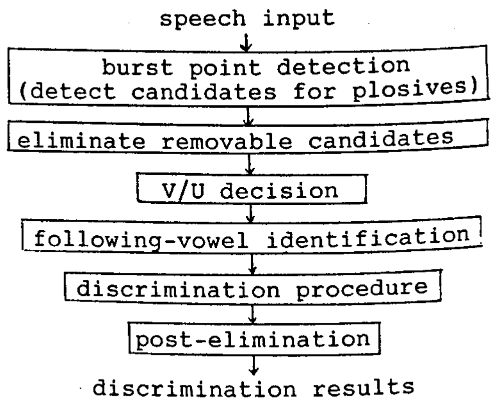


Fig.1 Flow chart of the process.

Burst Point Detection

Detection of plosive sounds can be realized by observing temporal changes of acoustic parameters. Temporal changes of acoustic parameters have been usually evaluated by comparing the difference between the analysis results of successive frames of the same frame length. However, the analysis scheme has problems of binary ambiguity and poor sensitivity because of its double-sided shifting gaps. In order to clear the these difficulties, an odd pair of frames with the starting points coincided as depicted in Fig.2 is introduced for detection of burst points or temporal change of spectral parameters. LPC cepstrum coefficient is employed here for the spectral parameter, and the distance d(t) between the analysis results of the odd pair of frames in the LPC cepstrum space is called <u>cepstral distance</u> where t denotes the center of the frame gap Lb.

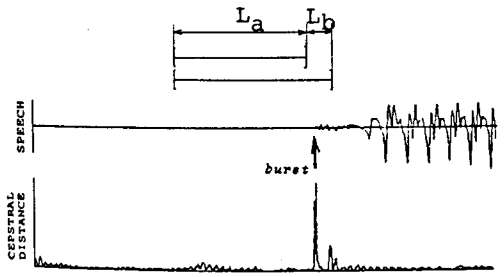


Fig.2 An odd pair of frames for detection of burst points.
Frame length: 20 ms for the longer frame
17 ms for the shorter frame
Shifting interval: 1 ms

Elimination of Removable Candidates

It seems impossible to detect plosive sounds only by the cepstral distance even though it is a good cue for burst point detection. Since the cepstral distance shows sharp peaks not only at burst points of plosive sounds but also at other occasions like beginning of fricatives, nasals and vowels. Auxiliary parameters introduced are short time power, pole position obtained by low-order LP analyses and a relative disposition of input sample in the discrimination space. The additional conditions to keep up as acceptable candidates are as follows:

a)Short time power P(t) grows up more than 40% of its local maximum P_m at the point t_p where the cepstral distance d(t) shows its local peak d_p.

$$P(t_p) / P_m > 0.4 \qquad (1)$$

where $d(t_p) = \max d(t)$

This is to eliminate the spectral transition part which shows a phenomenal local peak on the cepstral distance although it is not a plosive sound.

b)The average build-up rate r of ~~the~~ short time power P(t) is greater than a threshold r_o if t_p is at word-initial.

$$r = \frac{P_m - P_p}{t_m - t_p} > r_o \qquad (2)$$

This is to eliminate some phonemes which show slow power build-up in word-initial position like vowels.

c)LP analysis of order 2 yields no stable pole beyond 4kHz extending over successive 6 frames out of 28 frames shifted every 5ms around t_p.

This is to eliminate /s/ from the candidate list.

d)In case of voiced sounds, LP analysis of order 4 gives stable pole in the frequency range 80-300 Hz with narrower band-width than 200 Hz.

This is to eliminate /r/ from the candidate list.

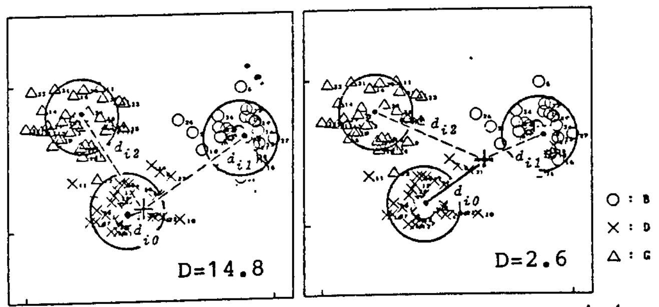
e)In word initial position, the following measure D, representing inclination to a particular plosive, is greater than a certain threshold D_o.

$$D = \sum_{i \neq i_o}^{N} \frac{d_i}{d_{i_o}} > D_o \qquad (3)$$

where di : distance from input sample to class i.

$$d_{i_o} = \min_i d_i$$

This is to eliminate phonemes of non-plosive disposition from the candidate list for plosive sounds. Fig.3 shows examples of relative disposition of the input samples /d/ and /r/ in the Fisher space with corresponding D values.



(a) input sample /d/　(b) input sample /r/
Fig.3 Relative disposition of /d/ and /r/ in the Fisher space.

This condition is applied after the discrimination procedure as post-elimination only to candidates in word-initial position.

The cepstral distance measure d(t) does not show any remarkable peak for sound /g/ in words since the temporal change of spectrum envelope is rather slow for the phoneme though it belongs to the plosive group. However, as /tʃi/ and /tsu/ yield evident peaks on the cepstral distance like plosives, the present report includes them in the set of phonemes to be detected and identified as /ti/ and /tu/, respectively.

DISCRIMINATION OF PLOSIVE SOUNDS

Fig.4 shows the discrimination process for the detected sounds.

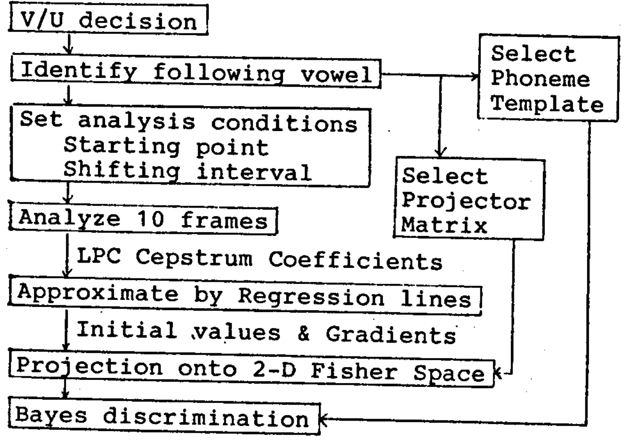


Fig.4 Discrimination among plosive sounds.

Voiced/Unvoiced Decision

V/U decision is made based on existence of buzz bar before the burst and the autocorrelation pattern of the prediction residual signal.

Identification of Following Vowels

Plosive sounds in Japanese are always followed by one of five vowels and the phoneme templates to discriminate plosives are prepared for each following vowel in the present system. So, identification of the following vowel should be performed before discrimination of the leading plosives. For identification of the following vowel two-stage decision by majority is employed. The first decision by majority is made among the 5-nearest neighbors in the discrimination space or a two-dimensional Fisher space derived from the LPC cepstrum coefficients, and the second decision by majority is made on the five decision results obtained from successive five frames shifted by 1 ms each, with the center frame at 70 ms after the burst.

Discrimination of Plosive Sounds

Both dynamic and instantaneous properties of acoustic parameters are employed for discrimination among plosives. Gradients of the regression lines representing temporal change of short time power and LPC cepstrum coefficients during the transition period from the burst to the following vowel are employed as the dynamic parameters. The number of frames to be fitted by regression lines is fixed to be 10. The first frame is located at the position fr1 starting at Td ms after the burst and the succeeding frames fr2, fr3, shifting interval Ts as depicted in Fig.5. The delay time Td representing the relative position of the first frame from the burst point and the shifting interval Ts are set optimal for each following vowel to give the best discrimination score for isolated CV utterances by preliminary experiments. The interpolated values on these regression lines at the specified positions are used as the instantaneous parameters.

The acoustic parameters here are C0, the frame power and C1, C2, ..., C12, the LPC cepstrum coefficients of order 12, and the parameters for discrimination among plosives are their gradients and the interpolated values for the first frame position. So the total number of parameters for discrimination is (1+12)x2=26.

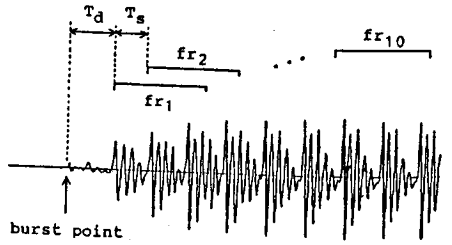
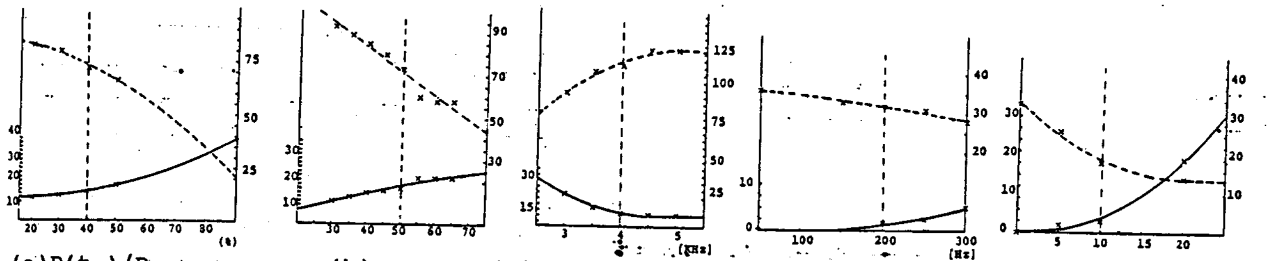


Fig.5 Set up of succeeding analysis frames.

Discrimination is performed in the two-dimensional Fisher space obtained from the 26-dimensional parameter space described above after V/U decision and following vowel identification. The phoneme templates are prepared for each following vowel and for voiced and unvoiced case respectively, that means there are 10 sets of phoneme templates containing three standard phonemes each. The resultant vector y in the Fisher space is obtained from the original parameter vector x as follows by the projector matrix W that maximizes the Fisher ratio J(W).

$$y = W^t x \qquad (4)$$

$$J(W) = \frac{|W^t S_b W|}{|W^t S_w W|} \longrightarrow \max \qquad (5)$$



(a)P(t_p)/P_m (b) r (c)Pole by LP(2) (d)Pole by LP(4) (e) D

Fig.6 Performance of the parameters to eliminate removable candidates from the list.
solid lines : missing error
dotted lines : excess error
vertical broken lines : threshold value

where t denotes transpose
S_w: within-class covariance matrix
S_b: between-class covariance matrix

Discrimination here assumes the normal distribution of each phoneme in the Fisher space and decision is made according to Bayes rule.

EXPERIMENTAL RESULTS

Speech Data

Phoneme templates in the Fisher space is obtained from isolated CV syllables uttered by 38 male adults. Test samples are 30 Japanese city names uttered by other 5 male adults. The number of phonemes to be detected and identified is 160 out of 535. Frequency of occurrence for each phoneme is not well balanced because of special feature of city names.

Results

Performance of the proposed method on above-mentioned test samples is shown in Table 1. It shows the statistics of missing errors and excess errors by adding conditions one by one on the test data.

Table-1 Detection and discrimination results of plosives in words. The total number of plosives to be detected = 160.
(0) LPC cepstrum distance d(t) > d_o
(a) Normalized power at the burst $P(t_p)/P_m > 0.4$
(b) Power build-up rate $r > r_o$ in word-initial
(c) Pole position in high frequency range by LP(2)
(d) Pole position in low frequency range by LP(4)
(e) Relative disposition $D > D_o$ in word-initial

Condition	Dis/Det/Tot	Missing	Excess
(0)	120/153/160	7	121
+(a)	116/147/160	13	71
+(b)	111/140/160	20	41
+(c)	111/140/160	20	36
+(d)	111/140/160	20	33
+(e)	109/137/160	23	18

Table-1 shows that the final detection rate is 86% (137 out of 160) with 15% excess-detection error, and 85%(109 out of 137) of the detected plosives are correctly discriminated.

Discussions about elimination conditions

Performance of the five parameters introduced in each elimination condition is shown in Fig.6 with a vertical broken line in each diagram representing the threshold value set in the present system.

CONCLUSION

Automatic detection and discrimination of plosive sounds are presented. LPC cepstral distance between an odd pair of frames is introduced as the primary cue for detection of plosives. Some other additional conditions are discussed to eliminate excess candidates for plosives.

Acknowledgment
This work is due to devotional help by Messrs.Mitsuhiro Tsunoda and Yukihiro Okada of Kansai University.

References

[1]S.Kitazawa et al.:J. Acoust. Soc. Jpn., 40, 5, 332-339(1984).
[2]M.Tominaga et al.: Trans. Committee on Speech Res., Ac.Soc. Jpn., S81-72(1982).
[3]K.Ide et al.: JAS Jpn, 39, 5, 321-329 (1983).
[4]Y.Yamashita et al. : Trans. IECE, Jpn., J69-A, 2, 282-290(1986).
[5]Y.Yamashita et al. : Trans. IECE Jpn., J70-A, 1, 132-134(1987).
[6]M.Tsunoda et al. : Autumn Meeting of Acoust. Soc. Jpn., 1-3-6(1986).
[7]Y.Takahashi et al. : Autumn Meeting of Acoust. Soc. Jpn., 2-1-5(1984).
[8]R.O.Duda and P.E.Hart:"Pattern classification and scene analysis", John Wiley, New York, pp.114-121(1973).

LOCATION AND RECOGNITION OF PLOSIVE CONSONANTS IN CONTINUOUS SPEECH

MARY O'KANE

School of Information Sciences and Engineering
Canberra College of Advanced Education
Belconnen 2616, Canberra, Australia

ABSTRACT

Algorithms to locate and classify plosive consonants in continuous speech have been incorporated and tested in the FOPHO continuous speech recognition system [1]. These algorithms are based on a detailed study of the speech of ten Australian English speakers (five male, five female) who participated in a word game in which each speaker produced continuous speech versions of all possible VCV combinations where the vowels were either the high front vowel /i/ or the low back vowel /ɔ/ and where the consonant was one of the six plosive consonants of Australian English /p,b,t,d,k,g/. The study showed that a complete plosive classification algorithm must be context-dependent because it was found that generally all speakers produced the plosives in a heavily coarticulated manner with systematically varying coarticulation phenomena being observed in formant transitions, timing effects, bursts and pseudo-loci as one progresses from bilabial through alveolar to velar plosives. Another important factor that has to be taken into account in the recognition algorithm is the speaker's sex.

INTRODUCTION

An algorithm for the classification of plosive consonants in continuous speech was developed from a study of ten speakers, five male and five female, producing plosives in set contexts. This algorithm was tested and then generalised and incorporated into the FOPHO speech recognition system. In this paper the development of this algorithm is described and some results of using it in the recognition system are given.

FEATURES OF PLOSIVE CONSONANTS

When developing a recognition algorithm for the plosive consonants one has a large amount of literature of which to draw in order to determine what features of the plosive consonants are likely to be important for distinguishing between the plosives produced at different places of articulation. This literature can for convenience be grouped in three categories: perception experiments using synthetic speech, perception and production experiments using real speech, and classification studies. Synthetic speech experiments have highlighted the fundamental characteristics of plosives while the real speech experiments have tended to clarify the interaction of these characteristics and the variability of their realisations in real speech. Automatic recognition studies give some guide as to which features can be located automatically with reasonable efficiency and robustness. Here we give only a very cursory guide to these three types of literature.

In synthetic speech experiments in the early 1950's Cooper, Delattre, Liberman, Borst, and Gerstman [2] found that certain stop burst spectra were characteristic of the three stop types. In later experiments Liberman, Delattre and Cooper, [3], found that the formant transitions from the plosive consonant to the following vowel produced a successful synthetic voiced plosive. The first formant transition appeared to contribute to the voicing of the stop while the second formant transition provided a basis for distinguishing between the stop types. Further experiments led to development of the now famous 'locus theory' which postulates that the second formant transitions should point to a frequency locus no matter what the following vowel is. Delattre, Liberman, and Cooper [4] found that this was particularly characteristic of /d/, not quite so reliable for /b/, while for /k/ there were two loci, a high one if the following vowel was a front vowel and a low one if the following vowel was a back vowel. Hoffman [5] further refined much of the previous plosive research on stop consonants to see how the two main types of cues for plosive conconants, burst and formant transitions, interacted. He concluded that all the cues are perceptually independent of the other cues present and that for some stops, notably /b/ and /g/ the burst provided a weak cue and the transitions a strong cue while for /d/ the burst was a strong cue and the transitions were a weak cue. Later research has demonstrated the significance of these conclusions.

In real speech experiments, Halle, Hughes and Radley, [6] considered plosives occurring not only in conjunction with vowels but also in consonant clusters and at the beginnings and ends of words. They concluded that a complex array of cues was needed to characterise the plosives and that the locus theory was somewhat inadequate for this task. In this they were supported by Öhman [7] who studied spectra of VCV coarticulations for Swedish, using all possible combinations in which the consonant was a plosive. He deduced that each VCV coarticulation was a 'basic dipthongal gesture with an independent stop consonant gesture superimposed on its transitional portion'. The relative importance of burst and transitions in signalling plosive consonants has been heavily debated. However there is evidence (see [8] for example) that different speakers signal various plosives in different ways. Despite this Stevens and Blumstein [9] showed that stimuli as short as 10-20 msec sampled from the onset of consonant-vowel syllables can be reliably classified according to place of articulation using gross spectral shape wave 85% of the time. Studies such as that by Lisker and Abranson [10] have shown that timing phenomena such as VOT are also crucial to the correct production of plosives.

There have been several published descriptions of algorithms for the automatic discrimination of plosive consonants in various languages (e.g. [6], [11], [12], [13], [14], [15], [16], [17]). Considering this group of algorithms as a class, the most favoured aspect of the plosive for plosive discrimination is the burst which is often analysed according to some frequency-band scheme. Measurements of transitions and timing tend to be used as secondary recognition cues.

PLOSIVES IN CONTINUOUS SPEECH

Many of the published algorithms for the recognition of plosive consonants were developed from studies of plosives produced in citation syllables. From studying the literature described in the previous section we concluded that citation-form syllables were unlikely to be fully representative of the range of plosive production phenomena that speakers might use in continuous speech. Therefore, in order to develop an algorithm for the recognition of plosive consonants in Australian English continuous speech we designed an experiment in which both male and female speakers produced all the plosives occurring in English in a range of VCV coarticulatory settings with the added complication of junctural effects occurring within the VCV triplet.

This experiment was conducted as follows:
Lists of two-word sequences were prepared. Each of these two-word sequences was one of two forms:

(1) The first word ended in a VC combination and the second word began with a V; where the vowels could be either the high front vowel /i/ or the low back vowel /ɔ/ and the consonant was one of the six plosives e.g. 'heat ought', 'morgue awful';

(2) The first word ended with V and the second word began with a CV combination where the vowels could be either /i/ or /ɔ/ and the consonant was one of the six plosives e.g. 'he taught', 'more gory'.

Thus with the two vowels and six plosive consonants and two juncture positions there were a total of forty-eight two-word combinations. Five male and five female speakers who all spoke standard educated Australian were each presented with the list of two-word sequences and instructed not to study the list but to immediately begin saying sentences containing the word sequences. It was impressed on the subjects that the sentences they produced were to be spoken at a conversational speed and that the semantic content of the sentences was of no particular importance. This was to keep the subject speaking at as conversational a rate as possible. This aim was largely achieved. With this experimental paradigm the phonetic and junctural contexts were controlled but the speech used was reasonably representative of continuous speech.

The sentences containing the two-word sequences were recorded and then the required VCV tokens were excised and digitised (with a 10 kHz sampling rate for male voices and 16 kHz for female voices). These tokens were then analysed using an autocorrelation-based linear prediction technique. Timing, formant transition and burst phenomena were all investigated. The results of these investigations are briefly described below.

TIMING PHENOMENA

The particular timing parameter measured, chosen because it was easily amenable to automatic measurement, was the interval which began at the point in time corresponding to the minimum gradient point of the waveform rms energy curve in the region in which the rms energy decreases from its value for the steady state of the vowel to the closure for the stop consonant, and ended at the point in time corresponding to the point of maximum slope of the waveform energy curve in the region in which the rms energy curve increases after the plosive burst to its (much higher) value during the steady state of the vowel following the consonant. It was found that this measurement reflected a complex interaction of junctural, voicing, place of articulation and speaker differences. Some of these effects are illustrated in figure 1. This interaction of effects meant that this timing parameter is of little use as a primary recognition determiner although it can be used as a check on extreme cases.

FORMANT TRANSITION PHENOMENA

A detailed description of the results for formant transitions has been given elsewhere [18]. In summary, it was found that the second formant transitions did indeed display strong locus effects with a low locus range for labial plosives, an intermediate locus range for alveolar plosives and two locus ranges for velar plosives - a high locus range if the vowel receding the consonant was the high front vowel /i/ and a low locus range if the vowel preceding the locus range was the low back vowel /ɔ/. Indeed it was found that the position of the locus was primarily determined by the vowel preceding the consonant with the vowel following the consonant having a modifying effect on this locus. Within sex groupings inter-speaker differences of locus positions were slight while the differences between male and female locus

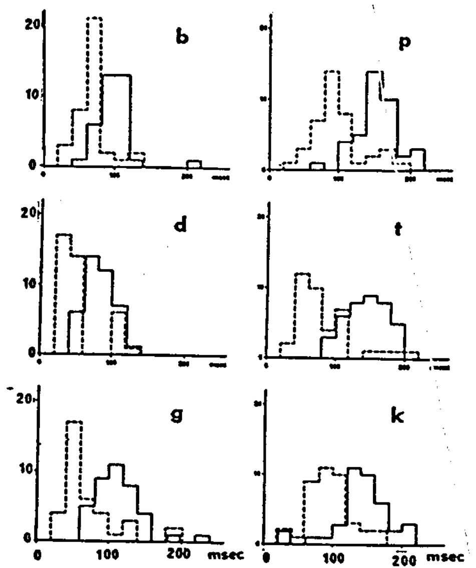


Figure 1 : Typical examples of bilabial, alveolar and velar bursts for male speakers in a/ɔ-ɔ/ context.

positions are reflections of the differences between vowel second formant positions for male and female voices.

Transition timing effects were also investigated. Transition lengths were found to be somewhat dependent on the place of articulation of the consonant. In particular it is very noticeable that transitions from vowels to the velar plosives (/k/, /g/) and from velar plosives to vowels are on average much shorter than transitions to and from labial and alveolar plosives.

BURST PHENOMENA

It was found that the burst spectral shapes fell into the three categories. Labial bursts were diffuse and of low level energy. Alveolar bursts were generally not so diffuse and were higher level energy bursts than the labial bursts. Alveolar bursts are most prominent in the 2.5-4 kHz region while labial bursts tend to be in the 1-3 kHz region. Velar bursts characteristically display two narrow-bandwidth peaks. The more prominent of these occurs in the 0.7-2.8 kHz region, and the (generally) smaller one occurs in the range 3.5-5 kHz. The exact position of these peaks is dependent on the nature of the surrounding vowels. Typical examples of the three burst shapes are given in figure 2.

With regard to the effects of coarticulation and the burst spectrum the most noticeable effect is that coarticulation effects are more evident and more consistent in velar than in alveolar plosives. And coarticulation effects are more evident and more consistent in a alveolar plosives than in bilabial plosives. Another effect is that different speakers can manifest coarticulation effects in bursts in a variety of ways and to a variety of degrees. However coarticulation effects in velar bursts are quite spectacular for all speakers. Such effects are most strongly indicated by the position of the lower in frequency (and generally higher in amplitude) of the two peaks of the typical velar spectrum. An example of this phenomenon is displayed in figure 3. Male/female differences in bursts again tend to reflect male/female differences in vowel formants. It should be noted that 7% of voiced plosive consonants were produced without any discernable burst. Certain speakers are more prone to this mode of production than others. Generally however these speakers produce very clear formant transitions.

A PLOSIVE RECOGNITION ALGORITHM

A algorithm for automatically classifying the plosive consonants was developed from this study. In particular it was developed from the results for eight speakers and tested on the remaining two. This algorithm is described in detail in [19]. The rules were speaker-independent within each sex grouping and produced a fuzzy estimate of the likelihood that any unknown plosive was produced in any of the three possible places of articulation. The rules primarily involved three fundamental measurements - measurements of bursts, measurements of formant transition endpoints and measurements of formant transition slopes. The rules are constructed such that a variety of speaker variation effects (such as burst non-production) are allowed for. Using these rules it was found that 92% of the plosive consonants tested were correctly and uniquely classified with membership greater than or equal to 0.5. The velar consonants were the most successfully recognised class of sounds with an average of 95% correct recognition. The labial consonants were recognised on average 90% of the time and the alveolar consonants were correctly recognised 91% of the time. In about 6% of all cases consonants were classified correctly but also received a simultaneously high rating in an incorrect category.

GENERALISING THE RULES

The plosive classification rules described above were incorporated in the FOPHO recognition system along with a plosive location algorithm which depended primarily on burst location and to a lesser extent on energy contour and timing effects. It was found that the overall recognition of plosives was poorer in the more general situation and that many of the refinements in the classification rules (such as those allowing for burst non-production) were rarely not invoked as many plosives productions that might have been classified by these rules were not located. Thus the usefulness of studies such as the one described here is limited unless plosive location algorithms are good.

Also as only about 25% of all plosive productions in continuous speech occur in VCV contexts the sophisticated coarticulation phenomena noted for these situations is only of limited immediate usefulness. Nevertheless they have proved a useful guide for other contexts and results for VCC contexts (where the plosive is the consonant after the vowel) tend to confirm that the vowel before the consonant is the primary determiner of transition coarticulation phenomena.

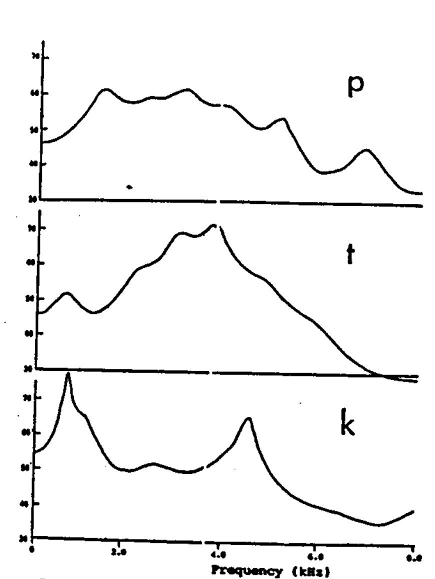


Figure 2 : Histograms showing the distribution of the timing parameter. The dotted lines represent the /VC V/ cases and full lines represent the /V CV/ cases.

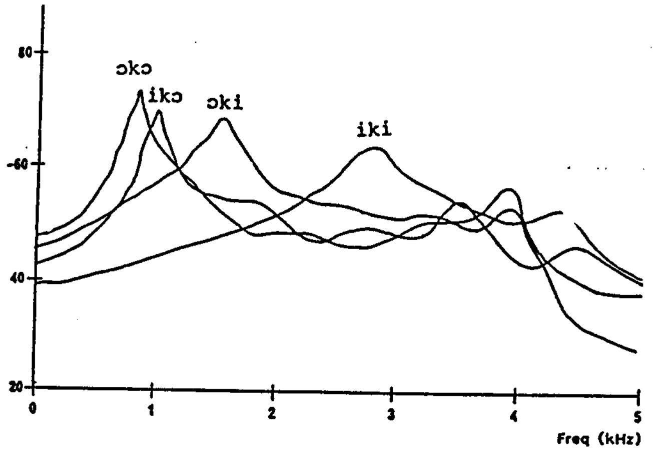


Figure 3 : Velar bursts showing coarticulation effects for a male speaker.

REFERENCES

[1] M. O'Kane, 'The FOPHO Speech Recognition Project', Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, 1983, pp.630-632.

[2] F.S., Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst & L.J. Gerstman, 'Some experiments in the perception of synthetic speech sounds', J. Acoust. Soc. Amer., 24, 1952, pp.597-606.

[3] A.M. Liberman, P.C. Delattre & F.S. Cooper, 'The role of selected variables in the perception of the unvoiced stop consonants', American Journal of Psychology, 65, 1952, pp.497-516.

[4] P.C. Delattre, A.M. Liberman & F.S. Cooper, 'Acoustic loci and transitional cues for consonants', J. Acoust. Soc. Amer., 27, 1955, pp.769-773.

[5] H.S. Hoffman, 'Study of some cues in the perception of voiced stop consonants', J. Acoust. Soc. Amer., 30, 1958, pp.1035-1041.

[6] M. Halle, G.W. Hughes & J.P.A. Radley, 'Acoustic properties of stop consonants', J. Acoust. Soc. Amer., 29, 1, 1957, pp.107-116.

[7] S.E.G. Öhman, 'Coarticulation in VCV utterances: Spectrographic measurements', J. Acoust. Soc. Amer., 39, 1966, pp.151-168.

[8] M.,Dorman, M. Studdert-Kennedy & L. Raphael, 'Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context sensitive cues', Perception and Psychophysics, 22, 1977, pp.109-122.

[9] K.N. Stevens & S.E. Blumstein, 'Invariant cues for place of articulation in stop consonants', J. Acoust. Soc. Amer., 64, 1978, pp.1358-1368.

[10] L. Lisker & A.S. Abramson, 'Some effects of context on voice onset time in English stops', Language and Speech, 19, 1967, pp.1-28,

[11] A.K. Datta, N.R. Ganguli, S. Ray & B. Mukherjee 'Computer recognition of plosive speech sounds', IEEE Conference on Computers, Session on Pattern Recognition and Learning Methods, 1978, pp.122-134.

[12] P. Alinat, 'Edute du trait permettant de distinguer entre les 3 classes de consonnes explosives PB, TD, KG', Textes des exposes de 9emes Journees d'Etude sur la Parole, Lanion, France, May-June 1978, pp.297-303.

[13] C.L. Searle, J.Z. Jacobson & S.G. Rayment, 'Stop consonant discrimination based on human audition', J. Acoust. Soc. Amer., 65, 1979, pp.799-809.

[14] H. Fujisaki, H. Tanaka & N. Higuchi, 'Analysis and feature extraction of voiced stop consonants in Japanese', Trans Committee on Speech Research, Acoustics Society of Japan, No. S79-12, May 1979, pp.89-96.

[15] C.J. Weinstein, S. McCandless, L.F. Mondshein & V.W. Zue, 'A system for acoustic-phonetic analysis of continuous speech', IEEE Trans Acoust Speech Sig. Proc., Vol.ASSP-23, 1985, pp.54-72.

[16] W.A. Woods (Principal author), 'Speech understanding system final report', BBN Report No.3438, November 1974-October 1976.

[17] P. Demichaelis, R. De Mori, P. Laface & M. O'Kane, 'Computer recognition of plosive consonants using contestual information', IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-31, 1983, pp.359-377.

[18] M. O'Kane, 'Making the Locus Theory useful for automatic speech recognition', Proceedings of the Tenth International Congress of Phonetic Sciences, edited by A. Cohen and M.P.R. Van den Broecke, Foris Publications, Dordrecht, 1984, pp.331-337.

[19] M. O'Kane, 'Acoustic-phonetic processing for continuous speech recognition', Ph.D. Thesis, ANU, Canberra, 1981.

CONTEXT DEPENDENT RECOGNITION OF SPANISH STOPS

HORACIO FRANCO

JORGE A. GURLEKIAN

Laboratorio de Investigaciones Sensoriales. CONICET.
CC 53 (1453) Buenos Aires, Argentina.

ABSTRACT

The recognition performance using either the entire or selected portions of running spectra obtained from intervocalic Spanish stops is presented in the framework of a recognition system based on segmentation and context dependent statistical classification.

In order to capture the dynamic nature of this context dependent sounds, critical band spectra were uniformly grouped in a temporal sequence of contiguous segments.

A set of conditional probability density functions were estimated for the spectra belonging to each segment for every different vocalic context.

The contribution to the recognition performance of different segments was evaluated and related to the relevant acoustic features.

The performance of the recognizer was also studied under different degrees of context dependence.

Results for unvoiced and voiced stops were obtained in a speaker dependent manner using a data base consisting of 2592 emissions of the stops / p, t, k, b, d, g/, embedded in VCVCVCV sequences where the V's were the vowels /a, i, u/ uttered by two male Argentine speakers.

1. INTRODUCTION

Among the stop consonant recognition systems the best performance at present was obtained in those works which used contextual information /1/ or sequences of short time spectra as features for recognition /2/.

In this work we combined these two characteristics in the framework of a statistical approach. Our basic objective was to evaluate the recognition performance of an automatic recognition system for the Argentine Spanish stops uttered in intervocalic position based on the use of a context-dependent statistical classifier, and using the whole set of spectra along the acoustical realization as features for the classification.

With this objective we also evaluated the contribution of portions of the spectral sequence for the recognition of the stops under different degrees of context dependence, i.e., either the following or preceeding or both vowel classes which act as the a priori information for the classifier.

2. ACOUSTICAL DESCRIPTION

In the case of voiced stops, the intervocalic realization is an approximant as was defined by Ladefoged /4/ with no signs of burst and with a week evidence of consonant release. Moreover, the stops appear as a slight acoustic variation of the transition between two particular vowels. The acoustic correlates of the coarticulation of the VCV sequences follow aproximately the typical formant patterns obtained by Ohman /3/ for Swedish and American English. Quasi-stationary portions of vowels are slightly affected by the consonant and transconsonantal vowel, while the formant patterns of the occlusive portions along the energy dip are dependent of the two, initial and final, vowel classes.

On the other hand, unvoiced stops show a burst following the silent gap but weaker than in American English emissions.

3. SPEECH DATA AND SPEECH ANALYSIS

The speech data base consisted of 2592 emissions of the intervocalic stops /p, t, k, b, d, g/ combined with the vowels /a, i, u/ in all combinations, i.e., 9 vocalic contexts for the VCV sequences. The speech data was uttered by two male Argentine speakers.

As a step to test the system with running speech, three different consonants sharing the context vowels were produced embedded in VCVCVCV nonsense utterances. In this way, continuous utterances of 800 msec (on the average) were produced and the stress pattern of the VCV sequences had a greater degree of variation.

The speech samples were recorded in a low noise environment, low pass filtered to 5 kHz, sampled at 10 kHz and digitized with a 12 bit A/D converter. Each utterance was preceeded and followed by short segments of background noise and stored in separate waveform files. From the speech waveform stored on disk the following parameters were extracted every 10 msec through a sliding 25.6 msec Hamming window of the preemphasized waveform.

a) Short time energy expressed in dB.

b) Critical Band Spectra. From a 128 point short-time DFT spectra (40 Hz resolution), the energy output of an auditory filter bank was obtained following the method described by Moore /5/ but using the Zwicker's /6/ critical bandwidths and critical band rate scales instead (Fig. 1). Energy at each critical filter was obtained as a weighted sum of DFT energy values. The energy transfer function for each filter was the rounded exponential:

$$W(g) = (1 + g p) \mathrm{Exp}(- g p)$$

were $g = \mathrm{abs}(f - fc) / fc$ and $p = 4 fc / BW(fc)$ with BW(fc) the bandwidth of the critical filter centered at fc. Center frequencies were spaced one critical bandwidth starting from 100 Hz till 4500 Hz, in this way 18 filters resulted. These spectra were expressed in

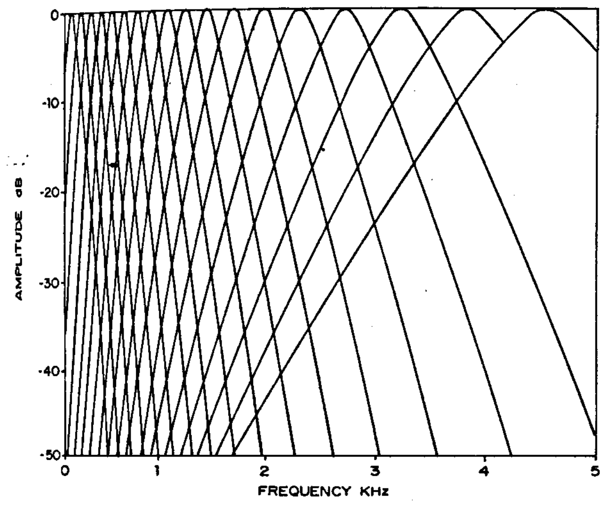


Fig. 1. Responses of the 18 individual channels of the critical filter bank.

dB and normalized by substracting the linear mean of each in logarithmic scale.

4. SEGMENTATION

The proposed scheme was: first, to find reliable points for the temporal location of the consonant using the short time energy contour, and secondly to characterize its dynamic nature through different probability distributions associated with portions of the spectral sequence.

Previous research /7/ showed the relevance of the temporal amplitude contour dips for the perceptual detection of the voiced stops, so, the segmentation strategy was based on the use of this acoustical event.

The segmentation was accomplished as follows: first, a dip detection was performed finding the local maxima and minima in the log-energy contour wich was smoothed via six passes through a zero phase digital filter with a three point triangular impulse response. With this degree of smoothing a great percentage of spurious peaks and dips were removed although the peaks and valleys associated with vowels and this kind of consonants were preserved. Following this, a dip classification was performed. For each dip and associated left and right peaks a set of features characterizing its width, depth, and abruptness were measured from the log-energy contour and its unsmoothed derivative (Fig. 2).

A statistical classifier /8/ assuming a Gaussian multivariate probability distribution was trained using each half of the data, then the dips from the other half were classified in the following categories: unvoiced dip, voiced dip, and dip non valid.

The consonantal portions were defined in valid peak-dip-peak sequences, for each of them, the points of maximum slope at the consonant closure and release were located in the log-energy contour.

In order to characterize the time varying spectral pattern of the consonant, the time scale was linearly mapped to seven segments and the spectra belonging to each segment were associated with a single probability distribution. The two points of maximum slope were used as anchor points of the seven linearly distributed segments with the second and sixth segment tied to those

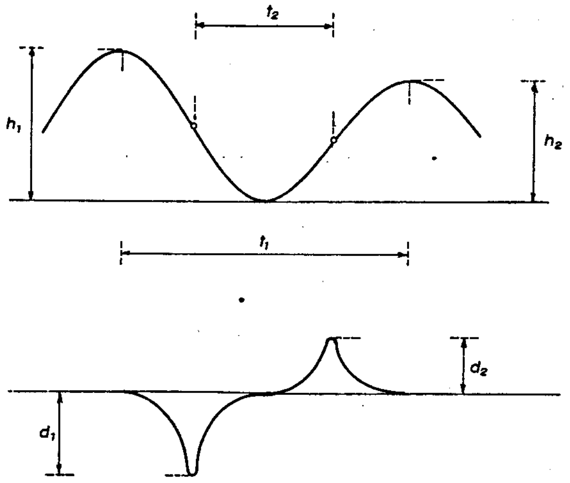


Fig. 2. On top: log-energy contour showing the measurements used for the definition of dip features (Fn); F1 = h1+h2; F2 = t1, and F3 = t2. Bottom: First derivative contour of the log-energy showing one additional feature: F4 = d1+d2.

anchor points. Five segments span between the anchor points including them, and two others were extrapolated at the extremes.

Even though the unvoiced stops in Argentine Spanish present a clear release, the same scheme as for the voiced stops was used, given that in VCV contexts, the spectra belonging to both the VC transition and the CV transition could provide evidence for the recognition.

This procedure provided an aproximate time normalization for the dynamic portions and a simple method of alignment for the test and reference patterns.

Vocalic segments were defined either between two consecutive consonantal portions or between the endpoints of the utterance and the closest consonantal portion.

5. THE CONSONANT CLASSIFIER.

The features selected for training and classification were normalized critical band spectra obtained at the seven segments located along the log-energy dip as defined above. Given the unstationary and context-dependent nature of the consonantal realizations, different probability distributions were assumed for the spectra at each one of the seven segments, for each one of the different vowel contexts considered.

A bayesian context-dependent classifier /8/ was designed according to the following assumptions:

a) the class conditional probability density functions of the spectra at each segment are Gaussian independent,

b) there is statistical independence between spectra,

c) the vowel and consonant classes are equiprobable and independent.

In the training phase the speech data from each speaker were split in two halves. From each half, maximun likelihood estimates of the mean vector and the assumed diagonal covariance matrix of each class conditional probability density function were obtained after a supervised segmentation.

In the recognition phase the classifier was run over each data half with the parameters obtained from the other half. Under the assumption that the classes of adjacent vowels are known, the parameters of the consonant classifier were chosen among those obtained in the training phase for the different vowel contexts considered. To accomplish the recognition, a conditional log-likelihood Lj/c for each consonant class j and the given context c, was obtained as:

$$L_{j/c} = - \sum_{k} \sum_{i=1}^{18} \left[\frac{(Y_{ik} - M_{isjc})^2}{V_{isjc}} + \log(V_{isjc}) \right]$$

with s = S(k)

where Yik is the value of the normalized spectra corresponding to the ith critical band at time index k, and Misjc and Visjc are respectively the mean and variance associated with the ith critical band of the spectra corresponding to the segment s for the jth consonant class in the assumed known context c. The segment index s is obtained from the time index k through the segmentation mapping s = S(k).

So, L is like a "global distance" computed as the sum over all the segments, of the weighted euclidean distances between the spectra belonging to each segment and the corresponding mean spectra, plus a segment dependent term. The weights are the inverses of the variances of the corresponding spectral samples.

The classifier performed three-way /p, t, k/ or /b, d, g/ discriminations. The recognized consonant corresponded to the largest likelihood.

The classifier was run using the spectra belonging to all the segments and alternatively using the spectra obtained from single segments to evaluate their particular recognition performance.

The degree of context dependence was given by the classifier training and operation according to the consideration of three alternative cases. In the first, the information of the preceeding and following vowel classes is used to select different probability distributions. This case will be referred to as VCV recognition. In the second and third cases the information of only the preceeding or the following vowel classes is used. These cases will be referred to respectively as VC or CV recognition.

The context vowels were recognized, using the spectra belonging to the vocalic segments, by means of a similar statistical classifier that uses a single probability distribution for each vowel class given its assumed quasi-stationarity and context independent nature.

6. RESULTS

6.1 Segmentation.

The performance of the dip detector and classifier to discriminate between valid consonant dips or invalid dips was of 99.7% for speaker 1 and 98.7% for speaker 2. The voiced-unvoiced discrimination among the valid consonant dips detected, reached a 96,1% for speaker 1 and 97.5% for speaker 2.

6.2 Consonant Recognition.

The recognition rate of the VCV, CV and VC cases using the spectra corresponding to all and single segments are presented for each speaker and for the unvoiced and voiced stops in Table I and Figs. 1 to 4.

	VCV	CV	VC
Sp.1			
Unvoiced stops	92.7	90.4	80.3
Voiced stops	88.5	76.4	73.6
Sp.2			
Unvoiced stops	94.3	94.2	88.3
Voiced stops	90.8	75.5	77.0

Table I. Recognition scores (%) corresponding to the performance when the information of all segments is used.

Under the VCV context condition an evaluation of the individual spectral segments was obtained. As it can be seen the recognition performance is relatively uniform over the seven segments in the case of voiced sounds. For the unvoiced stops it is clear the highest contribution showed by the segments corresponding to the stop release and the transitional part towards the following vowel.

When the knowledge of the context that was accounted for the classifier training and operation was restricted to only the following vowel (CV recognition), the voiced stops aproximately doubled the error rate obtained for the VCV case. However for the unvoiced stops the performance held similar values.

On the other hand for the VC recognition, the voiced stops gave similar results as for the CV recognition showing that in this case there is no clear preference for the accounting of the following or previous vowel context. This was not the case for the unvoiced stops. Accounting of the following vowel clearly gave a better performance than using the preceeding vowel.

With reference to the performance of the individual spectral segments under the different degrees of context dependence, it can be observed that the segments close to the vowel not considered in the VC and CV context conditions significatively lowered their recognition scores, while the segments located near the opposite vowel approached the values corresponding to the VCV case.

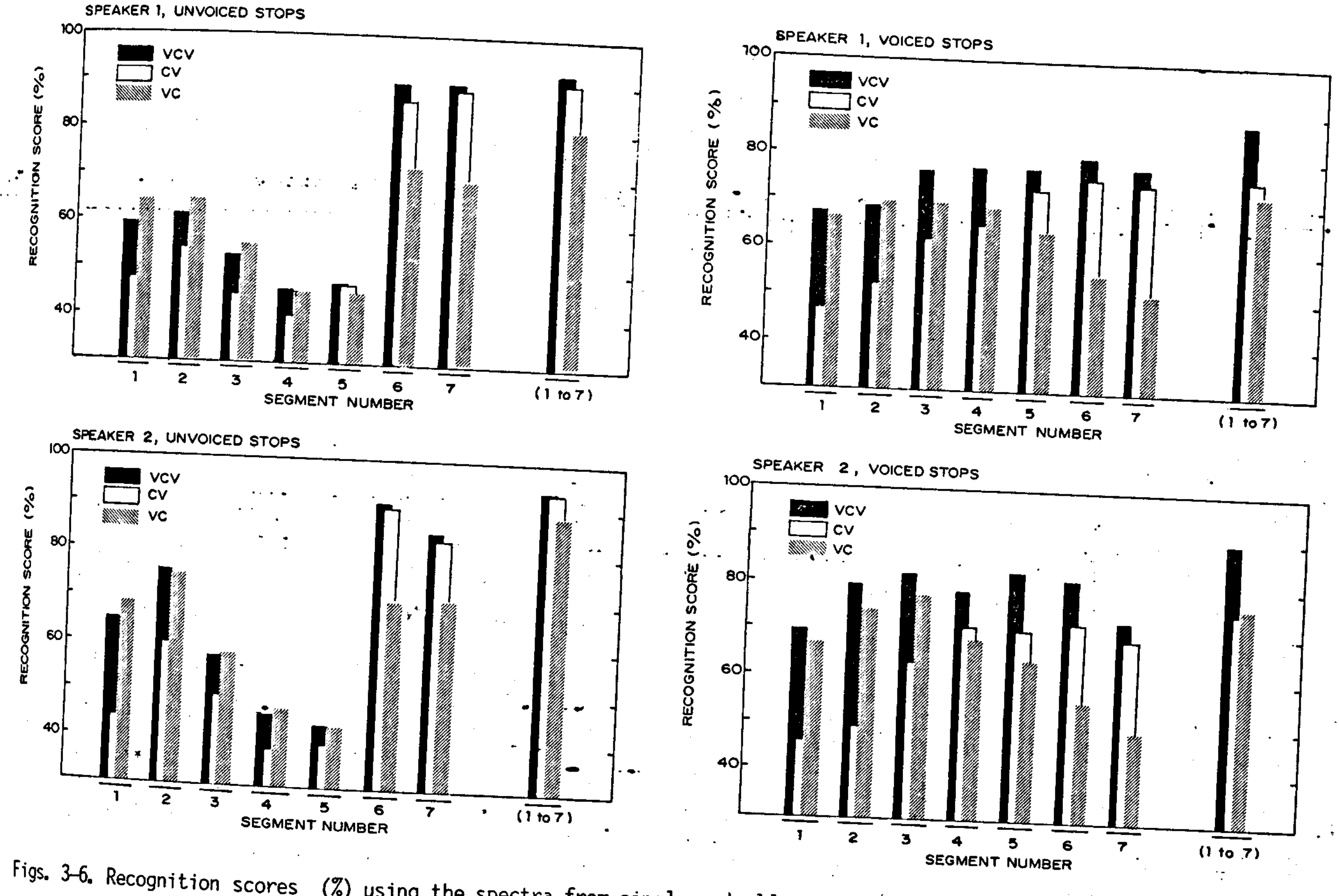
The performance results discriminated for every vocalic context for the VCV recognizer ordered by decreasing performance are presented in Table II. The average recognition rate ranged from 99.6% for the /a-a/ context to 81.8% for the /i-u/ context.

	a-a	a-i	i-a	u-a	i-i	a-u	u-u	u-i	i-u
Sp.1									
U	100	98.6	91.7	98.6	94.4	95.8	81.9	81.9	91.7
V	100	90.2	100	88.9	94.4	84.7	76.4	77.8	83.3
Sp.2									
U	100	98.6	93.1	93.1	93.1	95.8	94.4	93.1	87.5
V	98.6	95.8	97.2	98.6	87.5	91.7	95.8	86.1	65.3

Table II. Recognition scores (%) for unvoiced (U) and voiced (V) stops discriminated by vocalic contexts.

Considering that the contribution of segments corresponding to the silent gap in the unvoiced sounds could introduce noisy information to the classifier the recognizer was also run using only selected segments such as, numbers 2, 6, and 7 which presented the best individual scores. For this case the recognition scores are presented in Table III.

Given the limited data available to train the recognizer the supression of the noisy information effectively improved the scores for unvoiced stops.



Figs. 3-6. Recognition scores (%) using the spectra from single and all segments (1-7) under different degrees of context dependence (VCV, CV, and VC), presented for both speakers.

The recognition of the context vowel classes was accomplished with a high performance giving only one error over the total of 3456 vowels recognized for both speakers.

(VCV)		
Unvoiced stops	Speaker 1 95.1	Speaker 2 96.5

Table III. Recognition scores using selected segments.

8. CONCLUSION

In this work we have tested a statistical approach to the recognition of unvoiced and voiced stops.

The best results for the voiced stops were obtained when the "two vowel" context dependent classifier was used. For the unvoiced stops the classifier achieved similar higher scores when using either the "two vowel" context or only the following vowel context.

Regarding the performance of the individual segments of the spectral sequence, the results obtained under the VCV vowel context condition showed that for the voiced stops the information was quite uniformly distributed along the whole VCV pattern. For the unvoiced stops there was a clear dominance of segments around the burst and transitions to the following vowel.

These results suggest that at least for Spanish, an intervocalic consonant recognition strategy could be based in the use of CV units for unvoiced stops but it should be based on larger units as the VCV for the voiced stops.

The authors wish to thank Dr. B. Cernuschi for his helpful suggestions.

10. REFERENCES

/1/ P. Demichelis, R. De Mori, P. Laface and M. O'Kane, "Computer Recognition of Plosive Sounds Using Contextual Information", IEEE Trans. Acoust., Speech and Signal Processing, Vol ASSP-31, 359-377, 1983.
/2/ G. E. Kopec, "Voiceless Stop Consonant Identification Using LPC Spectra", IEEE Trans. Acoust. Speech and Signal Processing, ICASSP84. March 19-21, 1984.
/3/ S. E. G. Ohman, "Coarticulation in VCV Utterances: Spectrographic measurements", J. Acoust. Soc. Am., 39, 151-168, 1966.
/4/ P. Ladefoged, "A Course in Phonetics", Nueva York, Harcout Brace Jovanovich, Inc., 1975.
/5/ B. Moore, B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acoust. Soc. Am., 74, 750-753, 1983.
/6/ E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am.,68, 1523-1525, 1980.
/7/ H. Franco, J. A. Gurlekian, "Recognition of Spanish intervocalic consonants", J. Acoust. Soc. Am., Vol. 77 S1, S27, 1985.
/8/ R. Duda, P. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, 1973.

CARACTERISTIQUES ACOUSTIQUES DE LA PREMIERE CONSONNE DANS UN GROUPE DE DEUX OCCLUSIVES

Alain MARCHAL et Anne FOTI

U.A. 261, CNRS, Institut de Phonétique,
29, av. Robert-Schuman, 13621 Aix-en-Provence, France

INTRODUCTION

L'étude détaillée de l'organisation des appuis linguo-palatins dans des groupes d'occlusives observée à l'aide de l'électropalatographie **(1)** nous a permis de mettre en évidence des phénomènes de :

- coproduction consonnantique
- coproduction vocalique
- double occlusion
- relâchement de C1 caractéristique de la production de "clics".

Nous nous attacherons dans cette communication à décrire les caractéristiques acoustiques de la première consonne dans un groupe d'occlusives. Nous essayerons de voir s'il est possible de reconnaître automatiquement C1 à partir des informations spectrales ou temporelles apportées par le bruit de relâchement.

Corpus et méthode

Les données proviennent de 3 corpus différents qui ont été répétés de 3 à 6 fois par 3 locuteurs. Il s'agit d'un ensemble de 800 phrases naturelles de 4 à 6 syllabes où les groupes d'occlusives formées par les consonnes /t, d, k, g/ apparaissent dans plusieurs contextes syntaxiques, précédées et suivies des voyelles /i, a, u/. Le débit rapide favorisait les phénomènes de coarticulation.

La segmentation des événements articulatoires a été opérée à partir de la séquence des appuis linguo-palatins relevés par le système de palatographie dynamique de Montréal **(2)**. Nous avons délimité les événements suivants :

1 - Début de la préparation de C1; de C2.
2 - Implosion de C1; de C2.
3 - Tenue de C1; de C2.
4 - Relâchement de C1; de C2.

Nous avons examiné l'évolution correspondante du signal acoustique. Le son synchrone a été digitalisé et analysé. Compte tenu de la sensibilité des extracteurs de maxima fréquenciels au bruit (conditions d'enregistrements, rapport signal sur bruit, nombre de points d'analyse, gain, fenêtrage, nombre de paramètres, ...), nous avons jugé dans la perspective d'une application en reconnaissance de la parole qu'il fallait

1 - utiliser des procédures automatiques
2 - admettre la possibilité d'erreurs de détection
3 - comparer les résultats d'analyse provenant de systèmes différents pour éviter les artéfacts induits par une méthode et une situation expérimentale donnée.

Nous avons relevé les données acoustiques suivantes :
les valeurs des trois premiers maxima de fréquence pour les voyelles et les consonnes et leur amplitude relative dans les parties stables et dans les transitions, les durées des tenues de C1 et de C2, le VOT, et la durée de bruit de constriction, soit un ensemble de 39 mesures pour chaque suite V1C1C2V2.

Nous avons utilisé pour l'analyse : 1) le logiciel de traitement de signal de parole "SIGNEX" implanté à l'Institut de Phonétique d'Aix-en-Provence **(3)**, 2) le logiciel commercial "ILS" de SIGNAL TECHNOLOGY, 3) un banc de filtres numériques **(4)**.

Résultats et discussion

Vu l'importance du corpus et le très grand nombre de mesures à effectuer, le travail d'analyse et d'interprétation des données se poursuit encore actuellement. L'état de la recherche est le suivant : le signal acoustique de tous les enregistrements a été segmenté et les 8000 marqueurs correspondant aux événements articulatoires ont été placés. L'analyse acoustique complète a porté pour deux locuteurs sur les phénomènes temporels et sur la structure acoustique du bruit de relâchement pout /t, k/. Nous avons interprété les résultats en fonction de la possibilité d'écrire des règles de production simples et sûres.

Un invariant spectral ?

La thèse voulant qu'un invariant spectral permette d'identifier la consonne occlusive a été examinée **(5)**. Lorsqu'on calcule les valeurs moyennes des trois premiers pics depuis les basses fréquences indépendamment de leur amplitude relative, tous corpus confondus, on trouve

/t/ :		/k/ :	
/t/ :	1304 Hz;	/k/ :	700 Hz
	2887 Hz		1985 Hz
	4416 Hz		3806 Hz

si l'on considère seulement les valeurs absolues, il semblerait possible de distinguer ainsi les deux occlusives; il apparaît toutefois que les valeurs d'écarts-type sont trop importantes; respectivement pour /t/ : 926 Hz; 1320 Hz; 1670 Hz et pour /k/ : 547 Hz; 904 Hz; 1569 Hz. Celles-ci font apparaître une zone de recouvrement importante entre les maxima 1 et 2 et entre les maxima 2 et 3.

On pourrait objecter que l'amplitude relative des zones de bruit est essentielle pour les consonnes et que les valeurs de fréquence doivent être ordonnées en fonction de ce paramètre.

Nous avons remarqué que le spectre de bruit de relâchement de C1 dans un groupe d'occlusives était caractérisé par une amplitude et par une dynamique faibles. Les valeurs de maxima d'énergie par ordre décroissant et les écarts-type font apparaître une zone de recouvrement moins grande pour /t/ et /k/ mais encore trop importante pour pouvoir distinguer avec confiance les deux consonnes.

	/t/			/k/	
	moy.	écart	:	moy.	écart
Max 1 :	1983 Hz	1607 Hz	:	996 Hz	787 Hz
Max 2 :	2873 Hz	1716 Hz	:	1860 Hz	1157 Hz
Max 3 :	3776 Hz	1789 Hz	:	3623 Hz	1714 Hz

Nous n'avons pas pu identifier un ou plusieurs facteurs stables permettant de réduire la dispersion autour de la valeur centrale. L'influence de V1 et V2 ne suffit pas à expliquer la variabilité. La méthode d'extraction des maxima ne peut non plus être invoquée car la dispersion se retrouve à peu de choses près identique dans les trois analyses.

Un invariant temporel ?

La physiologie de l'articulation (Hardcastle 1976) nous enseigne que la masse linguale mise en mouvement et que l'activité musculaire requise pour la production de /t/ et /k/ sont très différentes. On pourrait alors faire l'hypothèse qu'une cible peut être atteinte plus rapidement que l'autre et que cette différence cinétique, due à des contraintes fonctionnelles pourrait se retrouver dans le signal acoustique. Nous avons examiné les durées des tenues seules et les durées des bruits de relâchement. Il est apparu assez rapidement que la durée du bruit de relâchement variait considérablement et que dans un corpus important les différences phonémiques intrinsèques /k>t/ étaient annulées par des variables mal contrôlées (conditions d'enregistrement, proximité du micro, gain..., conditions phonétiques : débit, ...).

Nous avons alors concentré notre attention sur les tenues silencieuses; soit l'intervalle séparant l'implosion de C1 de l'apparition de la trace du bruit de relâchement sur les tracés acoustiques. Nous avons aussi relevé la tenue silencieuse totale de C1 et C2.

Les résultats complets en ms apparaissent dans le tableau suivant :

	/t/	/tk/	/k/	/kt/
	C1	C1C2	C1	C1C2
Moy. :	75	177	75	155
Ecart :	12	27	14	29

On note que les moyennes des durées des tenues silencieuses de /t/ et /k/ sont très comparables. Ce qui importe toutefois c'est que le rapport de la durée de C1 sur la durée totale de C1 et C2 est significativement différent.

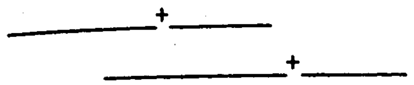


Schéma des durées relatives des tenues de C1 et C2

On observe la même tendance pour l'enchaînement de /d/ à /g/ ou de /g/ à /d/. Notre observation sur le français rejoint les résultats de Hardcastle & Roach (1977) **(6)** sur l'anglais; ceci semble bien indiquer que ce phénomène temporel est lié à une contrainte d'ordre physiologique de bas niveau. Il pourrait s'agir d'une boucle tactile déclenchant le mouvement vers la deuxième occlusive dès que l'information tactile appropriée est transmise par le contact linguo-palatal de la première occlusive. Dans le cas d'un enchaînement "alvéo-dentale---vélaire", l'information pourrait être envoyée par des mécano-récepteurs de la langue quand elle fait contact sur les côtés du palais et sur les alvéoles. A partir de cette position, tout ce qui est nécessaire pour la production d'un /k/ est une contraction d'un muscle intrinsèque de la langue : le longitudinal inférieur. Un tel mouvement peut se produire relativement rapidement parce que les muscles intrinsèques ont des temps de contraction rapides. Dans le cas d'un enchaînement "vélaire---alvéo-dental", la situation n'est pas aussi simple et plusieurs gestes deviennent antagonistes. Le mouvement vers le /t/ implique un mouvement vers l'avant et vers le haut. Celui-ci est dû essentiellement à l'activité du génioglosse, muscle extrinsèque lent qui provoque une remise en position de tout le corps de la langue. L'élévation de l'apex étant produite par la contraction du longitudinal supérieur et du muscle transverse. Il n'est donc pas étonnant que cette différence fonctionnelle se retrouve dans les données temporelles.

CONCLUSION

Dans le cas de l'enchaînement de deux occlusives, il ne semble pas possible d'identifier un invariant spectral permettant la reconnaissance de la première consonne. L'organisation des données temporelles et notamment le rapport des tenues silencieuses de /t/ et /k/ mettent en évidence une contrainte physiologique universelle qui pourrait être utilisée en reconnaissance automatique de la parole. Le groupe /tk/ est plus long que le groupe /kt/ et surtout le rapport de la durée de la tenue silencieuse de C1 sur la durée de la tenue silencieuse totale est environ de 0,5 lorsque C1 est un /k/ et s'abaisse autour de 0,4 lorsque C1 est un /t/.

REFERENCES BIBLIOGRAPHIQUES

(1) Marchal, A. (1985), "L'électropalatographie : contribution à l'étude de la coarticulation dans les groupes d'occlusives", th. de doct. d'Etat, Nancy.

(2) Marchal, A. (1984), "Le système d'électropalatographie de Montréal : contribution à l'étude des occlusives du français", **Trav. Inst. Phon. Aix**, 9 : 267-341.

(3) Espesser, R. & Nishinuma, Y. (1984), "Traitement de signal sous le système UNIX : description des commandes", Institut de Phonétique d'Aix-en-Provence.

(4) Manceron, F. (1982), "Contribution à l'analyse spectro-temporelle du signal de parole considéré comme une unité d'impulsion acoustique", th. doct. ing. LIMSI, Paris.

(5) Blumstein, S. & Stevens, K. (1979), "Acoustic invariance in speech production : evidence from measurement of the spectral characteristics of stop consonants", **J. Acoust. Soc. Am.**, 66 (4) : 1001-1013.

(6) Hardcastle, W. & Roach, P. (1977), "An instrumental investigation of coarticulation in stop consonant sequences"; **Work Prog. Phonet. Lab. Univ. Reading**, 1 : 27-44.

ОБ ОДНОМ ПОДХОДЕ К ВОПРОСУ ФОНЕТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ГРУППЫ ЩЕЛЕВЫХ СОГЛАСНЫХ И АФФРИКАТ РУССКОГО ЯЗЫКА

М.Ф. БОНДАРЕНКО, А.Н. ГАВРАШЕНКО

Кафедра вычислительной техники
Институт радиоэлектроники
Харьков, Украина, СССР, 310141

АННОТАЦИЯ

В докладе рассматривается группа щелевых согласных (С,З,Ш,Ж,Ф,Х') и аффрикат (Ц,Ч) русского языка, предлагаются методы сегментации и фонетической идентификации этой категории звуков, надёжно работающие как при обработке изолированных слов, так и слитной речи. Предлагаемые алгоритмы используют результаты временного и спектрального анализов речи, а также результаты анализа тонкой структуры речевых сигналов, дискретизированных с учётом эффекта сглаживания в слухе /1/.

При разработке надёжных и эффективных методов распознавания отдельных фонем в речевом потоке важное значение приобретает решение задачи сегментации речи на составляющие её звуки. От успешности её решения зависит успех в решении задачи надёжного распознавания фонем и всего распознаваемого сообщения в целом.

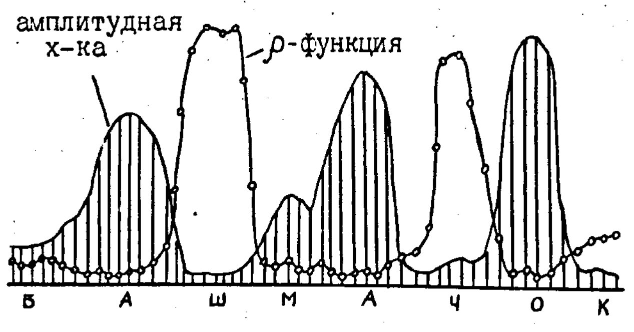
Для надёжного выделения рассматриваемых в докладе фонем, из изолированно произносимых слов или слитной речи, предлагается метод сегментации речевого сигнала на участки, фонетически соответствующие шумовым согласным. В практике распознавания речи задача сегментации решается различными методами. Наибольшее распространение получили спектральные методы и методы анализа клиппированного сигнала. Однако, использование для сегментации процедур спектрального анализа, требующих больших вычислительных затрат или специальных устройств, не всегда целесообразно в практических системах.

Для решения задачи сегментации речевого сигнала на участки, соответствующие шумовым звукам, предлагается следующий алгоритм сегментации, обладающий высокой помехоустойчивостью к воздействию аддитивных и локальных помех. Разработанный алгоритм ориентирован на использование в условиях с высоким уровнем окружающих акустических шумов, вплоть до 80 дБ.

При разработке алгоритма сегментации была использована графическая форма представления речевых сигналов. Для этого необходимо выполнить кратковременный анализ интенсивности сигнала и функции нулевых пересечений. Такие графические изображения, названные условно "динамическими портретами, отражают динамику исследуемых параметров речевого сигнала во времени. Отличительной особенностью рассматриваемой в докладе категории звуков является ярко выраженный шумовой характер этих фонем, что отличает их от всех остальных звуков русского языка. Для выделения шумной части звука анализ функции нулевых пересечений выполнялся на интервале скользящего с шагом 5 мс временного окна в 25 мс. Количество всех пересечений нуля в таком окне определяет одну точку графика. Метод перекрывающегося окна был использован с целью устранения присутствующих в речевом сигнале неоднородностей. Для устранения некоторой "лохматости" картинки "динамический портрет" сглаживается методом скользящего усреднения с интервалом сглаживания, равным 7 точкам графика. Полученное таким образом изображение достаточно наглядно показывает места расположения шумовых согласных в потоке речи.Для определения границ таких участков используется метод сечений, который в общей сложности предполагает построение двух сечений. Одно из них проводится на некотором усреднённом уровне, уровень проведения второго зависит от индивидуальных особенностей диктора. Срез по сечениям и определяет границы шумовых участков в речевом сигнале.

Выделением из речевого потока шумовых участков завершается первый шаг работы алгоритма сегментации. Известно, что аффрикаты являются более сложными по своему составу звуками по сравнению со щелевыми согласными ввиду присутствия кроме шумного участка ещё и участка смычки.Поэтому выделение из акустического сигнала шумового сегмента ещё не определяет всего согласного. Это обстоятельство, а также необходимость уточнения границ щелевых согласных, предусматривает переход ко второму шагу работы алгоритма сегментации, на котором строятся "динамические портреты" другого рода. Для этого используются результаты кратковременного анализа не только функции нулевых пересечений, но и интенсивности речевого сигнала на интервалах непересекающихся 10 мс сегментов. Выбор интервала обусловлен особенностями тракта речеобразования человека. Результатом кратковременного анализа интенсивности акустического сигнала является набор значений абсолютных максимумов сигнала на каждом из 10 мс сегментов. Неоднородности, как правило присутствующие в изображении, сглаживаются процедурой скользящего усреднения с окном, равным 5 точкам графика. Построенный таким образом "динамический портрет" даёт возможность разграничить участки расположения в речевом сигнале гласных и согласных звуков по изменениям амплитудной характеристики.

В подавляющем большинстве случаев согласные звуки в речи располагаются в окружении гласных. Находясь между гласными, шумовые согласные довольно отчётливо определяют своё место на "динамическом портрете". При этом поиск их границ осуществляется на основе анализа характера изменения интенсивности окружающих гласных и ρ -функции на участке шумового согласного. В качестве примера можно рассмотреть слово "БАШМАЧОК". "Динамический портрет" на втором шаге работы алгоритма сегментации для этого слова представлен на рисунке 1.



"Динамический портрет" слова "БАШМАЧОК"

Рис. 1

Из изображения видно, что границы между шумовыми согласными и окружающими их гласными расположены на пересечении огибающей интенсивности и ρ -функции. Слуховой анализ выделенных участков речевого сигнала подтверждает правильность выделения границ.

Если в речевом потоке шумовые согласные расположены в окружении согласных звуков (кроме исследуемых шумовых), то в качестве границ принимаются границы шумовой части фонемы, чего вполне достаточно для щелевых согласных, содержащих в своём составе только шумный участок. Выделение же участка смычки у аффрикат осуществляется на следующих этапах анализа.

Таким образом, использование описанного алгоритма сегментации позволяет выделять из потока речи шумовые согласные независимо от контекста при условии, что две шумовые согласные не могут располагаться в речи рядом. Выделяемый алгоритмом звук "Щ" исключается из рассмотрения по признаку характерной длительности шумового участка, свойственной только этой фонеме.

Для осуществления предварительной фонетической идентификации выделенных из речи шумовых участков был разработан алгоритм, позволивший на базе небольшого количества признаков осуществить грубую маркировку исследуемых согласных.

Как показал анализ множества осциллограмм и "динамических портретов" словосочетаний с шумовыми согласными в различных контекстах и положениях в слове, даже стационарная часть звуков характеризуется большой вариабельностью и неоднородностью характеристик. В связи с этим из всего количества присутствующих на стационарном участке элементарных сегментов необходимо выбрать группу таких рядом расположенных сегментов, которые характеризуются наибольшей однородностью и стабильностью в рамках описанного ниже критерия. Оценка фонетического качества шумовых согласных на таких акустически однородных участках является наиболее устойчивой. Как показали эксперименты, шумовые фонемы можно надёжно идентифицировать на интервале трёх рядом расположенных акустически однородных 10 мс сегментах. При этом не требуется больших вычислительных затрат на определение характерных сегментных признаков.

Объединение сегментов в акустически однородные области производилось по следующему критерию. Была введена мера близости между тремя соседними сегментами стационарной части шумовых согласных. При этом на каждом сегменте вычислялись значения двух признаков: коэффициент монотонности, характеризуемый общим количеством участков монотонного изменения сигнала и количество мгновенных значений речевого сигнала, превышающих некоторый уровень С. Три рядом стоящих сегмента объединяются в акустически однородную область по минимуму среднеквадратического отклонения значений признаков на этих сегментах от их среднего уровня. На выделенном таким образом участке звука будет вестись его дальнейшее распознавание.

Для предварительной фонетической идентификации выделенных из речи шумовых согласных сформируем систему признаков. В эту систему были включены ряд широко используемых (энергия сигнала, число переходов сигнала через нуль, нормализованный коэффициент автокорреляции, максимальная амплитуда сигнала на сегменте и другие), а также такие признаки, как коэффициент монотонности и количество значений сигнала на сегменте, превышающих некоторый порог.

Как показали проведенные эксперименты, для целей предварительной классификации шумовых согласных на группы важную информацию несут такие признаки, как коэффициент монотонности огибающей речевого сигнала (F), количество нулевых интервалов меньших 50 мкс (L) и общее количество мгновенных значений сигнала, расположенных выше некоторого уровня $D_{пор}$ (G). Значения признаков вычисляются на интервале 10 мс сегментов. Так, признаки F и L позволили разделить все исследуемые шумовые согласные на классы

$$\left.\begin{array}{l}(\omega^{с}, \omega^{з}, \omega^{ч}) \in \Omega_1 \\ (\omega^{ш}, \omega^{ж}, \omega^{ф}, \omega^{2}, \omega^{х'}) \in \Omega_2\end{array}\right\} \quad (1)$$

Признак G позволил выделить из класса Ω_2 фонемы "Ш", "Ж", "Ч". В результате получили систему из трёх классов

$$\left.\begin{array}{l}(\omega^{с}, \omega^{з}, \omega^{ч}) \in \Omega_1 \\ (\omega^{ш}, \omega^{ж}, \omega^{2}) \in \Omega_2 \\ (\omega^{ф}, \omega^{х'}) \in \Omega_3\end{array}\right\} \quad (2)$$

Аффрикаты "Ц" и "Ч" надёжно выделяются из классов Ω_1 и Ω_2 по значению признака G, вычисленному на 10 мс сегментах всего звука и n сегментах предшествующих ему ($n = 9$). Значение порога $D_{пор}$ для каждой из аффрикат "Ц" и "Ч" выбираются с учётом различий в их интенсивности. Распределение значений признака G на анализируемом участке речевого сигнала надёжно характеризует смычку, свойственную этим звукам. Для остальных фонем из классов Ω_1 и Ω_2 распределение признака G имеет совсем иной характер.

Таким образом, по результатам предварительной фонетической идентификации удалось надёжно разделить шумовые согласные на следующие группы

$$\left.\begin{array}{l}(\omega^{с}, \omega^{з}) \in \Omega_1; \quad (\omega^{ш}, \omega^{ж}) \in \Omega_2; \\ (\omega^{ф}, \omega^{х'}) \in \Omega_3; (\omega^{ч}) \in \Omega_4; (\omega^{2}) \in \Omega_5\end{array}\right\} \quad (3)$$

Методы временного анализа речевого сигнала позволяют с высокой надёжностью выполнять сегментацию и предварительную фонетическую идентификацию шумовых согласных в потоке речи. Но информации, получаемой на этом уровне обработки недостаточно для разделения фонемных пар внутри выделенных групп. Для решения этой задачи необходимо осуществить анализ более тонкой структуры анализируемых речевых сигналов. С этой целью были использованы средства спектрального анализа, позволяющие исследовать микроструктуру фонетических пар звонкий-глухой и определить их отличительные признаки.

Известно /2/, что в образовании звонких шумовых согласных принимает участие фонация. Поэтому такие звуки характеризуются наличием в их спектре составляющих основного тона голоса (СТГ). Из ранних работ Варшавского /3/ следует, что частота 400 Гц может быть принята как абсолютная верхняя граница наличия основного тона (ОТ) в естественно произнесенном речевом сигнале. С учётом этого для анализа был использован полосовой фильтр с диапазоном 80-400 Гц для выделения ОТ. Спектральному анализу будем подвергать интервал протяжённостью 40 мс. Этот интервал включал акустически однородную область и один из примыкающих к ней сегментов, менее других отличающийся от неё по своим характеристикам. Таким образом, на анализируемом участке будет содержаться не менее четырёх периодов ОТ.

Прежде чем речевой сигнал поступит на спектральный анализатор он подвергался нормированию по динамическому диапазону. Правило, позволяющее отделять звонкие фонемы "З" и "Ж" от их глухих аналогов "С" и "Ш", использует значения двух признаков - энергию сигнала в канале 80-400 Гц и признак периодичности. Известно /4/, что процесс восприятия СТГ человеком представляет собой определение периодичности огибающих во всех или некоторых каналах слухового анализатора.

При дальнейшем анализе за меру периодичности принимается степень близости различных замеров периодов ОТ. На участках звонких согласных соседние замеры ОТ очень близки. На глухих согласных - замеры периодов ОТ случайны и сильно отличаются друг от друга.

Для определения значений периодов ОТ был использован анализ микроструктуры шумовых фонем в частотном канале 80-400 Гц. С этой целью речевой сигнал дискретизировался с учётом эффекта сглаживания в слухе /1/. Речевой сигнал, представленный в бинарном виде, является весьма удобным объектом исследования прежде всего в том смысле, что единственным информативным элементом в данном случае является только временной интервал между смежными единицами. Эксперименты показали, что всё многообразие комбинаций нулей и единиц, определяющих структуру временных отношений в пространстве выборки, можно свести к конечному и при том довольно небольшому числу основных или базовых блоков по критерию регулярности последовательности двоичных элементов. Таким образом, были выявлены регулярные стационарные блоки единиц Y (длительностью K), регулярные блоки нулей N (длительностью m) и нерегулярные нестационарные блоки X (произвольные комбинации двоичных элементов).

Для упрощения анализа двоичного представления сигнала бинарный код подвергался двойному медианному сглаживанию с различными окнами (3 и 5 шагов).

Исследование наиболее общих закономерностей структурной организации периодов ОТГ позволили выявить стабильные комбинации блоков

$$\left.\begin{array}{l} Y\ N_1\ Y'(X)\ N_2 \\ Y\ X_1 N_1 Y'(X_2) N_2 \\ Y\ X_1 N_1 X_2 \\ Y X N \end{array}\right\} \qquad (4).$$

Проведенные исследования показали, что период ОТ определяется по наличию в бинарном коде конкретной комбинации (4) и допустимых численных значений входящих в неё блоков. Для определённой реализации звонких шумовых согласных структурная организация периодов ОТ имеет довольно стабильный характер. Для глухих же согласных рядом стоящие периоды могут иметь любую из организаций (4).

Таким образом, по описанному методу на интервале анализа предъявляемого для распознавания звука определяется четыре периода ОТ, которые сравниваются между собой по некоторому критерию. Если различия между периодами ОТ меньше некоторого порога R, то это звонкий звук, если больше - то глухой.

Предложенный метод оценки ОТ обладает тем преимуществом, что позволяет идентифицировать глухой шумовой согласный не только по соотношению длительностей нескольких периодов ОТ, но и по специфике структурных комбинаций базовых блоков, составляющих ряд последовательных периодов ОТ. Большая изменчивость таких структур на интервале анализа является характерной для глухих шумовых согласных, что не свойственно звонким фонемам.

Для разрешения возникающих неопределённостей используется значение энергии сигнала в той же полосе частот 80-400 Гц.

Шумовые согласные "Ф" и "Х" из класса Ω_3 (3) обладают перекрывающимися в широком диапазоне спектрально-временными характеристиками и поэтому различаются между собой с очень низкой надёжностью (около 64%).

В заключение следует отметить, что предложенные в данном докладе методы сегментации и фонетической идентификации шумовых согласных обладают высокой помехоустойчивостью по отношению к аддитивным и локальным помехам. По результатам контрольной проверки для 9 дикторов (6 мужчин и 3 женщины) обеспечивается надёжность распознавания 98,2%.

Список литературы

[1] Абрамов О.М., Дрюченко А.Я., Усенко С.А., Шабанов-Кушнаренко Ю.П. Эффект сглаживания в слухе. - В сб. Проблемы бионики, Харьков, 1977, вып.19, с.31-37.

[2] Сапожков М.А., Михайлов В.Г. Вокодерн я связь. - М.: Радио и связь, 1983. - 248с.

[3] Варшавский Л.А., Литвак И.М. Исследование формантного состава и некоторых других физических характеристик звуков русской речи. - Проблемы физиологической акустики, М.-Л.: изд-во АН СССР, 1955, вып.3, с.5-17.

[4] Фант Г. Акустическая теория речеобразования. - М.: Наука, 1964.-284с.

DOUBLETS IN ARABIC: NOTES TOWARDS A DIACHRONIC PHONOLOGICAL STUDY

RADWAN S. MAHADIN

English Department
Yarmouk University
Irbid, Jordan

ABSTRACT

This article examines doublets in Arabic, discussing the alternations between the determinants in the doublets. Moreover, it shows that the alternations are the result of phonological changes. The directions of the phonological changes are suggested.
The conclusion shows that the phonological changes are in agreement with the changes that have occurred in other Semitic languages and in modern Arabic dialects. Moreover, it shows that the Classical Arabic is a mixture of different pre-Islamic dialects and that modern dialects are an extension of the old Arabic dialects.
Finally, the article shows the advantage of using the present to explain the past.

This article examines doublets in Arabic, discussing the alternations between the determinants in the doublets. Moreover, it shows the alternations are the result of phonological changes, the specific process being dialect borrowing [5]. Doublets can be definqd as two or more words in the same language, deriving from the same source and having similar meanings. They are phonemically identical except for one sound; that is, the stems share all sounds except one consonant. The differing consonant, called the determinant, could be any of the consonants in the root, but in a particular doublet, the determinant would have to appear in the same position, i.e. initial, medial, or final. The following examples illustrate the point:

1. (bakka) -- (makka) "Mecca"
2. (qunfuð) -- (qunfud) "hedgehog"
3. (ṭuxrūr) -- (ṭuħrūr) "thin cloud"

In these examples three alternations of the determinants occur.

1. (b) -- (m) The alternation is in the initial position. The determinants differ in the manner of articulation: (b) is an oral stop; (m) is a nasal stop.
2. (ð) -- (d) The alternation is in the final position. The determinants differ in both the place and manner of articulation: (ð) is an interdental fricative; (d) is a dental stop.
3. (X) -- (ħ) The alternation is in the medial position. The determinants differ in the place of articulation: (X) is uvular; (ħ) is pharyngeal.

Is it possible to determine the direction of the phonological change? To answer such a question, we need to examine the source of the doublets, for the source allows us the possibility of determining the direction of phonological change.
If the two forms of a certain doublet were found in two different areas, tribes, or dialects, then we may say that the two forms were introduced into Classical Arabic by the Arab grammarians when they tried to systematize the Arabic language [8]. In other words, the Classical Arabic is a koine [3]. On the basis of this assumption, it is not easy to trace the development of the two forms, especially the phonological changes.
On the other hand, we could assume that we have one form which was borrowed by another dialect or speaker and which then underwent the change. Or we could assume that the change happened within the same dialect due to the ease of articulation or to children's mistakes in language learning or to some other causes of linguistic change [11].
For example, the deletion of a glottal stop in Arabic is a very common process (ra? s → rās "head"), and it is also common in all Semitic languages [1]. If this last process is the case, we could trace the change with a reasonable certainty.
However, without denying the possibility of any or all of these processes occurring, my judgment of the direction of the change

is based on the following assumptions:

1. The Semitic languages have a tendency to change in the same direction. For example, dental fricatives have become either stops or alveolar fricatives in most Semitic languages [9]. Consequently, one could assume that similar changes could occur in Arabic.

2. Certain changes are more natural than others. For example, it is more natural for fricatives to become stops than for stops to become fricatives. For example, fricatives emerge after stops in child language [10]. Voiceless obstruents are more natural than voiced obstruents, and consonants with emphatic articulation tend toward plain articulation.

3. Certain assumptions could be deduced by comparing some forms in Modern Standard Arabic with other forms in Modern Colloquial Arabic and, at the same time, with forms which existed in old dialects and in Classical Arabic before Islam. In other words, we can compare forms before Islam with forms after Islam. This can be done by examining written records, i.e. the scattered writings of the Arab grammarians in which they attributed certain forms for certain tribes or dialects. By weighing which dialect was more prestigious or which form was more common (the relative usage of the forms), we can determine the old form. However, this last method is not always reliable because many of these items were reported on the authority of a number of grammarians and transmitters of poetry. However, we must be very cautious in our use of such testimony, for pre-Islamic literature is mainly poetry and is rather suspect in its authenticity [6].

4. By comparing the distribution of relative frequency of consonants, we can, with the help of other criteria, decide that the sound with a high frequency is more stable and less susceptible to change than the one used less frequently. The following are some representative examples of doublets in Arabic:

[b-m]	:	[?abada]-[?amada] "to linger"
[b-f]	:	[dabba]-[daffa] "to walk slowly"
[ð̌-d]	:	[dʒað̌ð̌afa]-[dʒaddafa] "to row"
[ð̌-z]	:	[ḥāð̌a]-[ḥāza] "to own"
[θ-t]	:	[raθama]-[ratama] "to utter"
[θ-s]	:	[maraθa]-[marasa] "to macerate"
[θ-š]	:	[nabīθa]-[nabīša] "Soil dugout from the earth"
[s-š]	:	[faqasa]-[faqaša] "to break"
[l-n]	:	[?ismāgīl]-[?ismāgīn] "Ishmael"
[l-r]	:	[šarama]-[šalama] "to split"
[n-r]	:	[wakn]-[wakr] "nest"
[ɣ-x]	:	[ɣafīr]-[xafīr] "guard"
[ɣ-9]	:	[musawwa9]-[musawwaɣ] "permitted"
[x-ḥ]	:	[ṭuxrūr]-[ṭuḥrūr] "a thin cloud"
[9-ḥ]	:	[daba9a]-[dabaḥa] "to lower the head in walking"
[ṭ-d-t]	:	[maṭṭa]-[madda]-[matta] "to stretch"
[ṣ-z-s]	:	[baṣaqa]-[bazaqa]-[basaqa] "to spit"
[k-q]	:	[dakka]-[daqqa] "to crush"
[ð̣-ḍ]	:	[ð̣aby]-[ḍaby] "gazelle"

On the basis of phonetic similarities and the alterations between consonants in the doublets and on the basis of the assumptions discussed above, the following phonological changes can be suggested:

I. Plain (non-emphatic) consonants:

1. Labial consonants:

[m] [f] > [b]

2. Front consonants:

[ð̌] > [d] [z]

[θ] > [t] [s] [š]

[š] > [s]

3. Alveolar resonants:

[l] > [n] [r]

[r] > [n]

4. Back consonants:

[ɣ] > [x] [9]

[x] [9] > [ḥ]

II. Emphatic consonants:

[ṭ] > [t] [d]

[ṣ] [s] [z]

[q] > [k] [dʒ]

[ð̣] > [ḍ]

The conclusion shows that the phonological changes are in agreement with the changes that have occurred in other Semitic languages and in modern Arabic dialects [9, 2]. Moreover, it shows that the Classical Arabic is a mixture of different pre-Islamic dialects and that

modern dialects are an extension of the old Arabic dialects.

Finally, the article shows that the advantage of using the present to explain the past [7, 4].

REFERENCES

[1] Anīs, Ibrāhīm, <<Fi al-Lahjāt al-garabiyya>>, Cairo-Egypt, 1978.

[2] Cantineau, Jean, <<Course de phonetique Arabe>>, University of Tunisia-Tunisia, 1966.

[3] Ferguson, Charles, <<The Arabic Koine>>, Language (35), pp. 616-668, 1959.

[4] Greenberg, Joseph, <<Synchronic and Diachronic Universals in Phonology>>, Language (42), pp. 508-517, 1966.

[5] Hoengiswāld, Henry, <<Language Change and Linguistic Reconstruction>>, University of Chicago Press-Chicago, 1960.

[6] Hussein, Ṭaha, <<Fi al-Adab al-Jahili>>, Cairo-Egypt, 1952.

[7] Labov, William, <<On the Use of the Present to Explain the Past>>, Proceedings of the Eleventh International Congress of Linguistics, pp. 825-851, 1974.

[8] Mahadin, Radwan, <<The Morphophonemics of the Standard Arabic Tri-consonantal Verbs>>, Unpublished Ph.D. Dissertation, University of Pennsylvania, 1982.

[9] Moscati, Sabatino, et. al., <<An Introduction to the Comparative Grammar of the Semitic Languages>>, Otto Harrassawitz-Wiesbaden, 1969.

[10] Schane, Sanford, <<Generative Phonology<<, Englewood Cliffs-New Jersey, 1973.

[11] Weinreich, Uriel, et. al., <<Empirical Foundations for a Theory of Language Change>>, in <<Directions for Historical Linguistics>>, ed. by W.P. Lehmann and Yakov Malkiel, University of Texas Press-Austin, 1968.

ON RECONSTRUCTING PHONOLOGICAL SYSTEMS AND PATTERNS OF THEIR DEVELOPMENT

K.V. Gorshkova

Moscow State University

ABSTRACT

In reconstructing phonological systems of Common Slavic dialects two types must be posited: one with the DF of palatalization, the other with the enriched palatal series. Eastern Slavic dialects belonged to the second type. The correlation of palatalized and non-palatalized consonants emerged there as a result of the secondary palatalization of consonants. This development characterized the system on which the Rostov-Suzdal' dialect was based and constituted the starting point of later phonetic processes. An archaic system was preserved in the Old Novgorodian dialect and its descendants. Relative and absolute chronology of these changes and of connected processes is discussed and some general assumptions conserning the reconstruction of phonological systems and the evaluation of its results are tested in this framework.

"In reconstructing... the basic linguistic patterns that must, in principle, correspond to the system of the initial common language there arise certain methodological problems: the probability of these reconstructions, the degree to to which they conform to the system that actually existed in time and space and was the ancestor of a certain group of related diatects" (1). To resolve this methodological problem comparative linguistics must rely on linguistic typology and the study of language universals. The approach to more concrete questions of this kind depends on a number of details: on the level of linguistic structure to which the reconstructed and evaluated patterns belong; on how minutely this level is analyzed in synchronic descriptions of the related dialects under consideration; on the extent to which the historical sources are researched and worked out, etc. In this paper I shall try to evaluate verious reconstructions of the Common Slavic phonological system in its dialectal variants (particularly, East Slavic) and to reconsider the course of its development on the basis of the historical and descriptive dialectology of the Russian language.

There has been a considerable progress in the study of the phonological system of the contemporary Russian dialectal language in relation to other Slavic languages, in the analysis of Old Russian written sources that reveals the phonological structures of the oral language, in constructing models of the Common Slavic phonological system. Nevertheless, the following observation made by R.I.Avanesov retains its significance: "Our historical phonology, even in considering the whole, deals essentially with a mere sum of individual phonetic phenomena. On the contrary, it is necessary to consider individual linguistic phenomena as parts of the whole, of the entire phonetic system -- even in studying them individually. An adequate knouledge of an object may be obtained only by means of establishing correlation between general and particular" (2).

The evaluation of reconstructions -- generally accepted or resulting from recent studies -- is of particular importance in preparing university courses in comparative and historical linguistics. The character of proto-level recpnstructions predetermines our account of dynamic processes underlying historical changes of the phonological system. In Slavistics, these considerations influence courses on the Old Church Slavonic language, Slavic comparative grammar, and historical grammar of individual Slavic languages. It is particularly important to discuss general outlines of these courses at the present time because the results of such discussion will define the standards of the higher philological education in the near future.

In examining phonological reconstructions I proceed from the general assumption that phonemes as units of the phonological system are represented by sets of distinctive (DF) and integrative (redundant, IF) features which are realized in the constitutive features of phonetic units (sounds of speech). At the same time, sounds of speech contain additional positionally-conditioned features that are imposed in their realization upon the phonemic features.

Phonemic features combine in different sets according to universally significant patterns of compatibility ~~-- in~~ "vertical" order (1). Besides this "vertical" order of features that are distributed between two main types of feature combinations -- "marked" and "unmarked", there exists evidently a "horizontal" compatibility of features: some features of different phonemes freely combine in phonemic systems; combinations of other features are greatly restricted; still other combinations are impossible. "Horizontal" compatibility of phoneme-forming features is subject to syntagmatical regularities, whereas "vertical" compatibility is subject to paradigmatical regularities. The general type of a linguistic system depends on the relation between paradigmatic regularities of sound alternations and syntagmatic regulariries of sound combinations (3).

To assess the reconstructed phonological system of the later Common Slavic which is the basis of all historical Slavic languages one has to consider in the first place the results of Slavic palatalizations of consonants. "The importance of palatalization in Slavic is based on the tendency to raise the tongue in the palatal region. This is one of the main features of the Slavic basis of articulation" (4).

Phonetically, the results of the three palatalizations and of the palatalization through jotation are generally treated as a replenishment by new sounds of the palatal series represented before the palatalizations only by [j] (5). On the contrary, phonological inter pretations of these results greatly vary. A critical survey of different solutions that may be found in the literature is not the purpose of this paper.

One is justified, I think, in rejecting solutions based on the introduction of an additional (besides the phoneme) phonemic unit such as syllabeme, groupophoneme, etc. Theories that introduce these units deal with Proto-Slavic syllabic structure, they reflect the existence of intrasyllabic harmony but they have nothing to do with characteristics of the Proto-Slavic phonological system which is constituted (irrespective of concrete period) solely by means of phonemes as sets of relevant features. These theories are not in accord with the idea of the limited autonomy of the phonological system and its hierarchically structured relations with the morphological system, with the idea of interdependance of phonemes and morphemes. The morphemic structure of Proto-Slavic word-forms points to one phonemic unit only, the phoneme.

In considering the Later Common Slavic phonological system and its various developments it is especially important to determine the phonological character of palatal consonants, to realize what DF underlies their phonetic manifestation. During the Common Slavic period this domain of the phonological system presented considerable dialectal variation, primarily in respect of the quality of labials plus jod reflexes. The dialects of Common Slavic where palatalized labials emerged as a result of jotation (i.e. *bj, *pj, *mj, *vj changed into [b'], [p'], [m'], [v']) could form a timbre correlation of the type /t' - t/ already in the earliest possible period; in those dialects DF frontness-backness of the vocalic system lost its phonemic character and the sounds [i] and [y] were united in the phoneme /i/ as its positional realizations.

In those Common Slavic dialects where labials were not palatalized as a result of jotation but were supplied by an additional palatal element (i.e. *bj, *pj, *mj, *vj changed into [bl'], [pl'], [ml'], [vl']) the newly arised palatal consonants became the representatives of palatal phonemes. Eastern Slavic dialects belong to this second group. Thus, during this earlier period Eastern Slavic dialects of Common Slavic did not develop a timbre correlation of palatalized and non-palatalized consonants /t' - t/ but the palatal series (one of the series of consonants differentiated by the place of articulation DF) was greatly enriched. Differences between Northern and Southern East Slavic dialects concerned only the choice of palatal phonemes: in the northern areas where the dialect of ancient Novgorod and Pskov would later emerge [ǩ], [ǧ], [x̌] belonged to the palatal series because *k, *g, *x had not changed into *c', *z', *s' before *ě, *i of diphthongal origin.

An additional argument in favor of this solution may be seen in the syntagmatical properties of these phonemes. Palatal phonemes cannot combine with mid labialized vowels, and indeed combinations such as /to/, /te/, /tъ/, /tь/, /tǫ/, /tě/ but only /t'e/, /t'ь/, and (on morpheme boundaries) /t'ě/ may occur. At the same time, the syntagmatical behaviour of vowels suggests that in the phonological system under consideration oppositions between /e/ and /o/, /ь/ and /ъ/, as well as between /i/ and /y/ were based on the DF frontness-backness. The phonetic feature of labialization had no phonemic value, it functioned as a concomitant feature of frontness-backness DF. A vocalic system with this DF could not combine with a consonantal system including the DF of palatalization; on the other hand, it did not preclude the DF of the palatal place of articulation.

Palatalized labials [b'], [p'], [m'], [v'], as well as palatalized dentals [t'], [d'] might have originated in the phonetical system of Eastern Slavic dialects in the process of secondary palatalization of consonants, i.e. the palatalization of semipalatalized consonants before front vowels. Recent studies in Eastern Slavic dialectology support the view that the secondary palatalization of consonants was not a process common to all Eastern Slavic dialects. This process was consistently represented in those Eastern Slavic dialects that later formed part of the Rostov-Suzdal' dialect of Old Russian. The emergence of palatalized labials [b'], [p'], [m'], [v'] in Eastern Slavic dialects that underwent the secondary palatalization made possible the subsequent phonologization of these sounds which resulted in the establishment of timbre correlation between /p/ - /p'/, /b/ - /b'/, /m/ - /m'/, /v/ - /v'/. The outcome of these changes was a system with /t - t'/ correlation that could not preserve DF of "palatality". The consonants [s'], [z'], [l'], [n'], [r'] and [t'], [d'] found new places in binary oppositions with [s], [z], [l], [n], [r], [t], [d] forming correlative series of phonemes differentiated by DF palatalized -- non-palatalized: /s'/ - /s/, /z'/ - /z/, /l'/ - /l/, /n'/ - /n/, /r'/ - /r/, /t'/ - /t/, /d'/ - /d/.

The timbre correlation of palatalized and non-palatalized consonants began before the fall of jers. The fall of jers revealed new tendencies in the development of the phonological system and at the same time created new strong positions and new weak positions (positions of neutralization). Among the strong positions were word-final position (/kon/ - /kon'/) and position before consonants (/banka/ - /ban'ka/), among the weak positions were positions of assimilative palatalization or dispalatalization of consonants.

At the same time, evaluation of reconstructed phonemic systems involves considerations of diachronic character since reconstructed elements take part in the sequence of phonetic changes and phonological transformations.

From this point of view, it is important to decide on the relative chronology of the secondary palatalization of consonants and the loss of the nasal vowels (i.e. the change *ǫ →*u and *ę → *a) The secondary palatalization of consonants testify to an active intrasyllabic harmony, whereas the loss of nasal vowels demonstrates the considerable autonomy of vowels and consonants within the syllable; it is reasonable, therefore, to date secondary palatalization of consonants before the loss of nasal vowels.

This sequence of phonetic changes promoted the development of a phonological system with /t' - t/ since the change t˙ę → t'ę → t'ȧ created a strong position for /t' - t/ (position before /a/) and transformed the [a] - [ȧ] opposition into a positional variation. During approximately the same period, there emerged a strong position for labials and dentals /t/ - /t'/, /d/ - /d'/, namely, the position before /u/; this was the result of historical changes in *ī-stem noun declension.

As stated above, the hypothesis that the secondary palatalization of consonants gave rise to a special phonemic unit -- a syllabeme (together with the phoneme) -- is ungrounded. The morphemic structure of word-forms such as *plod-ъ, *kon'-ь, *gos't'-ь makes it impossible to posit syllabemes /dъ/, /n'ь/, /s't'ь/ as units of phonemic order in these cases.

The formation of the consonant system with DF "palatalized -- non-palatalized" precluded the existence of the opposition of front and back vowels in the same phonological system. Vowels differentiated by frontness became allophones of one and the same phoneme. At first, allophonic relations took shape for /u/ ([u]//[u̇]) and for /a/ ([a]//[ȧ]), then for /i/ (['i]//[y]) and for /o/ (['e]//[o]). In this system, reduced vowels [ь] and [ъ] do not merge but are vocalized in different ways: 'ь → 'e, ъ→ o.

The phonetic change t'et → t'ȯt and the analogical formation of syllables of the type t'o and t'ot' created the possibility of a new phonemic structure: a new vocalic subsystem (/e/ - /o/) with a new DF "labialized -- non-labialized" arose. This DF is easily combined with a consonantal DF "palatalized -- non-palatalized".

The emergence of the binary opposition of palatalized and non-palatalized consonants and of the vocalic system /i/ - /u/, /e/ - /o/ with DF "labialized -- non-labialezed" prepared the ground for the development of /ē/ into /e/ and of /ō/ into /o/, i.e. for the transformation of a seven-member vocalic system into a five-member one. This development characterized the Rostov-Suzdal' dialect which was the main component of the Center dialectal zone and later the source of the Standard Russian pronunciation norm (based on the central Moscow dialect) (7).

The phonological system of the dialects of ancient Novgorod and Pskov developed in another direction. The Old Novgorodian dialect has been studied in detail by A.A.Zaliznjak (8). In retracing its historical development it is important to take into account modern Northern Russian and particularly North-Eastern Russian peripheral dialects which still retain archaic phonetic features. The validity of the opposition of central and peripheral dialects has been clearly demonstrated by the study of the Russian linguistic landscape (9) Initially, no linguistic significance has been attached to the notion of "peripheral dialects". Later historical and dialectological research (10) provided evidence that phonological systems of the peripheral dialects are of the vocalic type and differ in this respect from the central dialect which is consonantal. The system of the peripheral dialects is characterized by a seven-member vowel set, by the possibility of the vocalic DF "frontness - backness", by the presence of palatal consonants, by the opposition (in certain dialects) of tense and lax consonants that makes the voiced -- voiceless distinction a redundant phonetic feature, by vocalicity of sonorants, by the predominance of vocalic elements in phonetic sequences (11, 12).

Studies of the historical development of phonological systems demonstrate their relative stability. Apart from the fact that constant elements prevail over changing ones at every stage of development of a phonological system, it should be noted that a single phonological system can serve as a means of expression for semantic units for many generations of people that use the respective language or dialect in their communication. One may assume that important changes in phonological systems occur as a result of the interaction between dialects or languages.

History of phonological systems as well as their reconstructions deals with the notion of relative chronology (linguistic time proper). A language historian can raise the problem of absolute chronology only after establishing a relative one. This transition from relative to absolute chronology requires special methods and in some cases cannot be carried out. At the same time, rearrangements of relative chronology necessitate respective rearrangements of absolute one. Thus, if one dates the loss of nasal vowels before the secondary palatalization of consonants the latter process may be dated from the 10th or 11th century; if, however, the relative chronology of these two processes is reversed the secondary palatalization of consonants must be posited before the 9th century.

In the Rostov-Suzdal' dialect, the change of e into o ceased to operate at the turn of the 16th century. This event may be dated on the basis of depalatalization of the affricate [c'] since e does not change into o before this phoneme; at the same time, the depalatalization of [c'] is clearly reflected in datable written sources. The transformation of a seven-member vowel system into a five-member one presupposes as a necessary condition the suspension of the e → o change and the subsequent phonologization of /e - o/ distinction; it is reasonable therefore to attribute this transformation to the late 16th or the early 17th century. The reconstruction of dialectal phonological systems and corresponding chrono-topoisoglosses as well as their evaluation applies both to the pre-literate and later periods.

References

I. Гамкрелидзе Т.В., Иванов Вяч.Вс. Индоевропейский язык и индоевропейцы. I. Введение. Тбилиси, I984.

2. Аванесов Р.И. Русская литературная и диалектная фонетика. М., I974.
3. Панов М.В. О двух типах фонетических систем. - В кн.: Проблемы лингвистической типологии и структуры языка. Л., I977.
4. Мейе А. Общеславянский язык. М., I95I.
5. Бернштейн С.Б. Очерк сравнительной грамматики славянских языков. М., I96I
6. Komaček M. Historicka mluvnice česká. Praha, 1958.
7. Горшкова К.В., Хабургаев Г.А. Историческая грамматика русского языка. М., I98I.
8. Зализняк А.А. Новгородские берестяные грамоты с лингвистической точки зрения. - В кн.: Янин В.Л., Зализняк А.А. Новгородские грамоты на бересте. М., I986.
9. Захарова К.Ф., Орлова В.Г. Диалектное членение русского языка. М., I970.
IO. Горшкова К.В. Историческая диалектология русского языка. М., I972.
II. Касаткин Л.Л. Русский диалектный консонантизм как источник истории русского языка. М., I984.
I2. Бромлей С.В. Различия в степени вокализованности сонорных и их роль в противопоставлении центральных и периферийных говоров. - В кн.: Диалектография русского языка. М., I985.

ON THE ORIGIN OF MORPHONEMICIZED ACCENT SYSTEMS

VLADIMIR A. DYBO

The Institute of Slavistics and Balkanistics
of the USSR Academy of Sciences

According to the type of accent languages are usually divided into two cathegories: those possessing free accent and those possessing fixed accent. In languages of the latter cathegory the position of accent is conditioned by the phonetic structure of the word.

The accent is being fixed in various manners, most of which, however, are variations of one and the same basic principle - syllable counting. In the simplest case stress falls on a certain syllable counted from the word's beginning or end. All other ways of fixing the accent are apparently being derived from the basic principle by means of introducing into the fixation rule of some additional factor - structure, quantity of the syllable (by quality we mean both the segmental and the suprasegmental syllable quality - i. e. tone).

Thus, if we take into account the manner of accent fixation and rules determining the position of accent in accent-fixed languages, then following subdivisions may be distinguished:

1) Languages where the position of accent is determined by counting the number of syllables (syllable-counting languages), e. g. Polish, with its stress on the penultimate syllable, or Czech, with stress falling on the word's first syllable.

2) Languages where the position of accent is determined both by the syllable number and the syllable quantity (short vowel = 1 mora, long vowel = 2 morae) (mora-counting languages), e. g. Latin with its stress on the penultimate syllable if it contains 2 morae, but on the pre-penultimate if the penultimate contains 1 mora.

3) Languages where syllable structure is significant as well, e. g. having stress on the ultimate syllable if it is closed, but on the penultimate if it is open - the suggested (Chr. Sarauw) proto-Semitic system reflected (although with some deviations) in the old Arabic grammar by Pedro de Alcala; this basic rule (with some minor complications) is apparently preserved in Maghrib dialects of Arabic.

4) Languages where the position of accent is also determined by vowel quality, e. g. Moksha; similar systems probably underlie the Mari and Permic morphonemicized accent.

5) Languages where the position of accent is determined both by syllable number and the prosodic syllable features: East Saharan - Tubu, Kanuri (at least its Badavi dialect), Kanembu (stress falls on the first high-tonic syllable; the same rule functions in Hausa, Yoruba and Bambara); on the East Saharan accent and tonal systems see V.A.Dybo, The prosodic system of Tubu (Teda-Kanuri group) - a beginning of changing the tonal system to a system of paradigmatic accent? - in: Afrikanskoye istoricheskoye yazykoznaniye, Moscow, 1987; on stress in Yoruba and Bambara see I.Herms, Ton und Intensität im Yoruba - Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, Bd. 35, Hf. 2, 150-156 (1982); S.Brauner, Zur grammatischen Funktion prosodischer Merkmale im Bambara - ibd., 144-149. A more complicated rule is noted in Usarufa (New Guinea): accent is placed in the word's end if the word contains only low-tonic syllables, but in the beginning of a high-tonic syllable sequence (accordingly, on a single high-tonic syllable) if the word contains only (or at least some) high-tonic syllables; however, if a word beginning with a sequence of high-tonic syllables has a low-tonic end, then accent is placed on the last high-tonic syllable (see "Studies in New Guinea linguistics", Sidney, 1962, p. 114, see also the list of data on p. 115-127).

All cases when we know the prehistory of a given free accent system or of a morphonemicized accent system, or when sufficiently persuasive comparative evidence is available, reveal that systems of that kind develop from systems with fixed accent added by a factor deforming the fixation rule (i. e. from types 2, 3, 4, 5). It happens after the loss of the phonemic contrast, on which the specifical manner of accent fixation is based, this contrast being replaced by a contrast of accent positions - i. e. under conditions of neutralising phonemic oppositions which had earlier constituted the factor def-rming the basic rule (in other words, the distinctive role played earlier by segmental or suprasegmental components of a syllable, is being transferred to the accent contour of the whole word). Thus, free-local accent in some Romance languages is a result of quantity merger and therefore, of cancelling the contour rule based on mora counting. The surviving accent distinctions are losing their mora motivation and become phonemic. A similar free-local accent in the Yazva dialect of Komi has resulted from phonemicizing of accent contours which were previously motivated by vowel quality distinctions (see the works of V.I.Lytkin).

Free accent in Balto-Slavic and Abkhazo-Ubykh must be explained by phonemicizing the accent contours after the loss of tones which had earlier motivated those contours (I have given detailed arguments in favour of this proposal in a series of papers under a common title "The Balto-Slavic accent system from a typological point of view and the problem of reconstructing Indo-European accent" - in: "Balto-slavyanskiye etnoyazykovyye kontakty", Moscow, 1980 (publication incomplete); for a short account see V.A.Dybo, "The Balto-Slavic accent system from a typological point of view and the problem of reconstructing the Indo-European accent", in "Kuznetsovskiye chteniya 1973", Moscow, 1973, pp. 8-10; V.A.Dybo, The West Caucasian accent system and the problem of its origin. - "Konferentsiya "Nostraticheskiye yazyki i nostraticheskoye yazykoznaniye. Tezisy dokladov", Moscow, 1977, pp. 41-45; V.A.Dybo, "The tonological hypothesis on the origin of Indo-European accent systems" - in "Konferentsiya "Problemy rekonstruktsii", 23-25 October 1978, pp. 56-61; V.Dybo, S.Nikolayev, S.Starostin, A tonological hypothesis on the origin of paradigmatic accent systems, - in: Estonian papers in phonetics, Tallinn, 1978, pp. 16-20; see also S.L.Nikolayev, The Balto-Slavic accent system and its Indo-European sources, Moscow, 1986.

Depending on the relation of factors determining the stabilisation of accent to morphemic word-boundaries, two extreme types of morphonemicized accent systems can be distinguished: 1) if the morphonemicization results in lexical distribution of accent types (in case the accent-stabilising factor had been primarily connected with the root morpheme), there arises the so-called "paradigmatic accent", 2) if the morphonemicization results in a distribution of accent types among morphological forms and cathegories, one deals with the "cathegorial accent". The second result is being immediately obtained after morphonemicizing a fixed accent of the Latin kind. Here the reasons are obvious: the accent contour obtains the distinctive functions inherent in the penultimate vowel of the word (often being the suffix or ending vowel); further phonetic processes can simplify the situation (as in French, where the reduction of non-stressed final vowels had demolished the distinctive nature of the accent) or else make it somewhat more complicated by creating the word-final position for stress, parallel with its positions on the 2d and 3d syllables, etc.

In such cases we usually find a most frequent accent-type, which is common for the vast majority of words - it may be labelled as the "unmarked" (trivial) accent type - and several accent types characteristic for certain suffixed structures, cer tain cathegories and word-forms. This distribution, of course, can never be strict enough, since the non-derived (synchronically) lexemes include also a number of words having a non-trivial accent. In such cases we usually get following descriptions: one determines the accent types, defines the trivial (unmarked) type and states the connection of non-trivial types with certain cathegories of words and word-forms; then one gives the list of exceptions. Within the framework of generative morphonemics one usually reconstructs (on the so-called deep-structure level, with certain - simplifying or complicating - approximations) the primary rule of accent stabilisation, introduces certain marks denoting reconstructed positions of the rule's application, and then formulates rules translating the "deep" structures to the surface level (this whole system of rules is a "synchronisation" of diachronical processes which had brought the language's archaic system to its modern state). Whether the language, in its synchronic state, does really possess such a mechanism, is a separate problem which has to be explored, and its solution may vary depending on different kinds of generative models.

More complicated systems arise after morphonemicizing of the 5th type of fixed accent (i. e. the accent connected with prosodic, tonal features, of syllables). The character of such systems is determined to a large extent by the character of the tonal systems which they "reflect". Among tonal systems we find systems with extremely developed so-called grammatical tone (e. g., in Hausa). It is natural to expect that after the morphonemicization of accent such tonal systems change to systems with cathegorial accent.

When the accent of tonal systems with prevailing "lexical" tone is being morphonemicized, there appear systems with "paradigmatic" accent, often possessing rather complicated accent types (accent paradigms), such as Slavonic (Russian, Serbian, Slovene etc.), Baltic (Lithuanian), West Caucasian (Abkhaz, Abaza, Ubykh) accent systems. The usual method of their description is determining the number of accent types within every lexico-morphological word class and its particular subclasses, and describing the behaviour of word-accent within each of these accent types ("the accent curve").

When there exists a possibility, the (complementary) distribution of accent curves depending on morphological subdivision is established. In this case several accent curves can be united (on basis of certain characteristic features) within one class of complementary distributed accent curves. Such a class (or a cathegory characterised by this class) is usually called an "accent paradigm". Sometimes the notion of accent paradigm and accent curve are being confused; but we should stress that in different word-cathegories one and the same accent paradigm can be manifested by different accent curves. Then one determines the content of each accent type, i e. enumerates all lexemes accentuated according to a given accent pattern, and specifies the accent types of derived structures which may be either connected with the accentuation of underlying non-derived morphemes and with the morphonemic type of the affix, or else may depend only on the character of the affix. One outlines the parts of the system where exceptions from the paradigmatic principle of accent types distribution and a developing cathegorial accent can be observed. Finally one describes various transformations of accent curves within syntactic units (transfer of accent to enclitics, proclitics etc.)

Recent decades have shown that systems of this kind can be rather efficiently described by means of generative morphonemics, and such descriptions help discovering many obscure (from the "surface" point of view) relations. As an example of this kind of a discription, essentially structured along the lines of my internal reconstruction of the Balto-Slavic system of accent valencies, may be quoted the study of the Lithuanian accent system in the new Lithuanian Grammar (Vilnius, 1985, pp. 61-68; section written by A.Girdenis). A sufficiently complete description of the Russian accent system has been achieved by A.A.Zaliznyak ("Russian nominal word-formation", Moscow, 1967; "From the Proto-Slavic accentuation towards the Russian", Moscow, 1985). A "generative" description of the Abkhaz accent system based on the above principles was proposed in my paper "The typology and the reconstruction of paradigmatic accent systems" (in "Aktsentologiya i sravnitelno-istoricheskiy metod", in print).

Languages with "paradigmatic accent" in many cases reveal a tendency of shifting towards "cathegorial accent" by means of generalisation within certain lexical cathegories and especially within derivational types of some accent paradigms. Examples of different progressive stages of this tendency are the Russian, the Lithuanian and the Pushtu accent systems.

Results presented in this paper can probably serve as a basis for distinguishing between languages with morphonemicized accent systems and languages with tonal word-contours and tonal paradigms (patterns) which are possibly results of immediate development - and not of a "representation" - of systems with "classic" tones.

ФОНЕТИЧЕСКИЕ ДОЛГОТЫ ГЛАСНЫХ В РУССКОМ ЯЗЫКЕ: ФОНОЛОГИЧЕСКАЯ СУЩНОСТЬ И ИНТЕРПРЕТАЦИЯ СВЯЗАННЫХ С НИМИ ЯВЛЕНИЙ (ДИАХРОНИЧЕСКИЙ И СИНХРОНИЧЕСКИЙ АСПЕКТЫ)

АЛЕКСАНДР ПЕНЬКОВСКИЙ

Фак-т русского языка и литературы
Владимирского государственного педагогического института
Владимир, СССР, 600000

РЕЗЮМЕ

В докладе предлагается фонологическая интерпретация трех типов фонетических долгот гласных в качестве ключа для единого объяснения процессов контракции гласных в зияниях в истории русского языка, в живой диалектной и разговорной речи, а также широкого круга явлений междометной сферы, аффективного говорения, экстранормальной фонетики и фонетики певческой и стихотворной речи. Выдвигается предположение о существовании в русском языке периферийной реликтовой просодии счета по морам.

0.0. Существует два типа долгих ударных гласных и соответственно два типа фонетической долготы:

0.I. *n*-морная долгота с ударением на первой море: [а́̄] = [а́а], [о́̄] = [о́о], [э́̄] = [э́э], [и́̄] = [и́и], [у́̄] = [у́у] и т.п.

0.2. *n*-морная долгота с ударением на последней море: [а̄́] = [аа́], [о̄́] = [оо́], [э̄́] = [ээ́], [ӣ́] = [ии́], [ӯ́] = [уу́] и т.п.

0.3. Число мор реально не превышает семи, но как правило равно двум или трем. В докладе будут рассмотрены преимущественно двухморные единицы.

I.0. В системе с экспираторным ударением долгота с ударением на первой море может рассматриваться как аналог нисходящего ударения, тогда как долгота с ударением на последней (второй) море оказывается соответственно аналогом восходящего ударения. С фонетической точки зрения эти два типа долгот организованы по принципу зеркальной симметрии и должны быть признаны логически равноправными.

2.0. Фонологически же они неравноправны.

2.I. Первая – монофонемна или трактуется как монофонемная и тогда подвергается сокращению или контракции.

2.2. Вторая – *n*-фонемна (двухфонемна) и поэтому неспособна к сокращению, но зато допускает развитие внутренних интервокальных консонантных элементов.

3.0. Вполне последовательно эти закономерности проявляются в парадигматических рядах русских междометий.

3.I. Так, на базе пяти гласных фонем русского языка (как показано в работе [I], они могут рассматриваться как sui generis "фонетические корни" русских междометий) образованы симметрично организованные междометные пары, обнаруживающие противопоставление двух указанных выше типов долгот: А [а́а] – [аа́] → [а^{h}а́] АГА; О [о́о] – [оо́] → [о^{h}о́] ОГО; Э [э́э] – [ээ́] → [э^{h}э́] ЭГЕ; У [у́у] – [уу́] → [у^{h}у́] УГУ; И [и́и] – [ии́] → [и^{j}и́] ИИ. Ср. также ЭЙ и ЭГЕЙ, ИХ и ИИХ и др. под.

3.2. Показательно, что почти все левые члены таких междометных пар, связанные с нисходящей долготой (при ударении на первой море), обнаруживают в размытом спектре их смыслов значение отрицания (или выражение отрицательной оценки), тогда как их восходяще-долготные противочлены (с ударением на второй море) находятся в смысловой зоне утверждения и положительной оценки.

3.3. То же в паре назальных междометий, специализированных на выражении отрицания и утверждения: [~́~] 'нет' и [~~́] 'да', неточно и непоследовательно передаваемых на письме написаниями мм, мгм и др.

3.4. Особо должны быть выделены междометия на базе редуцированного непереднего ряда, потерянные русскими грамматиками и словарями в связи с особенностями русской фонологической системы и русской графики. Ср. иное положение в болгарском. Таковы общеупотребительные разговорные "нисходящее" отрицательное [ъ́ъ] 'нет' и противопоставленное ему "восходящее" [ъ^{h}ъ́] 'да'.

3.5. Свидетельствуемая этими фактами связь нисходящей долготы с отрицанием, а восходящей долготы с утверждением находит подтверждение в изоморфной организации системы акцентов в трех типах русских повторов: а)С усиленным ударением на втором члене, выражающие усиление: Он до́брый-до̋брый 'очень добрый'; б)С усиленным ударением на первом члене, имеющие уступительно-отрицательное значение: До̋брый-до́брый, а какое зло сделал; Глӳп-глу́п, а понял; в)С равными по силе ударениями на каждом члене, имеющие функцию логического подчеркивания: Он до́брый, до́брый.

4.0. Аналогом последнего акцентного типа в системе фонетических долгот являются долгие гласные с равными ударениями на каждой море. В этом случае имеет место нейтрализация противопоставления двух основных типов долгот, а долгие гласные обнаруживают возможность двоякой фонологической интерпретации и двоякого фонетического развития. Такую ситуацию представляет певческая речь, там, где растяжение гласных (независимо от движения тона) осуществляется на равных длительностях.

4.I. В русской певческой фонетике (народной и классической) долгие гласныес равными по силе ударениями на каждой море трактуются как монофонемные единицы, что исключает развитие интервокальных консонантных вставок.

4.2. Иную традицию представляет народное пение западнорусских областей и соседних Украины и Белоруссии, где равноударные гласные в пролонгационной цепи последовательно разделяются гортанными консонантными сужениями. Так, в Западной Брянщине начало песни "На (о)городе верба рясна..." поется: На го-ро-де ве-рба р'а́-ħа́-ħа́-сна́, там сто-я-ла дев-ка кра́-ħа́-ħа́-ħа́-сна́.

4.3. То же характерно и для идущей из древности традиции церковного пения, где растяжение гласных, широко отражаемое древнерусскими кондакарями [2], сопровождается развитием интервокальных задненебных и гортанных фрикативных прокладок [3]. Ср. с этим характерную для архаической системы ц-сл. произношения общую тенденцию к консонантному разделению зияний [4].

4.4. Усиленная пролонгация гласных такого типа возможна как нарушение нормы и в некоторых экстремальных условиях разговорной речи, связанных со сверхвысокой степенью эмфатического напряжения. Таков, например, один из произносительных вариантов эмфатически напряженного отрицания нет. Ср. в литературном отражении: – О, он шутить не любит! Не-хе-хет!.. (И.С.Тургенев).

4.5. Явлениям разрядки равноударных пролонгационных квазизияний (см.4.2 – 4.4) могут быть сопоставлены явления образования и стяжения зияний в сходных условиях равноударности. Такая ситуация отмечена в архаическом слое некоторых с-в-р. говоров, где в определенных интонационно-синтаксических условиях обычная система соотношения ударных и безударных слогов уступает место другой, в которой все слоги одинаково ударны и, следовательно, безударны. Только в таких условиях могло возникнуть, например, обычное для владимирских говоров [гл'а:ла // гл'ала] 'глядела'.

5.0. К "плоским" равноударным долготам, т.е. долготам с ударением на каждой море, могут быть приравнены "плоские" же безударные долготы, которые также представляют условия нейтрализации противопоставления долгот нисходящего и восходящего типа и, следовательно, также должны иметь двойственные возможности фонологической интерпретации и дальнейшего фонетического развития. В русском языке такова ситуация в безударных зияниях гласных в своей и – особенно – в заимствованной лексике.

5.I. "Плоская" заударная долгота: индиви́д[уу]м, перпе́т[уу]м: а)Разрешение неопределенности по восходящему типу с развитием побочного ударения и гортанной смычки (индиви́д[уˀу̀]м, перпе́т[уˀу̀]м); б) Разрешение неопределенности по нисходящему типу (индиви́д[у]м, перпе́т[у]м).

5.2. "Плоская" предударная долгота: кооперативный – к[ʌʌ]перати́вный. Двоякое разрешение неопределенности при переводе в аббревиатуры: а)по восходящему типу – рыбкооп – рыбк[оˀо́]п; б)по нисходящему типу – рыбк[о́о]п ⟶ рыбк[о́]п.

6.0. На основе указанных выше закономерностей могут быть поняты и объяснены не получившие фонологической интерпретации и единого непротиворечивого истолкования явления возникновения и преобразования зияний, освоения зияний в заимствованиях, направления и результатов ассимиляции гласных в зияниях, явлений аферезы и элизии гласных и др. в истории русского языка, в живой диалектной и разговорной речи. В докладе в этой связи будут рассмотрены:

6.I. Вопросы истории стяженных форм имперфекта в древнерусском языке;

6.2. Варианты старорусских заимствований типа саадак 'лук и колчан со стрелами' – садак – сагадак 'то же'.

6.3. Варианты заимствованных собственных имен с зияниями гласных.

7.0. Особый интерес среди имен названной группы представляют вариантные пары типа Даниил – Данила. Вопреки [5], варианты типа Данила, Гаврила не могли возникнуть "в процессе освоения русским языком имен Даниил и Гавриил", поскольку, как ясно из 2.2. и сл., восходящие долготы не сокращаются, а восходящие зияния не подвергаются контракции. Необходимо, следовательно, постулировать на предшествующем этапе исторического развития вариантов таких имен с способной к сокращению нисходящей долготой типа Дани́ил, Гаври́ил, реальность которых как великорусских народных форм дониконовской эпохи подтверждена специальным исследованием [6].

8.0. Аналогично должны объясняться варианты и в парах типа Исак – Исаа́к.

9.0. Исходя из установленного выше положения о монофонемности нисходящей и полифонемности восходящей долготы, можно понять ряд важных явлений русской певческой фонетики и – среди них – запрет на восходящую пролонгацию гласных при переводе говорного слова в слово певческой речи. Можно петь иду́-у́ и иду́-у, гла-аза́ и глаза́-а, по́-о́ле и по́-оле и т.п., но запрещены иду-у́, глаза-а́, по-о́ле и т.п. Распевы такого рода воспринимаются на русский языковой и музыкальный слух как нарушение некой – до сих пор не формулировавшейся – нормы.

I0.0. На этой же основе могут быть поняты также закономерности, определяющие функционирование долгот и зияний указанных выше типов в русской стихотворной речи.

I0.I. Все слова и формы с восходящими долготами и восходящими зияниями трактуются русским стихом в отношении их фонемного состава и слоговой структуры,исходя из количества мор в долготе и количества гласных в зиянии:

а)Междометия и звкоподражания с восходящими долготами: "Лелеет лишь меня прекрасная пастушка"// –"А!а́! вертушка! // Не отвертишься ты..." (А.П.Сумароков); – А-а́!.. Какая кисть, какая сила!.. (Я.Полонский); О-о́х! Беда мне с стариком... (И.С.Никитин) – Приветим гостью дорогую / Чем бог послал. – И-и́! родной! //Приветь хоть лаской-то одной... (И.С.Никитин);– И-и́х, старик!Побойся бога!.. (И.С.Никитин); Сельди разом собралися, Сундучок тащить взялися, Только слышно и всего, Что у-у да о-о-о́ (П.Ершов); Плыву! – "У-у́..." – разносится Вдоль вспененной реки...(М.Луконин).

б)Иноязычные заимствования с восходящими зияниями: Ваа́л, Граа́ль, Маа́с, Саа́р и т.п., Воо́з, Воо́т и др. (двусложные), Исаа́к, Саа́ди, Трансваа́ль, Ханаа́н (трехсложные) и т.д. То же в собственной лексике на морфемных стыках: соо́бщник, поо́даль и т.п.

10.2. Слова и формы с плоскими долготами и безударными зияниями - в соответствие с 5.0 - 5.2 - характеризуются неопределенностью и получают двоякую фонемную и слоговую трактовку: Быть может, с каждой исцеленья Он из далеких стран спешил, Чтоб Иисус его мученья Всесильным словом облегчил (С.Надсон) - И под фальшивой брошкой в небе // Иисус Христос... (Е.Евтушенко); Да, вас судьба дарила щедро! Досель не тщетный звук для вас // Баярд, и Сид, и Сааведра ... (К.Павлова) - За правду, как доблестный рыцарь, Сервантес Сааведра встает (Н. Максимов); Россия ждет вестей из Саардама (М.Ткачев) - И в бедной мастерской Сардама сколачивал свой первый бот (К.Павлова). Сюда же широко распространенные в русской поэтической речи варианты вообще - вобще, воодушевление - водушевление, вооружиться - воружиться, сообщить - собщить и др.

10.3. Все нисходящие долготы и нисходящие зияния трактуются русским стихом только как монофонемные единицы и покрывают в стопе пространство равное одному слогу.

а)Поэтому для русского стиха Трансва́аль и Трансваль, Иса́ак и Исак, Мо́ор и Мор, Са́ади и Сади и др.под. - всегда лишь графические варианты: ...С недавних пор Он был и Фердинннд и Мор (М.Стахович) - Так вот зачем ребята Карла Моора Шли на титанов, шли на штурм веков (П.Антокольский); И Лейпциг - день железной славы, И Ватерло в резне кровавой (Н.Тихонов) - Когда нам время подошло, Мы по Толстому и Стендалю Увидели за дальней далью Бородино и Ватерлоо (П.Антокольский) и т.п.

б)Поэтому же нисходящие долготы междометий, как и нисходящие долготы гласных при растяжении в случаях эмфазы, окликах на расстоянии, командах и т.д. всегда монофонемны и односложны: И только там, в каменоломне, Он крикнул: "Ма-а-арш!" - И побледнел (И.Уткин); И звал: "Наза-а-ад, отступница, наза-ад!" (Н.Матвеева); И громкий голос паровоза //позвал в вагон: // - На целину-у-у! (А.Жаров); Во-он сидит рыбак-старатель, Не дает уснуть костру... (Г.Горбовский) и др.

То, что русский стих, всю свою историю отвоевывавший себе права на "поэтические вольности", остановился перед запретом на неодносложную трактовку нисходящих долгот, чрезвычайно показательно и важно.

II.0. Русский язык является слогосчитающим языком [7]. Однако нужно признать, что наряду с господствующей слоговой просодией существует подчиненная ей и испытывающая ее давление - реликтовая или рецессивная - просодия счета по морам. Имея центром сферу русских междометий, она распространяет свое влияние на сопредельные сферы аффективного говорения, певческой и поэтической речи, а также выходит в новейшую область аббревиации, где она обнаруживает себя, в частности, отменой редукции гласных. Все это заслуживает специального изучения (с учетом особенностей сетенциональной интонации, явлений примыкания согласных и др.)

[I] А.Пеньковский. Фонологическая интерпретация фонетических долгот гласных в русском языке (в связи с особенностями словообразования первичных междометий). - Проблемы теоретической и прикладной фонетики. - М., I973.

[2] Б.Успенский. Древнерусские кондакари как фонетический источник. - "Кузнецовские чтения - I970". - М., I970.

[3] Н.Успенский. Древнерусское певческое искусство. - М.: Музыка, I965.

[4] Б.Успенский. Архаическая система церковнославянского произношения.-М., I968.

[5] А.Суперанская. Ударение в собственных именах. - М.: Наука, I966.

[6] Б.Успенский. Из истории русских канонических имен. - М., I969.

[7] Н.Трубецкой. Основы фонологии.-М.I960.

PHONETIC CHANGE, PHONEMIC STATUS AND MORPHOPHONOLOGICAL ALTERNATIONS

TAMÁS SZENDE

HUNGARIAN ACADEMY OF SCIENCES
INSTITUTE OF LINGUISTIC SCIENCES

ABSTRACT

In this paper sources of morphophonological alternations are discussed. It is argued that, as attestable from Hungarian, co-occurrences of several types of reduction phenomena in casual speech and in their phonemic representations result in morphophonemic variants of the language.

The main aim of this research is to shed light on how parallel autonomous lexemic variants, such as Hungarian aztán and azután 'then or mondta and mondotta 'he/she/it said', come into being in the language.

Changes in the morphophonemic structure of a lexeme may led back, in a number of cases at least, to the fact that allegro rules/processes in the language use cause historically determined idiomorphic, i.e. primary, phonemic representations (original or first variant = v_1) to take the form of a phonetic representation that happens to correspond with a possible distinct phonemic structure (subderivative or second variant = v_2). This is sometimes the case even if allegro rules result in morphemic homonymy, see, for example, Hungarian vállat 'enterprise' (derived from vállalat 'id.' by [syllabic] deletion) as opposed to vállat 'shoulder + Acc.', whereby 'iconicity' and '(morphological) naturalness (cf. Dressler 1981) will apparently be damaged. (As for such a type of changes in view of goal conflict of better perception and better articulation--cf., for example, Lindblom 1983--, the hearer can only draw some poor consolation from Kiparsky's (1974) words: "language practises 'therapy' rather than prophylaxis".)

When seeking to reach a better understanding of the way of how and why subderivative, secondary variants (v_2) in (lexemic) morphology can, and in many cases also will, emerge, we are in essence faced with problems surrounding the correspondence of 'phonemic units' the morphemes are built up of, i.e. abstracta, with what is called the 'elementary speech events', i.e. pragmata. Some aspects of this relation turn out to be crucial in the explanation of the phenomenon discussed herein. In this line of thought the most important facts to be taken into consideration are as follows.

(i) The correspondence of 'phonemic units', on the one hand, and 'elementary speech events', on the other, does not always cover a one-to-one relation existing hypothetically between one phoneme size unit and one single articulatory or acoustic unit, cf. Stampe's (1980) divinity → [də̦v'ĩ̂ĩ] and the problems of 'biuniqueness', see also Hungarian bántsd [bã:d͡ʒD] ↔ /ba:ntʃd/ 'hurt [+ Acc.] him/her/it'.

(ii) Not only in casual but also in formal speech certain elements of phonetic representations corresponding to the respective elements in distinct underlying phonemic representations may coincide such as in [ɱ] in the phonetic representations of the Hungarian lexemes hamvas 'bloomy; downy' [hɔɱɔʃ] and honvéd 'Hungarian soldier' /honve:d/. What is more, some constituents of an underlying phonemic representation may occasionally remain undetermined even for the native speaker (cf., again, Hungarian [ka:ɱ̥or] possibly derivable from, either, /ka:nfor/ or, else, /ka:mfor/ on the basis of the rules:

nf → ɱ̥ / V___V and mf → ɱ̥ / V___V).

(iii) Also relatively homogenous articulatory/acoustic segments--having no direct reference to a phonemic constituent in the original underlying representation--may occur in speech, see, for example, French word-final schwa or t-epenthesis in German eigentlich (← eigen + lich). Although it is quite a rare thing in Hungarian that v$_2$--type variants come about by means of phonematization of non-etymological segments occurring in between of two phoneme realizations, individual cases of volksetymologie may be found, viz. t-epenthesis in szentfedél (← szemfedél) 'face-cloth'.

(iv) In what follows I shall give an overview of allegro phenomena which play an important part in bringing about new morphophonemic variants. With regard to their main categories, i.e. 'lenition' and 'fortition' (cf., among others, Dressler 1984), special attention should be paid to the types of lenition. As a matter of fact, there is no need of considering instances of fortition in this context, either, since morphophonemic variations can not be detected, at least in Hungarian, with the exception of occasional hypercorrect alternants of roots occurring under special communicative circumstances, such as színeművészet ← [si·nəmy·ve:set] ↔ /si:nmy:ve:set/. So, on the basis of a collection of allegro phenomena taken from Hungarian casual speech I here give an outline of a typology of what I call 'reduction'. (As for this typology, it is to be remarked here that, first, it does not cover all the allegro phenomena that occur in spontaneous speech and, secondly, individual items of the typology in question are not superimposed on each other in the hierachy of an ordered set but, rather, they occupy alternative points on a gradual scale.)

(iv/a) 'Weakening', generally speaking, means the lenis-production of a segment instead of its normative articulation as in lento. Weakening is carried out by means of deleting at least one, but not all, of the primary distinctive features the segment consists of, such as, for example, the distinctive feature [rounded] in [o] which latter will in this way be made to be [ɒ] or [ɔ] in instances like hogy 'that'. Secondary distinctive features may either be replaced by other concomitant features or else deleted completely. A second characteristic of weakening and a criteriaon of its delineation from 'loss', at the same time, is that the syllabic structure of the word remains always the same as determined in the underlying phonemic representation.

Weakening may also spread over a longer sequence of segments with an identical effect and content of its modifying various distinct members of the series. In other words: due to weakening, every single unit may show alternations of the same kind like, for example, loosening the closure in other stops occurring within the word boundaries.

(iv/b) 'Deletion' in a qualified sense is meant to be the dropping of a segment from the sequence such a way as to leaving behind traces in the articulation of the neighbouring segments, both left and right. Deletion, by definition, eliminates all the features that characterize the segment in question. As a consequence of deletion, syllable structure of the word changes to the extent that the number of segments the syllable is built up of changes to be less than the number of the segments in the v$_1$, i.e. lento, equivalent. This means that closed syllables may be converted into open ones. However, the number of syllables in the word remains unchanged.

As for the phonetic traces left behind after deletion, a normative, i.e. lento type--and often also hypercorrect--articulation of the two adjacent segments is to be observed as, for instance, in [ε·ʃø:] ← [εlʃø:] ↔ első 'first'. In contrast with weakening, deletion affects one unit within the sequence only. Whenever deletion happens to have been carried out, the remnants of the original variant of the word will undergo no further reduction. Accordingly, we find that tehát 'hence' [teha:t] changes to be [tea·t] as a result of h-deletion. Notwithstanding, [tea·t] coincides with the lento realization of the non-homonymous teát 'tea + Acc.', i.e. [tea:t] ↔ /tea:t/. (Note that duration of [a:] may in this place vary between the grades Ø, ·, and : without neutralizing the short/long opposition of /ɔ/ and /a:/.) In spite of the phonetic isomorphy of [tea·t] ← tehát + h-deletion and [tea·t] of teát, the former will never be pronounced with an intrusive [j]. So, we find that:

phonemic representation:	/teha:t/	/tea:t/
lento:	[teha:t]	[tea:t]
allegro:	[tea·t]	[te^{j}a·t]
	[teat etc.	[teja·t]

(It is to be remarked that tehát--as a whole--may in fact undergo further alternations and may become reduced to a one-syllable allegro realization, i.e. [ta·t]. This case of reduction of tehát, nevertheless, falls under the category of 'truncation'--see point (iv/d)---and is due to its specific outside-of-focus position within the sequence it enters into; some details see below.)

(iv/c) By 'loss' of a segment I mean the special class of elision by means of which a segment is dropped from the sequence without fortyfying adjacent segments and, in addition, necessarily modifying the syllabic structure of the word it affects. As opposed to deletion, in loss simplification of the syllabic structure also may take the form of deminishing the number of syllables. This happens whenever a vowel is subject to loss, see [sotsɔliʃtɔ] ← szocialista 'socialist'.

Although loss represents an extreme instance of reduction, phonetic traces may also be detected in the sequence. This means that in Hungarian allegro, among other things, loss of a vowel goes with the effect that the word maintains the former state of affairs with respect to vowel harmony until chain reduction rules begin to apply. Furthermore, fragments of the eliminated segment are optionally left over in the sequence in the form of what is labeled by Schane (1984) as 'cloning' and 'droning'. The independent phonemic status of eliminated segments is lost.

(iv/d) 'Truncation' is the reduction of a sequence, i.e. the deletion of the magnitude of sequence size units. In this type of reduction a word is reduced to the realization of at least one but not more than n - 2 segments the sequence is built up of in terms of a phonemic representation. The original syllabic structure becomes entirely destroyed due to the fact that more than one elementary constituents are eliminated. Accordingly, also phonetic traces can be found only occasionally in the remnants of the original sequence, apparently because of a mutual interaction of the constituents lost crossing each other's effect. Instead, also seemingly arbitrary articulatory/acoustic units may occur in the sequence which cannot be directly tracked back to the underlying phonemic representation of the truncated word, see [a] in [sa] ← szóval 'then, well' Here, too, the phonemic independent status of the constituents eliminated from the phonemic representation of the lento equivalent ceases to exist.

In case one conceives of weakening, deletion, loss, and truncation (iv/a--d) as allegro processes/rules functioning under specific conditions in casual speech, the question of rule ordering is brought up. I can here give some glimpses of the problem only. I should point out that in certain instances rule ordering may not be stated at all--such as in cases of co-occurrence of 'devoicing' and 'lengthening'--whereas in some others it may. For example, truncation in the strict sense may, and also often is, followed by reduction, i.e. change in the original state in terms of vowel harmony, see [s^{o}ɔ, so, sɔ, sa] → [sœ, sə, s^{ə}] derived from szóval in two respective steps.

(v) With regard to the present considerations, also one of the allegro phenomena of the next greater size unit of speech should be taken into account. Morphemes, or even complexes of morphemes, outside of focus position, and particularly when put in a phrase-initial or phrase-final position and, furthermore, when endowed with the communicative role of expressing the speaker's attitudes towards the message, the partner or himself (szóval 'well', tudniillik 'namely', végül is 'actually', etc.) are exposed to reduction to a greater extent than those being organic parts of the semantic structure of the text.

(vi) Finally, mention should be made of a trend of probability according to which the greater the number of the segments in the morpheme falling under the above mentioned category (see point(v)) the more likely reduction phenomena--as briefly discussed under point (iv)--will be carried out.

With all this types (processes, rules or tendencies) of allegro phenomena in mind one can conclude that new v$_2$ type morphemic variants come about if the reduced form may enter into the cycle of reduction, da capo al fine, as an independent unit, see mer --- mert 'because' which, in fortition, takes the form [mer^{ə}] like [e:r^{ə}] ← ér 'reach'.

Co-occurrences of (i), (iv/c--d) and (v), or (ii) and (iii), or (ii) and (v) in the language use --facilitated by (vi)--result in morphemic alternation and eventually changes in the mophophonological system of the language as attestable in Hungarian, see mié and mér ← miért, aztán ← azután, etc.

References

Dressler, W. (1981): Outlines of a model of morphophonology. Dressler, W., Pfeiffer, O. and Rennison, J. (eds.): Phonologica 1980. Innsbruck. 113-122.

Dressler, W. (1984): Explaining natural phonology. Phonological Yearbook 1, 29--50.

Kiparsky, P. (1974): On the evaluation measure. Bruck, A., Fox, R. and La Galy, M. (eds.): Natural Phonology. Chicago. 328-337.

Lindblom, B. (1983): On the teleological nature of speech processes. Speech Communication 2, 115-158. (Cited in Dressler 1984.)

Schane, S. (1984): The fundamentals of particle phonology. Phonological Yearbook 1, 129-155.

Stampe, D. (1980): A dissertation on natural phonology. New York.

THE ULTIMATE PHONOLOGICAL UNIT AS THE SMALLEST MORPHEME SHAPE

VULF Y. PLOTKIN

Dept. of English Philology
Pedagogical Institute
Tula, USSR 300026

ABSTRACT

The phoneme is divisible not only because it consists of ultimate constituents traditionally known as distinctive features, here termed kinakemes, but also because morpheme boundaries can run through phonemes. This is made possible by the ability of a kinakeme not only to participate in distinguishing morpheme shapes as a phoneme constituent, but also to provide such a shape by itself. Instances of inflexional and derivational affixes with shapes consisting of a single kinakeme are found in various languages, e.g. Estonian, Gaelic, Latvian, Nivkh (Gilyak), Romanian, Russian. A morpheme boundary can also run through a phoneme when an affix shape consists of a kinakeme cluster smaller or larger than a phoneme; the boundary then dissects the phoneme in question or its neighbour.

Ever since the notion of the phoneme as the basic unit in the sound system of language came into being, the problem of its (in)divisibility has always been present, though not always explicit in phonological theory. The insistence on the unquestionable absolute indivisibility of the phoneme, so characteristic of phonology's early days, soon gave way to the recognition of the existence within the phoneme of smaller truly ultimate constituents, named distinctive features /11, 272, 422-5; 12, 25/, merisms, phononemes, subphonemes etc. For reasons explained elsewhere /14, 82-3/ the best term is Baudouin's coinage 'kinakeme' /1, 199, 290/.

Recent research has demonstrated that these entities possess all the fundamental properties of language units:

(1) They are language-specific and cannot therefore be items in a universal inventory, Jakobsonian /11, 484-6/ or Chomskyan /8, 335/, any more than phonemes, syllables or words could be listed so /9, 152/.

(2) In each language they are paradigmatically united in a kinakemic system whose structure follows universal principles, but provides, like any other language system, a unique way of segmenting and organizing extralinguistic substance, which is not sound, linguistically organized by the phonemic system, but the speaker's cerebral activity in initiating sound and the listener's subsequent perceptive cerebration /14, 83 ff.; 15, 277-83/.

(3) Each language has its specific syntagmatic patterns for kinakemic combination in phonemes, which is basically non-linear simultaneous /15, 283-7/.

(4) Kinakemic systems play a leading role in the phonological evolution of languages and determine the direction of phonemic change /7/.

The establishment of the kinakeme as the ultimate language unit does not, however, take the question of phoneme (in)divisibility off the phonological agenda, for the problem has more than just one facet. An analogy may be appropriate here with the atom, whose very name reflects its indivisibility: despite its decomposition into a host of particles it remains the ultimate quantum of a chemical element and is indivisible on that level. Likewise, the phoneme is segmentable in certain aspects and indivisible in others.

The discovery of the phoneme in 20th century phonology was in a sense a rediscovery, for the original discovery dates back to the invention of alphabetic writing, when letters were created as symbols for phonemes. As long as the sound substance behind them was not analysed, they were treated as representing indivisible units of sound. The advent of phonetics in the 19th century put an end to the notion of integral sound units symbolized by letters and led to a two-pronged attack against them, pointing out the wide range of their variation and the complexity of their production and perception. The emergence of phonology was stimulated above all by the urgent necessity to uphold the notion of sound quanta and to protect them from being disintegrated in a continuum of variable phonic realizations. Hence the firmness with which the founders of phonology rejected every infringement on the principle of phoneme indivisibility.

The two questions concerning the unity of the phoneme with its variability, its articulatory and auditory complexity, have been answered differently by modern phonology. Allophonic variation has found its place in phonemic theory and no longer threatens phonemic unity. As for the inner complexity of phoneme structure, the discovery of the kinakeme as its constituent has of course shown the phoneme to be divisible into units of a lower level. But there is another aspect of phoneme structure which in its time attracted attention in connection with the problem of phoneme divisibility - the question of monophonemicity for sounds with temporally varying articulation, i.e. diphthongs and affricates. It must be stressed that kinakemic divisibility of the phoneme does not affect their monophonemicity if it is established by the well-known criteria of classical phonology which remain fully valid. Since kinakemic combination in a phoneme is non-linear and the kinakemes, clustered to form a phoneme, are activated more or less simultaneously, none of them can occupy a temporal segment of its own.

The kinakemic level is ultimately responsible for the conversion of sense into sound and the reconversion of sound into sense. Sound as the physical vehicle for the externalization of the speech signal is obviously so different from the cerebral activities with which the kinakemic system is concerned, that it has to be represented in the language system by a separate level. Deprived of its classical status of ultimate phonological (indeed, linguistic) unit, the phoneme retains its ultimateness on that level. Its properties including (in)divisibility are determined by the needs of the level it belongs to. Its kinakemic divisibility, far from being an obstacle to its function in organizing sound, is absolutely indispensable for the purpose. The phoneme is divisible in two other respects, and in both cases the divisibility is determined by the needs of other language levels without affecting the functioning of the phoneme. First, a phoneme can be crossed by a syllabic boundary. That is not possible for vowels and many languages do not permit it in consonants either. But languages that make use of the kinakeme of vowel checking may put the syllabic boundary after checked vowels inside a single consonant, between its onset and release, as in English 'body', 'supper'. Secondly, a phoneme can be crossed by a morpheme boundary, when one of the bordering morphemes uses a kinakeme and not a whole phoneme as its sole exponent.

The employment of kinakemes as exponents for morphological categories is observed in several languages. The analysis of protensity in Estonian provides abundant instances of the phenomenon. Estonian protensity with its traditionally recognized three levels (short, long, over-long) is now treated as resulting from a combination of two separate oppositions: long vs. short, over-long vs. long /cf. 16, 17/. The difficult problem of the exact phonic natures of both must be put aside in the present paper. What concerns us is the functional difference between the two: the opposition of long vs. short is mostly active in distinguishing lexical items (aasta 'year' - aste 'degree', mure 'grief' - murre 'dialect'), while the predominant function of the opposition of over-long vs. long lies in the grammatical sphere, where it is widely used to distinguish noun cases: part.sg. sāāli 'hall' - gen.sg. saali, gen sg. hōōne 'building' - nom.sg. hoone, part sg. lin̄na 'town' - gen.sg. linna, elat.sg. kal̄list 'dear' - part.sg. kallist. It is obvious that the kinakemes of the former opposition, which may be described as lexical protensity, do not possess semantic values of their own, whereas the kinakemes of the latter opposition of grammatical protensity serve as sole exponents of the categorial meanings of cases. That shows that the kinakeme of grammatical protensity or over-length is in itself a morpheme shape - a kind of case infix. As such it must occupy a certain fixed position in the stem shape and therefore must be able to join any phoneme in that position. The phoneme that incorporates the infix always follows the syllabic peak and may be a vowel; it is either different from the syllabic vowel and forms a diphthongal cluster with it (part.sg. laūlu 'song' - gen.sg. laulu, gen.sg. laīne 'wave' - nom. sg. laine), or identical with it and forms a bimoric monophthong with it, which becomes trimoric when the infix is added to it (see: saāli, hoōne above). The infix can also enter consonants, single (part. sg. seppa 'smith' - gen.sg. sepa, elat.sg. rikkast 'rich' - part.sg. rikast), in a cluster (part.sg. ok̄sa 'branch' - gen.sg. oksa) or geminated (see: lin̄na, kal̄list above).

In Irish Gaelic /2, 81/ categorial distinction of number and case in noun paradigms is regularly achieved by the kinakeme of palatalization: bád - báid 'boat', béal - béil 'mouth', bonn - boinn 'coin', cnoc - cnoic 'hill'.

In some German dialects /3, 384-5/ noun number is distinguished by certain kinakemes alone, e.g. voicing (pl. barɣ - sg. barç 'mountain', pl. brev - sg. bref 'letter'), vowel checking (pl. fiš - sg. fīš 'fish') or vowel fronting (pl. hünd - sg. hund 'dog').

Before their dat.sg. inflexions -m, -i Latvian nouns display a stem-final vowel /a/, /u/, /e/, /i/ (galdam 'table', tirgum 'market', zemei 'land', sirdij 'heart'). The sole exponent for the locative case is the kinakeme of protensity included into that vowel (galdā, tirgū, zemē, sirdī); for the accusative the sole exponent is

the kinakeme of tongue-raising added to the vowel (galdu, tirgu, zemi, sirdi) /10; 5, 232/.

The use of kinakemes as categorial exponents is widespread in Romanian, where it is found in the paradigms of nouns and verbs. The best-known example is the use of the palatalization kinakeme to mark the plural of nouns (pl. lupi 'wolf' - sg. lup) and the 2nd person in verbs (dormi 'sleepest' - 1st sg. dorm). But there are other instances as well. The four canonical forms expressing number and gender, e.g. sg.m. vecin 'neighbour', mîndru 'proud', pl.m. vecini, mîndri, sg.f. vecină, mîndră, pl.f. vecine, mîndre, demonstrate that kinakemic distinctions also exist between the two genders: both masculine forms end in high vowels /u/, /i/, regularly reduced to zero representations after single consonants (as in vecin, vecini), whereas both feminine forms end in non-high vowels /ə/, /e/; the masculine gender is thus marked by the positive kinakeme of tongue-raising, the feminine remains unmarked and usually includes its negative counterpart into its final vowel. But the feminine gender uses the positive kinakeme of tongue-lowering for its definite article: def.sg. casa 'house', cartea 'book', ziua 'day' with low vowels - /a/, front /ea/, back /ua ∼ oa/, absent from the indefinite forms (casă, carte, zi). In the 3rd person of many Romanian verbs indicative and conjunctive forms are distinguished only by the presence or absence of the positive kinakeme of fronting in the inflexional vowel (ind. bate - conj. bată 'beat', conj. poarte - ind. poartă 'carry'). Thus, the positive kinakemes of consonant palatalization, of tongue-raising and tongue-lowering can alone act as exponents for the grammatical categories of number, gender, definiteness in Romanian nouns, of person and mood in verbs.

The ability of single kinakemes to serve as grammatical exponents is not confined to morphological categories. They may also function as indicators of syntactic relations. For instance, in the Nivkh (Gilyak) language /4; 6/ syntactic subordination is marked by changing a modal kinakeme in the initial consonant of the headword. When a verb takes an object or a noun takes an attribute, it reflects its syntactic domination by a kinakemic restructuring of its initial consonant: the kinakeme of occlusion is replaced by that of constriction and vice versa, voicing is interchangeable with devoicing (bod' ∼ vod' ∼ pod' - 'to hold'), the kinakeme of aspiration with that of constriction (k^{h}u ∼ xu 'arrow'). The role of kinakemes as markers of syntactic dominance in Nivkh is functionally analogous to the Persian eẓāfe, i.e. the suffix attached to the noun when it takes an attribute (pesar-e mard 'the man's son', āb-e garm 'hot water').

The same pattern of kinakeme interchange is widely used in Nivkh lexico-grammatical derivation. Replacement of initial occlusion by constriction may form a causative verb (kukud' "to fall" - ɣukud' "to drop"), a change in the opposite direction creates a deverbal noun (fuvd' "to saw" - p^{h}uf "a saw"). The kinakeme of voicing in the initial consonant of a qualitative word indicates intensity (tuzla 'cold' - duzla 'very cold') /2, 102; 6-I, 72; -II, 41/.

The use of a single kinakeme as an exponent of lexico-grammatical derivation is found in Russian, where the inclusion of the palatalization kinakeme into the stem-final consonant is highly productive in building deverbal and deadjectival nouns: подписать - подпись, связать - связь, обувать - обувь, бездарный - бездарь, удалой - удаль, новый - новь.

The conclusion can safely be reached that a wide range of kinakemes, both vocalic and consonantal, among them vowel checking and fronting, tongue raising and lowering, protensity, palatalization, occlusion, constriction, aspiration, voicing and devoicing, are quite capable of serving as the sole exponents of various derivational meanings - syntactic dominance in a phrase, morphological in the categories of case, number, gender, definiteness, person, mood; lexico-grammatical in shifting a word from one class to another.

In general linguistic terms the capability of kinakemes to function as morpheme exponents refutes the idea of the morphemic indivisibility of the phoneme. When a kinakeme, which is naturally incapable of externalization outside a phoneme, constitutes a morpheme shape in itself and is thus from the morphosemantic viewpoint independent of the phoneme it enters, it is separated from the other kinakemes in the phoneme structure by a morpheme boundary. Kinakemes are capable of building affix shapes not only singly, but also in clusters. Strictly speaking, every affix shape can be described as made of a kinakeme cluster, since a phoneme is always such a cluster. But there is no need to do so if the affix shape consists of entire phonemes. However, the identity between kinakeme clusters in their two constitutive functions - making up phonemes and morpheme shapes - is not obligatory, and a cluster as an affix shape may not equal any phoneme or phoneme combination.

The affix shape is often larger than a phoneme and contains an extra kinakeme which is incorporated into the adjacent phoneme of the stem. Linguistic tradition has treated such instances as sound alternations, phoneme replacements, which entails separate treatment for each case of replacement. Relegating the phenomenon to the kinakemic level brings more uniformity to linguistic description.

Of the languages discussed above Gaelic

and Russian abound in affix shapes that expand beyond the inflexional phonemes and penetrate into the phonemes of the stem. For instance, the Irish possessive noun prefix of the 3rd person contains, besides the entire phoneme /ə/, the kinakeme of constriction for the masculine sg. (port - a phort 'his port', cóta - a chóta 'his coat'), the kinakeme of voicing for the plural (a bport 'their port', a gcóta 'their coat'). In contrast the corresponding prefix for the fem.sg. is equal to the phoneme /ə/ (a port 'her port', a cóta 'her coat') /2, 79-83/.

In Russian the vowel /e/ is unable to begin a suffix shape alone and is therefore always accompanied in it by the kinakeme of palatalization implanted into the final consonant of the stem: loc.sg. столе, dat. sg. траве, inf. твердеть, where the palatalization kinakemes in /l'/, /v'/, /d'/ do not belong to the stem shapes, but to the affix shapes together with the vowel /e/. As a morpheme boundary separates the kinakeme of palatalization from the rest of the phoneme it joins, the stem shapes by themselves do not undergo any changes on the kinakemic level despite the changes in the kinakeme structures of their final consonants.

The Russian vowel /i/ is not always accompanied by the kinakeme of palatalization in suffix shapes. It equals the suffix shape in some noun inflexions (nom.pl. пилы), but in verb inflexions beginning with the same vowel phoneme it is accompanied by the kinakeme of palatalization placed in the last consonant of the stem: пилит.

Affix shapes may also be smaller than a phoneme, which then has to fill the resulting gap in its structure by admitting a certain kinakeme from the stem shape. This is the essence of synharmonism. For instance, in Finnish the kinakeme of vowel fronting or its negative counterpart is carried over from the stem vowels into the vowel of the suffix: inf. puhumaan 'speak' - leikkimään 'play'.

In English the suffix shape in 'hopes', 'moves' contains only the kinakemes common to both phonemes /s/, /z/, and the suffix shape in 'hoped', 'moved' likewise contains only the kinakemes common to /t/, /d/. In other words, the suffix shapes do not show any variation determined by the phonemic context. The kinakemes of voicing and devoicing which enter the suffixal consonants, belong to the stem shapes and not to the suffix shapes /13/.

Affix shapes larger or smaller than phonemes have a special role to play in strengthening the unity of the derived word, as the penetration of one morpheme shape into the phonemes that otherwise belong to the other morpheme, the resulting participation of a phoneme in two morpheme shapes at once are factors which help to cement the ties between the morphemes. But each direction of penetration has a typological significance of its own. When the affix shape is larger than the affixal phoneme and spills over into the stem, it serves to emphasize the constitutive role of the affix in the structure of the word and accordingly reduces the discernibility of the stem; this is a characteristic trend in syntheticized languages. On the other hand, analyticized languages show a typological propensity to emphasize the pivotal role of the stem by its easy separation from the affixes, and that requires the stability of the stem, well-defined morpheme boundaries; the unity of the word is also enhanced by morpheme boundaries running through phonemes, but the phonemes affected in such languages belong to affixes, whose shapes are smaller than the phonemes. It can be said that in the former type of languages the unity of the word is based on the power of its affixes, whereas in the latter type it uses the stem as its bulwark.

REFERENCES

/1/ И.А.Бодуэн де Куртенэ. Избранные труды по общему языкознанию, т.2. М., 1963.

/2/ Л.Г.Герценберг. "Морфологическая структура слова в ирландском языке". Морфологическая структура слова в индоевропейских языках. М., 1970.

/3/ В.М.Жирмунский. Немецкая диалектология. М.-Л., 1956.

/4/ Е.А.Крейнович. Фонетика нивхского (гиляцкого) языка. М.-Л., 1937.

/5/ Общее языкознание. (Внутренняя структура языка). М., 1972.

/6/ В.З.Панфилов. Грамматика нивхского языка, ч.I,1962; ч.II, 1965. М.-Л.

/7/ В.Я.Плоткин. Эволюция фонологических систем. М., 1982.

/8/ N.Chomsky, M.Halle. The Sound Pattern of English. Harper-Row, 1968.

/9/ E.Fischer-Jørgensen. Trends in Phonological Theory: A Historical Introduction. Akademisk Forlag. Copenhagen,1975.

/10/ M.Halle. Discussion of /13/, ibid.

/11/ R.Jakobson. Selected Writings, I. Phonological Studies. Mouton, 1962.

/12/ R.Jakobson, L.R.Waugh. The Sound Shape of Language. Harvester, 1979.

/13/ F.Palmer. "Grammatical Categories and their Phonetic Exponents". Proc. 9th Int.Congr.Linguists. Mouton, 1964.

/14/ V.Y.Plotkin. "Systems of Ultimate Phonological Units". Phonetica, 33 (1976), 2.

/15/ V.Y.Plotkin. "The Kinakeme as the Ultimate Unit of Language". Kwartalnik Neofilologiczny, XXV (1978), 3.

/16/ V.Tauli. Standard Estonian Grammar,I. Uppsala, 1973.

ON THE MORPHEME BOUNDARY AS A CRITERION OF PHONEMIC DIVISIBILITY

MARINA RUSAKOVA

Leningrad Bechterev
Psychoneurological Institu
Leningrad, USSR, 193019

ABSTRACT

Neither the morpheme, nor the word boundary can be regarded as the absolute criterion of phonemic divisibility of a phonetic complex. Phonemic divisibility in language does not fully coincide with a phonemic divisibility in speech. In most cases phonetic characteristics by themselves determine phonemic divisibility.

The assumption that a morpheme boundary can not lie within a phoneme is not argued by the representatives of different linguistic trends. Nevertheless, this statement doesn't seem to be as obvious as it is usually believed.

First of all, it must be noted that the present report is concerned with the branch of phonology in which the problems are treated in accordance with investigations of speech production and speech perception mechanizms. This phonological trend is bound with traditions of Leningrad Phonological school. The thesis that the morpheme boundary is the criterion of phonemic divisibility remains indisputable for modern representatives of Scherbian phonology /1/.

In the present report only the problems of the inflexional, fusion languages are treated.

The statement that a morpheme boundary can not lie within a phoneme is a consequence of Scherba's understanding of the phoneme as a linguistic unit which can be, on it's own, the signifier of morpheme. This definition of the phoneme is bound with the specific idea of the origin of the phonemic level in general and concrete phonemes in particular.

In his article "O diffuznyh zvukah" ("On Inarticulate Sounds") L.V.Scherba says that human speech originally consisted of inarticulate sounds, the latter being afterwards divided into phonemes. The morpheme boundary was "the cause" of this division /2/. That means that if a morpheme boundary does not lie within a phonetic complex, the latter is not divided into phonemes, as there are no reasons for this division. Scherba's statement, then, is connected, first of all, with the origin of phonemic level in human language, phonemes being differentiated later on the basis of inarticulate sounds. Thus, Scherba considered a morpheme boundary as the cause of phonemic divisibility, rather then a criterion of it.[1]

It seems that the problem of the origin of phonemes must not be identified with the problem of the divisibility of definite phonetic complexes in a language with a developed system of phonemes. Even accepting the assumption that in a developed language the speakers do not divide a phonetic complex which has no morpheme boundary within it (for there is no functional reason for such division according to Scherba), the reverse is not necessarilly true. In other words, if a morpheme boundary does lie within the complex in question, the latter may or may not be bi-phonemic.

The above is not an evaluation of Scherba's concept, but rather the tracing of the origin of the idea that a morpheme boundary determines the phonemic divisibility.

Let us define the phoneme as it is understood in the present report.

The phoneme is, no doubt, something sounding in speech, and a certain image in the psycholinguistic system. (We are not interested now in a very difficult problem of the correlation of various speech sounds and the corresponding linguistic and psycholinguistic inits). At the same time the phoneme is a constituent of signifiers of semantic units.

It seems that the phoneme as the constituent of the signifier in speech and phoneme as a unit of storage of signifiers in a speaker's lexicon should not be mixed up.

Being the constituent of the signifier in speech, the phoneme comes to the fore as a phonetic unit, characterised, first of all, by its "material" (acoustic, articulate, perceptive) qualities, as the "brick" of sounding. It means that a psycholinguistic system should include the set of "phonemes-sounds", the set of sound images.

As the unit of storage of signifiers in psycholinguistic lexicon the phoneme is a "brick" of the image of word sounding in the psycholinguistic system of an individual. If a signifier is kept in lexicon as the image of sounding then the "storage phoneme" is the image of the "phoneme-sound". It is possible, however, to suggest that the signifier is stored in lexicon as a chain of abstract units, a chain of indexes, not bound with the image of sounding. If so, the psycholinguistic system must include a set of "storage phonemes" and some mechanisms for re-coding "phonemes-sounds" into "storage phonemes".

In any case, the signifier of the word in "phonemes-sounds" may not, on the whole, coincide with the signifier of the word in "storage phonemes". For instance, the final "storage phoneme" in the Russian word ЛЕС is С . If in speech chain this word occurs before a voiced obstruent then the "phoneme-sound" З appears in the final position. To find the word in the inner lexicon, a speaker of the language should use some psycholinguistic rules to re-code the "phoneme-sound" chain ЛЕЗ into the "storage phoneme" chain ЛЕС.

It seems that Leningrad Phonological school, postulating the constant set of phonetic features for the phoneme, is oriented mostly to the "phoneme-sound". The concepts of Moscow Phonological school, postulating the constant phonemic organisation of the morpheme, are more applicable to the description of "storage phonemes", for it is quite possible that the basic allomorphs or phonetic allolexes represent semantic units in the internal lexicon; the latter, however, should be verified by experiment.

The question, whether the morpheme boundary is always connected with phonemic divisibility of phonetic complex, is solved according to the above-mentioned understanding of the phoneme.

The only functional reason to regard a boundary as a criterion of phonemic divisibility should be kept in mind: the boundary may be the place of coming together in the speech chain of two independent units, each of them represented in lexicon and, therefore, characterised by the permanent phonemic structure. So, if in the process of speech production or speech perception two independent semantic units occur side by side, then the phonetic complex, appearing at their juncture, is naturally biphonemic.

In derivatives, however, morphemes don't come together as independent units. A derivative, existing already in the language, is not "built" of morphemes in speech chain. This statement is confirmed by psycholinguistic experiments, carried out specially for the purpose. The experiment has shown that derivatives and non-derivatives required for their production and perception, while words made ad hoc, - and it is these words which rely on morpheme, - need more time, at least, for perception /3; 4/. The derivative's "life in language" is a gradual loss of motivation /5/. The phenomenon of "morphological absorption" was for the first time described by Bogoroditsky /6/. So, a morpheme juncture is usually not a boundary between independent units coming together in speech-processing. The more is the word assimilated by the language, the more its motivation is erased and the less important is its morpheme structure for phonological interpretation of sounds representing its signifier.

As it is not necessary to preserve the inner form of the word, the replacement of "inconvinient" combinations of phonemes at the morpheme juncture seems to be natural. Quite natural is also the fact that the combination incovinient for pronounciation is replaced by a phonetic complex coinciding with a phoneme of the language. And so, it is no suprise that in the word ДЕТСКИЙ the morpheme boundary lies within Ц . This Ц seems to be the "same" (in acoustic, articulatory, perceptive aspects) as the other Ц in the same position (e.g. in the word СТРЕЛЕЦКИЙ). So, if we speak about "phonemes-sounds", Ц in ДЕТСКИЙ can be nothing but the phonological complex Ц. Yet the fact that the morpheme boundary lies within Ц in ДЕТСКИЙ gives no reasons to consider Russian Ц as biphonemic, for phonetically Ц≠ТС and morpheme boundary does not often lie within Ц, besides we should just call for common sense. The appearance of Ц in ДЕТСКИЙ is a manifestation of morphological absorption of signifier. This process takes place on a word level and is the replacement of the "phoneme-sound" combination by one phoneme and has nothing to do with the phonological interpretation of the "phoneme-sound"

Some other conclusions can be drawn concerning "storage phonemes". The interpretation of Ц in ДЕТСКИЙ depends upon the morpheme organization of this word in internal lexicon.

The problem of morphological organisation of the word ДЕТСКИЙ is rather difficult. Regarding linguistic description as a model corresponding to speech behavior, the essence of the matter is as follows: does the speaker of language re-code the set of "phonemes-sounds" ЦК into the set of "storage phonemes" ТСК , restoring in this way the signifiers of morphemes in derivative. If so, the "storage phoneme" chain ТСК and if it is not so, the "storage phoneme" chain ЦК is the correlate of "phoneme-sound" chain ЦК in the word ДЕТСКИЙ. It depends, in its turn, on the degree of the loss of motivation in concrete words. It is also possible that the restoration of morpheme signifiers and the re-coding of "phoneme-sound" chain take place only in special speech situations, for example in the process of derivation, when the word stored in the lexicon serves as a model for a new derivative. It is also well known that new derivatives are not constructed as a sum of morphemes, they are formed by analogy with the derivatives already existing in the language.

As there is no need to preserve the inner form of the word, the way of accomodation of phoneme combinations at the morpheme juncture is also adopted from the pattern-word.

Thus the way of re-coding of juncture "phonemes-sounds" into "storage phonemes" in one derivative does not seem to contain information about the "storage phoneme" chain (correlating to the same "phoneme-sound" chain) of other derivatives. In each particular case the solution lies in the psycholinguistic lexicon.

All these facts make it possible to come to the following conclusion: the morpheme boundary is not an indisputable criterion of phonemic divisibility for either "phonemes-sounds" or "storage phonemes".

Now let us turn to the problem of word boundary as the criterion of phonemic divisibility.

Words are, no doubt, independent linguistic and psycholinguistic units. Phonetic complex appearing at their juncture is biphonemic "storage phoneme" complex, if we do not, of course, assume that every word-combination is included as a unit in the internal lexicon. So, word boundary shows that phonetic complex appearing at a word juncture is biphonemic.

The language speaker can never break the limits of his language habits, so it is no wonder that sounds, corresponding to the "phonemes-sounds" appear at word juncture. It seems, however, that the language speaker must have some special rules for production and perception of word juncture phonemes. The latter statement is supported by existence of phonemes which appear only at the juncture. Besides, it is possible that solutions concerning word juncture phonemes may be provided by higher linguistic levels, i.e. after reaching a solution concerning the whole word, the stage of recognition of "phonemes-sounds" being omitted. So, conclusions about the phonemic organization of juncture phonetic complexes should not be transferred to the coinciding complexes (maybe even coinciding by chance). The word boundary, then, cannot serve as a reliable criterion of phonemic divisibility for "phonemes-sounds". Summing up, we can conclude that neither morpheme nor word boundary is a criterion of phonemic divisibility of phonetic complex, if we speak about "phoneme-sound". The morpheme boundary is not the criterion of phonemic divisibility either if we speak of "storage phoneme". The problem of phonemic set of a derivative should be specially solved for each word. If we speak about "storage phonemes", then word boundary shows phonemic divisibility of phonetic complex appearing at a word juncture.[2]

How can be, then, solved the problem of phonemic divisibility for phonemes-sounds? V.Kasevic, who proceeds from the fact that speech is a continuum, concludes: "...the problem of segmentation belongs to linguistics rather than to natural sciences, i.e. acoustics or physiology" /1, p.17/. This statement, however, seems to be unacceptable.

If the speech chain were absolutely undivisible "materially", articulatorily, acoustically, and perceptively, the linguistic divisibility would also be absolutely impossible, for the latter can be nothing but an interpretation of information contained in the "material" side of speech. In fact, there are no gaps in speech chain, but the problem is not to find the very place where a phoneme boundary lies, but to identify the number of phonemes included in a phonetic complex. This problem is easily solved, if only "material" characteristics are taken into account. Speech is a successive changing of states of speech organs, and, hence, of acoustic and perceptive characteristics. In most cases there are no problems in identifying the quantity and character of such units. Of course, this approach, in the case of sounding speech, can cause some difficulties, because of coarticulation, reduction and so on. But the question is the divisibility of phonetic complex in language, not in speech. If a segment of speech chain does not contain enough phonetic information for its divisibility, coinciding with segmentation as it takes place in language, then the information of higher levels is used. Generally, however, there is no doubt that the chain CV (let us take Russian as an example) consist of at least two articulations. As both consonants and vowels can themselves be semantic units, every CV chain undoubtedly consists of two phonemes. Looking for a morpheme boundary for a phonemic interpretation of each CV chain seems to be naive as the method of quasihomonyms.

It should be noticed that the problem of phonemic divisibility rarely arises at all, outside affricates and diphtongs.

It seems that this problem cannot be solved with the help of phonological methods. We think that psycholinguistic experiments should be called to do it.

NOTES

1 It can be supposed that mechanisms of this kind are operative in the process of child mastering phonemic system.
2 Morpheme boundaries differ. A phonetic complex with a boundary between a stem and a flexion is more probable divided phonemically than a phonetic complex with a morpheme boundary of some other type. A boundary between the stem and the flexion, here, is closer to a word boundary.

REFERENCES

I. В.Б.Касевич. Фонологические проблемы общего и восточного языкознания. М., I983.

2. Л.В.Щерба. О диффузных звуках.- Языковая система и речевая деятельность. Л., I974, с.I47-I49.

3. М.В.Русакова. К вопросу о лингвистической и психолингвистической функции морфемы.- Семантические аспекты языка. Л., I98I, с.92-99.

4. М.В.Русакова, А.Ю.Русаков. Слово и морфема в речевой деятельности.- Лингвистические исследования I983. Функциональный анализ языковых единиц. М., I983, с.I68-I77.

5. Л.В.Сахарный. Словообразование в речевой деятельности: Образование и функционирование производного слова в русском языке. Автореф. докт. дисс. Л., I980.

6. В.А.Богородицкий. Лингвистические заметки о морфологической абсорбции.- Русский филологический вестник, вып.I, I88I.

ASPECTS OF THE SOUND FORM OF THE WORD

LUDMILA ZUBKOVA

Dept. of General Linguistics
P.Lumumba People's Friendship University
Moscow, URSS 117198

ABSTRACT

The paper seeks to develop the systemic approach to the analysis of the sound form of the word (SFW) comprising the unity of the universal, group and individual properties of the language, hierarchal stratification of its structural levels, hierarchal organization of lexis. Accordingly the SFW includes characterological, constitutive and paradigmatic aspects. A word is structured phonetically as a meaningful unit connected by constitutive relations with the morpheme and the sentence. The material structure of the word is correlated with its formal-semantic organization and reflects the degree of generalization proper to different classes it enters. Hence the hierarchal nature of the SFW. The phonological typology of the word reflects its systemic characteristics and correlates with morphological typology.

The SFW in terms of the characterological aspect is the unity of the general, peculiar and individual. The specific traits of the SFW in every language depend not so much on its unique features but primarily on the interaction of universal, typological, genetic and areal characteristics.

Though its own supersegmental features may not necessarily be present the word is always organized by the segmental means. These are its universal characteristics. The word phonemic structure as well as its supersegmental peculiarities (if any) reflect, on the one hand, the position of the word in the hierarchy of language units and in the hierarchal organization of lexis, on the other. In other words, constitutive and paradigmatic aspects are always present in the SFW. The constitutive aspect characterizes the word as part of the system of interlevel relations and determines its inner and outer form. The inner form is inherent in the word as a particular type of composition of morphemes. The outer form characterizes the word as a syntactically indivisible integral part of the sentence-utterance. The inner form is discrete, the outer form is indiscrete. The constitutive aspect is inseparable from the paradigmatic one: words of different semantic and grammatical classes differ in terms of their inner and outer forms. Therefore the morphological and syntactic features of the given class of the words as well as interrelations of this class with other classes of greater and/or less degree of generalization may play an important role in the SFW analysis.

The constitutive relations between the morpheme, the word and the sentence make for close connections between phonetic, morphological and syntactic properties of positions within the word. The statistic approach to the segmental structure of simple (root) full words in the languages of different typology and genesis /I/ has clearly demonstrated that the degree of activity of individual phonemes and phonemic classes in a given position depends on the above mentioned characteristics of this position. In accordance with the stratification of phonological oppositions, primary phonemes (in the sence of R.Jakobson, T.Milewski) play a leading role in the segmental structure of the word. That's why the above mentioned correlation characterizes first and foremost the consonantal structure of the word in general and distribution of the modal classes of consonants, in particular. The most contrasting types of primary consonants - voiceless stops and liquids reveal the strongest correlation with the position.

Being "a syntactic atom" (J.Baudouin de Courtenay) and "the potential minimum of the phrase" (E.D.Polivanov), the full word in its segmental structure reflects universal regularities of speech production, which are revealed in the universal tendency towards rising/rising-falling sonority of the word's segmental structure. This tendency manifests itself in mainly consonantal beginnings and vocalic ends of words, in preferable location of noise consonants in the initial and sonants in non-initial positions, in correspondence between synchronic sound positional modifications as well as diachronic phonetic changes and general dynamics of the word-utterance articulation. The sonorous structure characterizes the word as a whole and brings out its indiscreteness, the "contour" character of its segmental structure.

Since different consonant classes (modal as well as local) are preferably used in different positions, contrast tendencies are rather typical of consonant combinatorics in every language.

The type and degree of contrasts in consonantal structure are determined by the group characteristics of languages. Genetic affinity of languages specifies the features of consonant contrasts within the root word. Though other contrasts are possible the following features come to the fore in different language families: noise/sonant in the Indo-European languages, peripheral/medial in the Indonesian languages, forward-flanged/backward-flanged in the Altaic languages. The degree of contrast (thus, the degree of positions differentiation) is determined by language typology and depends on syntactic and especially morphological characteristics of the word. The phonemic structure of a simple word reflects the canonic type of its morphological structure, including presence/absence of affixation, its type and functional load.

Consequently the consonantal structure of a simple (root) word includes 3 types of positions, thus revealing its discrete character: within-root position, potential morphemes juncture, potential words juncture. The morphological status of position largely determines distribution and semiologically relevant potentials of phonemes, also their division according to markedness. As a reflection of phonological oppositions hierarchy and their development, this relationship is more typical of modal consonant correlations. In particular, the hierarchy of opposed voiceless-voiced (tense-lax) consonants depends mainly on their position regarding word boundaries. Higher distributional activity of "naturally" unmarked voiceless (tense) consonant corresponds to its position in the sonorous structure of a simple word and is observed in the position of potential words juncture even in the absence of phonological neutralization. This position coincides with the end of a simple word in prefixing languages and with its beginning in suffixing languages.

The degree of positions differentiation in consonantal structure of a simple word weakens as its semantic and syntactic independence is lessening and the functional load of affixation (primarily, post-root affixation) is increasing. The technique of morphemes connection (agglutination or fusion) is less important. Consonants in the position of potential morphemes juncture are similar to those of within-root, consonants in the position of potential words juncture are contrasting to those of within-root.

According to the degree of positions differentiation in consonantal structure of a simple (root) word different types of languages constitute a successive graduation, thus revealing structural isomorphism. The strongest differentiation of positions is typical of root-isolating languages. Then come languages with mostly or solely unilateral affixation. The languages with bilateral affixation are characterized by the weakest positions differentiation.

The word sonorous structure is modified in accordance with its canonic morphological structure. The suffixing language type promotes the tendency for rising sonority. This tendency is supressed to a certain degree in prefixing languages and languages with bilateral affixation.

Due to diagramic correspodence between consonantal structure of a simple (root) word and canonic morphological word structure of every language, phonological word typology clearly reflects morphological typology and therefore interrelations of all meaningful units - morphemes, words and sentences. The key role of the typological criterion for word segmental organization may be clearly seen in related languages of different types. Thus, the analitical Tajik language which possesses definite agglutinative traits differs in its consonantal structure of the word from the synthetic fusion-inlexional languages of the same Indo-European family - Russian and Czech, but resembles agglutinative languages of Altaic family - Turkish and Mongolian.

The universal tendencies in stratification of phonological oppositions also influence the word vocalic structure, which is testified by high activity of "primary" vowels, particularly "optimal" vowels of the /a/ type as the most open. As far as typology is concerned the vocalic structure of the word as opposite its consonantal structure is of little informative value. This may be accounted for a greater functional load of vowels in the word supersegmental structure, the type of which definitely correlates with morphological typology of the language (tone with isolating structure, synharmonism with agglutinative structure, free mobile stress with inflexional structure).

The constitutive aspect of the SFW deals not only with the above mentioned universal and typological tendencies in simple words and not only with phonemic structure, but with syllabic structure as well as supersegmental structure (if any). The sound form of the word as a syntactic unity (the outer form) shows the degree of potential isolation, preferable location of the word in the utterance, preferable type of syntactic connection with other words in accordance with its function, meaning and the part of speech it belongs to. The outer form of the word is revealed

not only in definite modification of its sonorous structure, but also in the presence/absence of external sandhi, coincidence/non-coincidence of word and syllable boundaries. The outer aspect of the SFW in the languages with the word stress can also be vividly seen in a degree of influence of phrase prosody on word prosody and respectively in the mode of prosodic emphasing a stressed syllable, in fixation/non-fixation and a degree of stability of word stress in speech, in a degree of accentual prominence of the word, in dependence of words and word-forms accent structure on their syntactic linkage and functions.

The inner side of the SFW reflects its morphemic composition and word-formation structure, the degree of synthesis and grammaticality. The sound form of the word as a morphological unit appears: I) in a specific phonemic structure for different types of morphemes depending on their number in morphemic inventory, meaning and position in the word. It can be seen in the quality and quantity of phonemes used, their distribution within morphemes and in the positions of words and morphemes juncture as well as in the degree of differentiation of the given positions, in phonemic combinatorics depending on location in morpheme and type of morphemes juncture, in length of morpheme in syllables and phonemes; 2) in fusion or agglutination technique of morphemes connection depending on their functional and semantic characteristics as well as degree of independence and, hence, in presence or absence of morphonological modifications, in the direction and force of assimilation, in interrelations of morphological division and syllabification in different types of morphemes junctures; 3) in supersegmental characteristics of morphemes. For example, in Russian it appears in different accent properties of the derivational base and formant, in different accentual activity of morphemes according to the stage and mode of derivation.

Thus, every phonetic characteristics of a word (segmental, syllabic, supersegmental) have inner and outer aspects, i.e. both syntactic and morphological value. The sound form characterizes the word as a syntactic whole as well as a complex morphological unit, therefore its segmental and supersegmental structure is the unity of indiscreteness and discreteness.

Both inner and outer aspects of the form are closely interlinked. Consequently the tendency towards rising sonority and interrelations of word boundaries with syllabic boundaries may be realised in various ways depending on the morphological structure of a given language, and the phonemic structure of morphemes corresponds to the word structure as a whole, and thus, to its outer form as a syntactic unit.

It is not by mere chance that quantitative typological approach to different types of languages showed good or medium essential correlation between the frequency of monophonemic morphs and the frequency of morphs juncture within the syllable, on the one hand, and indexes of lexical/grammatical, agglutination/fusion and synthesis characterizing the word. As grammatical, fusion and synthesis indexes increase, the frequency of monophonemic morphs grows and morphs division more frequently diverges from syllabification. The frequency of phonemes, mostly vowels, which can make up a morph by themselves increases respectively. The degree of phonemic autonomy in relation to morpheme is going down as grammaticality of the morpheme and the word is increasing. Specifically, in Russian the degree of phonemic autonomy in relation to the morpheme is higher in nouns which perform the nominative function and therefore can be regarded as lexical units to a greater extent than verbs which express the predicative function /2/. It follows that phonemes constitute a morpheme not as a separate (autonomous) element but as an integral part of the word as a whole which has a certain meaning and performs a certain function. Different functional load of phonemic classes in the constitution of different types of morphemes and, thus, in the expression of meanings can be vividly seen in phonemes and morphemes correspondence in regard to their markedness. The greater the degree of phonemic markedness, the less the occurrence of phonemes in marked, syncategorematic, morphemes. Phonological oppositions stratification is thus shown to correspond to stratification of morphological differences, lexical and grammatical meanings.

Due to the unity of inner and outer aspects, the SFW appears as a result of interaction of all systemic characteristics of the word: word-formation structure, morphemic constitution, inflexional type, syntactic linkage as well as functional and semantic properties. As the word is a many-sided unity, every characteristics of its sound form can be viewed upon from different angles. Thus, regarding the word accentuation from the point of view of word-formation structure the accentual properties of the derivational base and formant come to the fore ground. And from the second stage of derivation and on, the depth of accentual motivation (that is not only immediate but also distant accentual connections of derived and deriving words) get important. For the word as a system of word forms the location of stress on stem or flexion is relevant. In the word (word form) as a unity of morphemes (morphs) the stress marks one of the morphemes. In the utterance the rhythmical structure of word-forms and the degree of their accentual prominence is getting dominant.

Interrelations of outer and inner form, word-formation and inflexion are clearly reflected in Russian in allomorphic root variations and dynamics of accentuation depending on the stage of derivation. With the increase of the stage of derivation the length of allomorphs alternation series gets shorter, the number of alternative phonemes in the allomorph is reduced, the frequency of alternative allomorphs is going down, the frequency of non-alternative allomorphs as well as of the basic allomorphs increases. Consequently the clarification of week root phonemes in inflexional paradigm is growing more complicated. On the other hand, less accentual activity of the derivational affix, lower frequency of immediate accentual motivation and the growing frequency of distant accentual motivation on higher stages of derivation result in reduction of mobile and inflexional stress and domination of fixed stress on the inflexional stem.

As a result of interrelation of all systemic characteristics of the word, its sound shape apart from individual features includes definite class features which are closely linked with other properties of a given word class. The character and the degree of phonetic differences among different word groups, as well as groups qualification and means of their formal expression depend on language typology and the degree of morphological development of language in particular. However the biggest word groups - full and form words, nouns and verbs - seem to be more or less phonetically different in every language, including isolating languages.

Phonetic differences of parts of speech include all above mentioned characteristics. Especially modification of sonorous structure, the degree of positions differentiation, correlation between morphological and syllabic division change in line with different degree of syntactic independence, word-order and morphemic structure of parts of speech /2; 3/. Supersegmental differences of parts of speech rather various. They may be observed in the preferable use of a definite supersegmental means (not in all, but in one or several parts of speech), in type and number of supersegmental patterns, in functional load of supersegmental characteristics for the expression of lexical and grammatical meanings, in mobile or fixed supersegmental structure and in case of mobility - in its sphere, type and function, in predictability of the supersegmental pattern on the basis of the word morphological structure, in degree of stability of supersegmental structure of the word, in word's ability to retain it in speech /2/.

As the word enters various semantic and grammatical word groups which are different in terms of degree of generalization and in their scope, the SFW is hierarchal. It could be proved by comparing the length in syllables and the accentuation of different groups of words in Russian, starting from most general classes - full and form words. Form words, first of all the primary ones, lack the nominative and significative function, possess a relative and highly generalized meaning, are syntactically dependable, they usually lack phonetic independence. They assimilate with full words and may have a non-syllabic structure. In case of syllabic structure form words are usually unstressed and are liable to re-syllabification in the juncture with full words. Full words are of greater semantic, syntactic and accordingly of phonetic independence. Due to potential isolation (the full word can make up an utterance) they always have the syllabic form and are usually stressed. Substitutive words (pronouns, in particular) have intermediate position: being more abstract they differ from full words by fewer number of syllables and a weaker accent in the utterance. Among full words nominative word-signs (nouns) differ from predicative word-signs (verbs) by a greater length of the root and more stable word stress in the utterance. Accentual differences between these two major parts of speech also applied to distribution of accentual paradigms, a type of a stressed morpheme within a stem, rhythmical structure of word-forms. Further gradual division of nouns in terms of the categories: abstract/concrete, animate/inanimate, person/non-person, countable/uncountable - also display differences in all above mentioned accentual characteristics.

Different sound forms of different word classes prove the categorial nature of the relationship between the sound structure of a word and its meaning. Systemic correspondence of the material (phonetic) and semantic structures of the word provides for the unity of sound and meaning and brings the arbitrariness of the word-sign to its limitation /4/.

REFERENCES

/I/ Л.Г.Зубкова, "Сегментная организация простого слова в языках различных типов", Дис. на соискание уч. ст. докт. филол. наук, Москва, I978.

/2/ Л.Г.Зубкова, "Части речи в фонетическом и морфонологическом освещении", Москва, I984.

/3/ Л.Г.Зубкова, "Звуковая форма частей речи", Народы Азии и Африки, I978, № I.

/4/ Л.Г.Зубкова, "О соотношении звучания и значения слова в системе языка (К проблеме "произвольности" языкового знака)", Вопросы языкознания, I986, № 5.

LA PALATALISATION ET LA PHARYNGALISATION EN POLONAIS DANS UNE COMPARAISON ENTRE PLUSIEURS LANGUES

Éva Földi

Département de Phonétique
Université Eötvös Loránd, Budapest

Résumé

Cette étude a pour but d'examiner la réalisation articulatoire et acoustique de la palatalisation et pharyngalisation à la base d'un matériel linguistique polonais en confrontation avec d'autres langues /et notamment le russe et le hongrois/. L'analyse a été faite essentiellement au moyen des méthodes radiocinématographiques et spectrographiques, complétée par une analyse assistée par ordinateur.

Introduction

L'opposition des consonnes dures et molles, phénomène présent dans plusieurs langues slaves et ayant des effets considérables sur les systemes de sons et de phonèmes, a donné jusque'à présent naissance a des vues plutôt divergeantes au cours des recherches. On connait des points de vue très différents surtout concernant le statut phonétique et phonologique des sons et phonèmes palatals. La polonistique n'est pas encore parvenue à tirer au clair les questions du déroulement articulatoire et de la projection acoustique de la palatalisation et pharyngalisation. C'est surtout l'aspect synchrone ou asynchrone ainsi que le degré de la palatalisation qui est en question; quant à la pharyngalisation, elle n'est même pas beaucoup étudiée. La discussion est devenue particulièrement animée au sujet des occlusives bilabiales molles [b', p', m']. Plusieurs chercheurs nient l'existence de ces sons et phonèmes, leur articulation serait même physiologiquement impossible /5, 6, 7/; d'autres pourtant sont de l'avis contraire /1, 8, 9, 10/.

Dans la présente étude je me penche sur la réalisation articulatoire et acoustique de la palatalisation et de la pharyngalisation tout au long du déroulement articulatoire des sons concernés. L. matériel de l'analyse a été constitué par des mots polonais comportant les consonnes en question en position initiale. Ce corpus a été composé de manière que toutes les modes articulatoires soient répresentées /occlusive, fricative, affriquée, latérale, vibrante; par exemple: pasek - piasek, wara - wiara, dżonką - dzionka, lato - list, rym - ring etc./. J'ai analysé le déroulement articulatoire et les traits acoustiques par les méthodes respectivement radiocinématographique et spectrographique, en les complétant par une analyse assistée par ordinateur. La prise de vue radiocinématographique a été exécutée au moyen d'un appareil radiographique de type Siemens Sirescop 2 et d'un magnétoscope de type Siemens Sirecord S. L'analyse spectrale a été faite à l'aide d'un spectrograph te-type 700, et finalement j'ai procédé à l'analyse assistée par un ordinateur individuel de type Commodore 64. Les résultats ainsi obtenus ont été confrontés à ceux dégagés de deux autres langues, l'une typologiquement semblable /le russe/, l'autre différente /le hongrois/. L'analyse de ces langues était menée avec une méthodologie identique, basée sur la concéption de Kálmán Bolla dans le cadre des recherches de phonétique contrastive /2, 3/.

Les expériences ont été effectuées avec 19 sujets de langue maternelle polonaise, ayant une prononciation conforme aux normes du polonais standard des milieux cultivés. Les radiogrammes reproduits dans cette étude ont été faits à partir des réalisations de l'informateur masculin noté 'St'.

Développement

Le polonais, tout comme le russe, est de caractère consonantique. Cela veut dire entre autres que dans la chaîne parlée ce ne sont pas les voyelles à conditionner la qualité des consonnes environnantes, mais bien au contraire: le type de la voyelle dépend de la qualité de la consonne avoisinante. Par exemple, une voyelle palatale ne peut pas se trouver après une consonne pharyngale /cf. 1. [bɨwɨ] -- 2. [b'il'i]/. Cette condition, comme il s'ensuit aussi de l'exemple précédent, est également porteuse de pertinence linguistique /1. = 'ils étaient', 2. = 'ils battaient'/. Il est en outre nécessaire de faire remarquer que 1. le polonais standard ne connaıt que les voyelles monophtongues; 2. il est plus correct parler de pharyngalisation et non pas de vélarisation puisqu'au cours de l'articulation d'un tel son la langue se déplace dans la direction de la paroi pharyngale plutôt que vers le voile du palais.

Au cours des analyses, en tenant compte de ce qui vient d'être dit, j'ai cherché à répondre aux questions suivantes: 1. s'agit-il d'une palatalisation synchrone ou asynchrone en polonais, surtout en ce qui concerne les occlusives bilabiales molles; 2. quelles sont les conséquences acoustiques du déroulement articulatoire?

1. L'analyse du matériel sonore et visuel /radiocinématographique/ enregistré sur casette vidéo, permet de faire les constatations suivantes:

-- On peut bien suivre la pharyngalisation tout le long de l'articulation, voir fig. 1. /Nous avons divisé chaque son en 5 frame ou tranches phoniques en succession dans le temps, cela permet de se faire une image plus précise de toutes les phases de l'articulation./ Cela est confirmé aussi par les données obtenues par l'analyse /assisté par ordinateur/ de la surface totale et des surfaces partielles des cavités supraglottiques, divisées en 5 parties /labiale, palatale, vélaire, pharyngale et nasale/. J'ai examiné les données de surface des parties, calculées en pourcentage par suite de la comparaison à la surface totale et aux autres parties prises individuellement.

-- Dans l'étude de l'aspect synchrone ou asynchrone de la palatalisation j'ai choisi des mots où la consonne soit suivie de [i] mais aussi d'une voyelle différente. Les analyses ont prouvé la présence d'une palatalisation synchrone dans

l'articulation des consonnes molles polonaises, même dans le cas si âprement discuté des occlusives bilabiales molles /voir fig. 2. et 3./, quelle que soit la voyelle suivante. La configuration buccale caractéristique de l'articulation palatale se déclare dès la phase initiale, et cela est valable même pour les occlusives bilabiales molles et sourdes /fig. 2. et 3./.

2. L'analyse des sonagrammes a fait ressortir le problème suivant: après les occlusives bilabiales molles, si la voyelle suivante était autre que [i] , on peut voir un segment de type [i̯] d'une durée de 50 à 80 msec, intercalé entre la consonne et la voyelle concernées et dont l' "appertenance" est discutée. J'ai également analysé la structure temporelle des sons, pour [p, p'] par exemple j'ai obtenu les données que voici: la durée totale de [p] est de 120 msec, l'explosion en représente 10 à 15 msec; la durée totale de [p'] /suivi de [a] / est de 160 msec, la phase de silence est de 90 à 100 msec, l'explosion est de 20 à 30 msec, et le segment [i̯] est de 60 msec. Il est évident que ce dernier segment ne peut pas être la réalisation du phonème /j/; il ne peut être l'élément d'une voyelle diphtongue constituant de syllabe non plus. En comparant les analyses radiocinématographiques et spectrographiques il est possible de concluse que le segment de type [i̯] se présente en tant que projection acoustique de l'articulation palatale de certaines consonnes /occlusives/. Cela est confirmé par les données de l'évolution historique de la langue polonaise /1/, et les règles orthographiques polonaises fournissent également un argument en faveur de cette analyse: pour noter les consonnes molles suivies d'une voyelle autre que [i] on se sert du graphème i /par exemple: piasek/, et la coupe syllabique ne peut jamais tomber entre ce graphème et la lettre consonantique ou vocalique voisine /pia-sek, ma-la-ria, et non ma-la-ri-a!/. Les résultats des analyses prouvent qu'en polonais les sons [b', p', m'] sont les réalisations des phonèmes /b', p', m'/, et non pas les variantes de /b, p, m/ avant /j/.

Quant au hongrois, les diagrammes du matériel linguistique reflètent l'absence de l'opposition entre palatalisation et pharyngalisation. Dans la confrontation à la langue russe, la palatalisation et pharyngalisation du polonais ne s'avèrent pas différentes, en pensant avant tout à la synchronie.

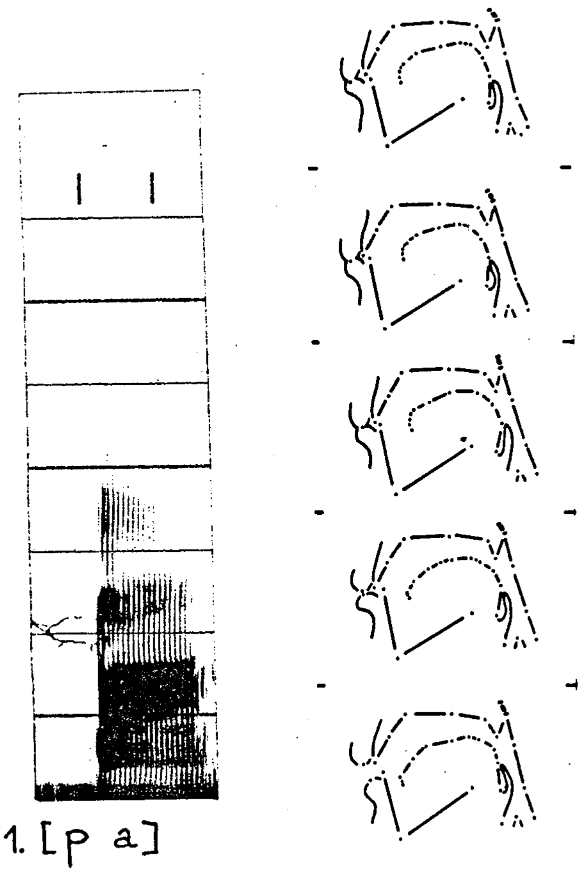
Conclusion

L'analyse combinée du déroulement articulatoire et du résultat acoustique prouve d'une façon évidente qu'en polonais la palatalisation s'étend sur la durée totale de l'articulation, même dans le cas tant discuté des occlusives bilabiales. Quoique le changement du processus acoustique reflété sur les sonagrammes n'exprime guère de façon manifeste l'existence de la palatalisation d'un bout à l'autre de la durée du son, les résultats obtenus à partir de l'analyse /assisté par ordinateur/ des données dégagées des tracés radiocinématographiques en témoignent sans laisser de place à la doute.

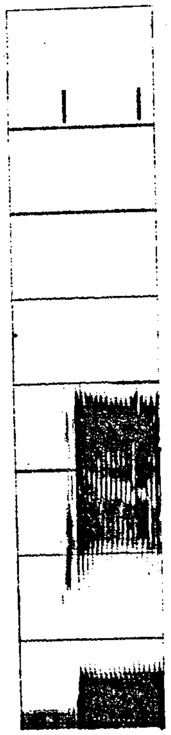
Références

Baudouin de Courtenay J.: Zarys historji jezyka polskiego. Warszawa 1922.
Bolla K.: Orosz hangalbum. MFF 11. 1982.
Bolla K.--Földi É.: A lengyel beszédhangok képzési és akusztikus sajátságairól. MFF 7. 1981, 91--139.
Bolla K.--Földi É.--Kincses Gy.: A toldalékcső artikulációs folyamatainak számítógépes vizsgálata. MFF 15. 1986, 155-66.
Jassem W.: Mowa a nauka o łączności. 1974.

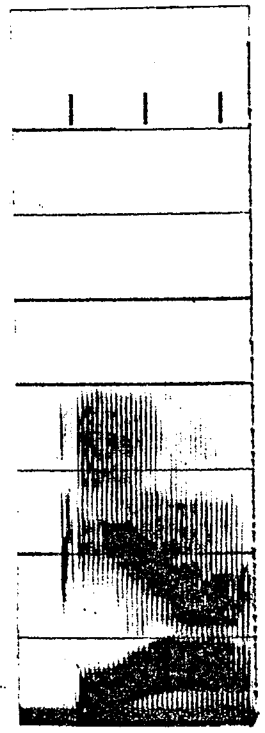
Koneczna H.: Charakterystyka fonetyczna języka polskiego... Warszawa 1965.
Rocławski B.: Istota miękkości. Język Polski LVI/1. 1976, 26--36.
Rocławski B.: Palatalność. Gdańsk 1984.
Szober S.: Gramatyka języka polskiego. Warszawa 1931.
Wierzchowska B.: Fonetyka i fonologia języka polskiego. Warszawa 1980.



1. [p a]



2.[p' i]



3.[p' ȧ]

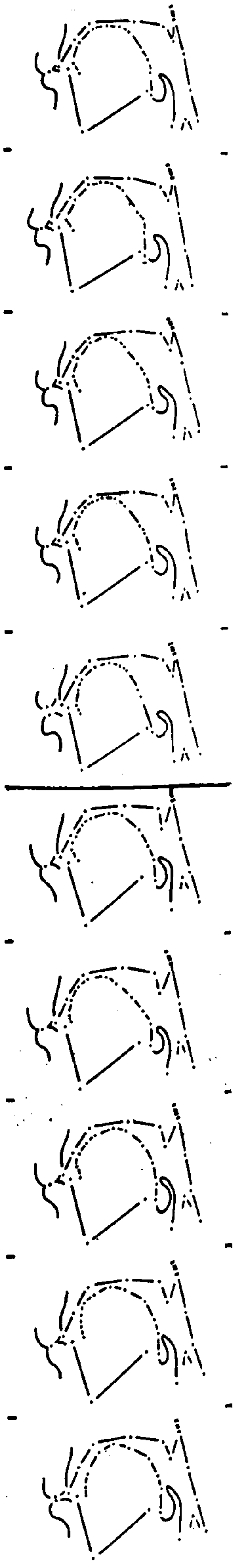


Fig. 1--2--3. Cinéradiogrammes de [p],[p'] [p'] des mots "pasek, pili, piasek" et extraits de sonagrammes des premières syllabes des mêmes mots.

EXPERIMENT IN INTERLINGUAL TYPOLOGICAL EXAMINATION OF VOWELS

by Gábor Kozma

Phonetic Department
of Eötvös Lorand University, Hungary

Abstract

We showed certain possibilities of the realization of the known connections between sound quality and the function of the organs and their relations. Our typological comparison pointed out the role of certain articulatory features of the examined sounds in forming the phonetic peculiarities.

The purpose of the examination

In our analysis we examined sounds belonging to five types /„i,e,a,o,u"/ in German, Hungarian, Russian, and Polish languages /see Fig./. On comparing the articulatory features we want to give an answer for the question of how to inter-

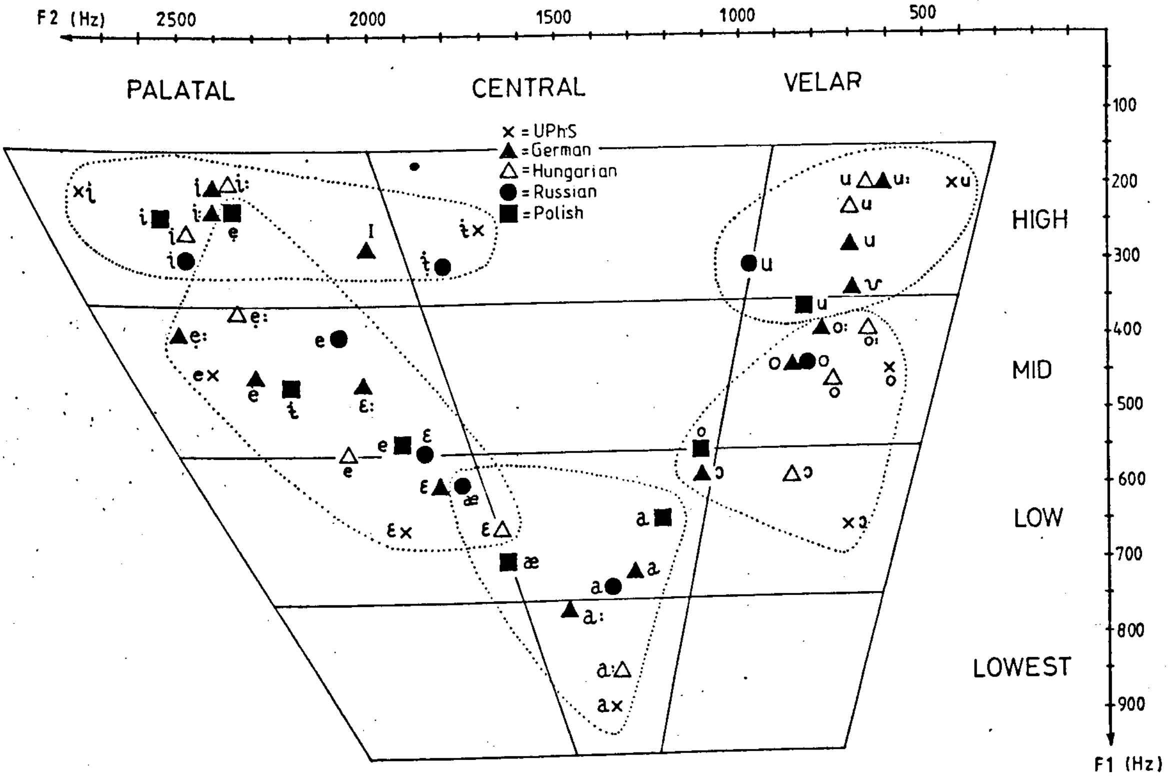


Fig. The analysed vowels according to the articulatory features and F1, F2 values pret the more palatal, more open or more labial articulatory character of a sound. We also examined what articulatory processes result in the sounds of near acoustical features /8/.

The course of the examination

We examined the connections between the articulatory features and acoustic characteristics of the sounds on the basis of data obtained from dynamic spectrum analysis and the analysis of the cinelabiographic and cineradiographic recordings of sounds by computer /1/,/2/, /4/,/5/,/6/,/7/. All of this was made according to the features given in the Universal Phonetic Sound Standard - ergo according to the exact etalon - elaborated by Kálmán Bolla /3/. The analysis of the position of the articulatory organs and the zone relations of the supraglottal cavities was by cineradiographic examination.

If we compare the proportion of the velar /T3/ and palatal /T2/ zones of a sound to the proportion value of another similar sound we can point out the strengthening of the palatal character or the weakening of the velar character.

If we compare the order of the sounds formed in this way according to the place of articulation with the order supposed on the basis of the F2 values, the conclusion on the connection between the articulatory features and acoustic characteristics, and on possible modifying influence can be drawn /Table 1./. The orders of two kinds of „e" type are different for example in Hungarian and Polish languages, almost the same in German, and the very same in Russian language. We can see on the Figure that the elements of the Hungarian and Polish „e" type are on different bond of tongue position, they scatter a lot.

	P ⟶ V	GERMAN	HUNGARIAN	RUSSIAN	POLISH
„i" type	F2	[i]=[i:],[I]	[i],[i:]	[i],[ɨ]	[i],[ɨ]
	T3/T2	[i:],[i],[I]	[i:],[i]	[i],[ɨ]	[ɨ],[i]
„e" type	F2	[e:],[e],[ɛ:],[ɛ]	[e:],[e],[ɛ]	[e],[ɛ]	[ę],[e]
	T3/T2	[e:],[e],[ɛ:]=[ɛ]	[e:],[ɛ],[e]	[e],[ɛ]	[e],[ę]
„u" type	F2	[ʊ]=[u],[u:]	[u],[u:]	[u]	[u]
	T3/T2	[ʊ],[u],[u:]	[u],[u:]	[u]	[u]
„o" type	F2	[ɔ],[o],[o:]	[ɔ],[o],[o:]	[o]	[o]
	T3/T2	[ɔ],[o],[o:]	[o],[o:],[ɔ]	[o]	[o]
„a" type	F2	[a:],[a]	[a:]	[æ],[a]	[æ],[a]
	T3/T2	[a],[a:]	[a:]	[a],[æ]	[æ],[a]

Table 1. The sound types according to the horizontal movement of the tongue

The other sound type also prove that in languages where the orders of two kinds according to the place of articulation are the same, the sounds are formed on almost the same height of the tongue, and the difference in the orders of two kinds is connectings with the difference of the height of the tongue. So there is a relationship in the same sound type between the F2 data and the place of articulation but it is valid only within the same height of the tongue.

We can see in the Fig. that the steepness of the straight line connecting the elements of the Hungarian „i" and Russian „a" has an opposite sign in comparison with the „i" and „a" types of other languages.

From the F1, F2 values we can come to the conclusion on the shape of the lips: smaller F1, F2 values refer to more labial character. Its realization can be characterized by proportion of the vertical and horizontal diameter of the shape of the lips / V/H /. The opposite sign change of F1, F2 values can be connected to opposite direction physiological processes.

If the angle between the jaws, that is the change of the distance of the upper and lower lips can be connected to the changes of the F1 value, then the change of the distance of the corners of the lips will have the contrary effect to this, and we can suppose that the F2 value has a connection with the change of the vertical diameter of the shape of the lips. On the basis of V/H values the change of the distance of the corners of the lips plays a dominant role in formation of the character of labialization.

The F1, F2 values of the Polish [ę] and [ɨ], the Russian [æ] sounds are close to the values of other sound-types. This phenomenon can be examined according to the Table 2.

On the basis of all this for exemple we can state that the Polish [ɨ] sound is articulated by other tongue position as it could be expected on the basis of the F1, F2 values. Only the F1,F2 values of the [ɨ] sound and its surroundings are acoustically the same, the other different acoustic features are based on the articulative base constituted by not only the different tongue positions but other articulatory marks /e.g. the degree of lip rounding/.

	ACOUSTIC FEATURES		ARTICULATION		ALL ACOUSTIC
	F1, F2	WHOLE	TONGUE	WHOLE	FEATURES
1.	+	+	+	+	+
2.	+	+	+	–	+
3.	+	+	–	–	+
4.	+	–	–	–	–
5.	+	+	–	+	+
6.	+	–	+	+	–
7.	+	–	–	+	–
8.	+	–	+	–	–

The comparison of the acoustic and articulatory features of two sounds (+ =similar, – =different)

Table 2.

Our methods of examination help the expressive and quick objective analysis of the lingual facts. These methods improve our knowledge of certain sounds not only by characteristics grantable by typological and interlingual comparisons, but they make it possible to elucidate the specific connection of articulatory features from new sides.

REFERENCES

Bolla, K.: A phonetic conspectus of Hungarian. MFF 6. 198o, 167 p.

Bolla, K.: A phonetic conspectus of Russian. MFF 11. 1982, 25o p.

Bolla, K.: A universal phonetic standard? Vowels. MFF 13. 1984, 71-12o.

Bolla, K. and Földi, É.: Articulatory and acoustic features of polish speech-sounds. MFF 7. 1981, 91-139.

Bolla, K. and Földi, É.: The labial articulation of Polish speech-sounds. MFF 8. 1981, 1o4-146.

Bolla, K. and Földi, É. and Kincses, Gy.. Examination of articulative processes of supraglottal cavities by computer. MFF 15. 1986, 155-165.

Bolla, K. and Valaczkai, L.: A phonetic conspectus of German. MFF 16. 1986, 21o p.

Joos, M.: Acoustic phonetics. Baltimore 1948. /Suppl. to Language: vol. 24, No. 2. suppl./

ОППОЗИЦИЯ ПАЛАТАЛЬНОСТЬ-НЕПАЛАТАЛЬНОСТЬ В РУССКОМ И БОЛГАРСКОМ КОНСОНАНТИЗМЕ

МАГДАЛИНА А. ГЕОРГИЕВА

Союз переводчиков Болгарии
1000 - София, Болгария
ул. Неофит Рилски № 5

РЕЗЮМЕ

Исследование оппозиции палатальность-непалатальность русских и болгарских согласных проводилось на акустическом и перцептивном уровнях с привлечением статистических данных. Целью работы было выявление контрастивной для двух языков специфики систем в целом, а также дистрибуции палатализованных и частных случаев артикуляции, способных привести к фонологическим ошибкам болгар, изучающих русский язык. Полученные результаты легли в основу пособия по русскому произношению и интонации, которое использовалось при проведении совместно с В.Г. Смирновой интенсивного курса для болгарских студентов-русистов Софийского университета.

Фонетические исследования проводились в Фонетической лаборатории Института болгарского языка БАН в Софии и в Лаборатории 1-го МГПИИЯ.

РАЗЛИЧИЯ В ОППОЗИЦИЯХ РУССКИХ И БОЛГАРСКИХ СОГЛАСНЫХ ПО ДАННОМУ ПРИЗНАКУ ПАЛАТАЛЬНОСТИ

Как известно, в русском и болгарском консонантизме существует оппозиция по признаку палатальности /хотя в двух языках некоторые согласные и остаются вне корреляции и наблюдается отсутствие симметрии в их реализации, см. дальше/.

Ниже приводятся статистические данные о соотношении болгарских и русских непалатализованных/палатализованных согласных. Болгарский материал дается на основе проведенного М. Мариновой и Ас. Мариновым исследования текстов 6 болгарских писателей - классиков и современников, а также статей из 3 крупных болгарских газет /1/. Русские данные взяты из работы Д. Спировой, проанализировавшей тексты из области художественной, общественно-публицистической и научно-технической литературы /2/.

Процентное соотношение непалатализованных и палатализованных согласных болгарского языка /в убывающем порядке/

з	:	з'	=	2,02	:	0,01	=	202,00
с	:	с'	=	5,45	:	0,03	=	181,67
п	:	п'	=	2,73	:	0,02	=	139,00
д	:	д'	=	2,95	:	0,06	=	49,17
р	:	р'	=	4,62	:	0,10	=	46,20
в	:	в'	=	4,08	:	0,09	=	45,33
м	:	м'	=	2,24	:	0,05	=	44,80
т	:	т'	=	7,95	:	0,19	=	41,84
н	:	н'	=	6,17	:	0,19	=	32,47
ц	:	ц'	=	0,58	:	0,02	=	29,00
б	:	б'	=	1,44	:	0,07	=	20,57
л	:	л'	=	3,12	:	0,21	=	14,85
х	:	х'	=	0,77	:	0,06	=	12,83
к	:	к'	=	3,43	:	0,28	=	12,25
г	:	г'	=	1,26	:	0,15	=	8,40

По-видимому ф' не встретился ни разу в исследованных текстах, поэтому оппозиция ф:ф' и не фигурирует в таблице. Некоторое недоумение вызывает сравнительно высокий процент палатализованных задненебных, хотя известна их ограниченная дистрибуция /в особенности, х'/; может быть авторы причислили сильно палатализованный аллофон к, г и х в позиции перед и и е к к', г' и х' /однако срв. 3,4/. Цифровые данные о дз дз' авторами не приводятся.

Процентное соотношение непалатолизованных и палатализованных согласных русского языка /в убывающем порядке/

ф	:	ф'	=	1,29	:	0,09	=	14,26
х	:	х'	=	1,18	:	0,10	=	12,04
з	:	з'	=	2,32	:	0,24	=	8,42
п	:	п'	=	4,47	:	0,73	=	6,15
г	:	г'	=	1,24	:	0,20	=	6,10
к	:	к'	=	4,42	:	0,79	=	5,59
в	:	в'	=	5,08	:	1,12	=	4,51
б	:	б'	=	1,66	:	0,47	=	3,53
т	:	т'	=	6,00	:	2,83	=	2,12
р	:	р'	=	4,04	:	2,15	=	1,88
н	:	н'	=	6,92	:	3,85	=	1,80
с	:	с'	=	4,45	:	2,49	=	1,79
д	:	д'	=	2,36	:	1,46	=	1,62
м	:	м'	=	2,98	:	1,88	=	1,58
л	:	л'	=	2,40	:	3,40	=	0,71

Сопоставление процентного соотношения русских и болгарских непалатализованных и палатализованных согласных

м:м'=	1,58 :	44,80	д:д'=	1,62 :	49,17
н:н'=	1,80 :	32,47	к:к'=	5,59 :	12,25
л:л'=	0,71 :	14,85	г:г'=	6,10 :	8,40
р:р'=	1,88 :	46,20	в:в'=	4,51 :	45,33
п:п'=	6,15 :	139,00	с:с'=	1,79 :	181,67
б:б'=	3,35 :	20,57	з:з'=	8,42 :	202,00
т:т'=	2,12 :	41,84	х:х'=	12,04 :	12,83

Как видно из представленных трех таблиц, единственное приближение соотношения дистрибуции русских и болгарских непалатализованных согласных наблюдается у х:х'= 12,04 для русских аллофонов и 12,83 для болгарских. Самые высокие /трехцифренные/ величины при сопоставлении болгарских непалатализованных согласных с палатализованными в процентах дают свистящие с и з, а также билабиальный аллофон п. В отношении оппозиции русских непалатализованных с палатализованными наиболее высокие показатели у ф и х, а наиболее низкий у л.

Общие заключения, которые можно сделать на основе анализа приведенных статистических данных говорят о перевесе /безусловном/ непалатализованных в болгарском консонантизме /максимальный процент - 202, а минимальный - 8,40/, тогда как в русском максимальный процент составляет 14,26, а минимальный - отрицательное число /0,71/.

Основные различия в русской и болгарской палатализации касаются, во первых, самой ее степени: у русских согласных она значительно выше, чем у болгарских, что в артикуляторном аспекте выражается в большей степени подъема обычно средней части спинки языка по направлению к твердому небу, а, во вторых, в их дистрибуции; есть три позиции, вполне частотные для русских согласных, но запретные для болгарских, а именно: перед передними гласными и и е, перед непалатализованными согласными и в абсолютном исходе слова.

Замена русского палатализованного согласного болгарским непалатализованным /а в некоторых случаях - только адаптированным к передним гласным, срв. комбинации болг. ки, ги, хи, ли, в которых при акустическом исследовании шумовые зоны и формантные величины согласных почти идентичны с аллофонами палатализованных фонем /перед задними гласными/, приводит к фонологическим ошибкам. С одной стороны, непалатализованный согласный в русских иноязычных заимствованиях звучит одинаково с соответствующими болгарскими, например: рус. и болг. реноме, макраме, кабаре,купе, артерия, дека, кеб, сет, кафе и т.д. С другой стороны, ошибочное произнесение слов с подменой палатализованного согласного непалатализованным по причине интерференции приводит к смысловым ошибкам, например: бил вместо был, сито вместо сыто, висок вместо высок, бери-бери /с непалатализованным б вместо б'/ вместо [б']ери, [б']ери, парез вместо по[р']ез, метр /"учитель"/ вместо [м']етр, для теста /от "тестировать"/ вместо для[т']еста и пр. К таким же фонологическим ошибкам относится подмена русского палатализованного согласного непалатализованным в остальных двух позициях, которая происходит в силу лексической близости двух языков /несмотря на наличие орфографического сигнала палатализации, выраженного мягким знаком/, например: танка вместо Танька,горка вместо горько, варка вместо Варька, полка вместо полька, удар вместо ударь, порол вместо пароль, дан вместо дань и пр.

Еще одна особенностьболгарского языка, характерная, прежде всего, для жителей Западной Болгарии, в диалекте которых отсутствует палатализация согласных, заключается в подмене палатализации йотообразным призвуком: б^{И}ал, н^{И}ама и пр.75/. При таком произнесении сочетаний палатализованного согласного с гласным в русском может произойти фонологическая ошибка, например: Мор[j]ак вместо моряк, по л[j]у вместо полю, об[j]еду вместо обеду, пен[j]е вместо пене и пр.

Вне корреляций по палатальности в двух языках остаются некоторые согласные, а кроме того не все коррелирующие пары болгарских непалатализованных/палатализованных согласных наличествуют в русском языке. Так, в отличие от русского консонантизма, в болгарском аффрикаты образуют пары не только по признаку звонкость-глухость /ц -дз, ч - дж/ но и по наличию палатализации /ц' - дз'/. В болгарском языке вне коррелятивного противопоставления по признаку палатальности остаются непалатализованные согласные ж, ш, ч, дж и палатальный йот. В русском же кроме непарного палатального йот аффриката ч'- палатализованный согласный, а ж, ш и ц не имеют палатализованных коррелятов. Существует также долгий палатализованный ш': /ш'ч'/. Эту несимметрию русского и болгарского непалатализованного/палатализованного консонантизма можно представить графически следующим образом:

согласный непарный	рус.	болг.	согласный непарный	р.	б.
ш	+	+	дз	-	+
ж	+	+	ц'	-	+
ч	-	+	дз'	-	+
дж	-	+	ш':	+	-
ч'	+	-	йот	+	+
ц	+	+			

Согласно таблице в русском и болгарском языках одинаковы 4 непарных по признаку палатальности согласных: ж, ш, ц и йот. В отношении шипящих как артикуляторный план, так и перцептивный в русском и болгарском неодинаковы. Они различаются положением языка - рус. какуминальные ш и ж против болг. дорсальных, степенью напряженности и местом его кончика, а

также сильным выпячиванием и напряжением губ при артикуляции русских шипящих и отсутствие этих положений при образовании соответствующих болгарских. Напряженность языка и губ при артикуляции рус. ш соответствует их состоянию при произнесении болг. непалатализованного ч. В сущности, различие между этими двумя звуками сводится к способу их образования - ш чистый фрикативный согласный, тогда как ч аффриката. Но поскольку и в русской и в болгарской разговорной речи намечается тенденция произнесения аффрикаты с ослабленной смычной фазой, различие между двумя шипящими снимается. Таким образом появляется фонологическая ошибка в произнесении болгарином на месте палатализованного ненапряженного русского ч'энергичного болгарского ч, воспринимаемого русским аудитором как непалатализованный рус. ш, например: засушить вместо засучить, замешает вместо замечает, пушок вместо пучок и пр.

В "обратном направлении" протекает смешение болг. ш с рус. ш':, в котором долгота не является релевантным признаком, например: пи[ш']и вместо пиши, плю[ш'] вместо плюш, про[ш']ение вместо прошение и пр.

Гораздо более сложные отношения существуют между ц в комбинации с и и е, когда следовало бы услышать сочетание цы и цэ. Но поскольку в болгарском консонантизме имеется парный палатализованный ц', а в русском языке перед передними гласными произносятся палатализованные согласные, то обычно в такой позиции болгары артикулируют палатализованный ц', перенесенный из болгарской фонологической системы. Для русского уха такой звук непривычен и самый близкий палатализованный согласный, субституируемый русским аудитором на его месте - рус. т', что вызывается особой его артикуляцией: в отличие от его непалатализованного коррелята, артикуляция которого завершается четким взрывом, у него раскрытие смычки происходит "более плавно /как у аффрикат/, что и воспринимается примерно как аффриката ц, но не твердая, как в русском языке, а мягкая, т.е. ц'" /6/. Здесь могут возникнуть фонологические ошибки следующего типа: тех вместо цех, стены вместо сцены, телом вместо целом и пр. /7/.

ВЫВОДЫ

Различия между палатализованным/непалатализованным консонантизмом в русском и болгарском языках распространяются не только на фонетический, но и на фонологический аспект.И если недостаточное овладение болгарами первым вызывает появление болгарского акцента в их речи, то ошибки фонологического плана приводят к искажению смысловой стороны русского текста. Вот почему при обучении болгарских учащихся на различных уровнях русскому консонантизму в плане палатальности в первую очередь следует обращать особое внимание на обработку тех согласных, которые могут стать источником фонологических ошибок.

БИБЛИОГРАФИЯ

/1/ Спирова, Денка. Возможности применения критерия частотности звуков в целях обучения /на материале болгарского и русского языков/. - Балканско езикознание, XXVIII, 1985, № 3, с.64-65.

/2/ Маринова, М. и Ас. Маринов. Статистически изследвания на фонемите в българския книжовен език. - Български език, XIV, 1964, кн.2-3, с.176.

/3/ Тилков, Димитър. Акустичният ефект от палатализиращото действие на гласните е и и върху съгласните к и г. - Български език, XX, 1970, кн.2-3, с.181.

/4/ Тилков, Димитър. Сонорните съгласни в книжовния български език. - Български език, XIX, 1969, кн.6, с.513.

/5/ Тилков, Димитър. Изследвания върху българския език. София, 19831 с.142.

/6/ Матусевич, М.И. Современный русский язык. Фонетика, Москва, 1976, с.134.

/7/ Георгиева, Магдалина. Фонологические ошибки болгарских учащихся в русской речи на уровне звуков. - Болгарская русистика, 1978, № 5, с.42-48.

/8/ Стойков, Стойко. Увод в българската фонетика. 2 прераб. изд. София, 1961.

/9/ Тилков, Димитър и Т. Бояджиев. Българска фонетика. София, 1977.

/10/ Панов, М.В. Русская фонетика. Москва, 1967.

/11/ Артемов, В.А. Экспериментальная фонетика. Москва, 1956.

/12/ Бондарко, Л.В.Звуковой строй современного русского языка. Москва, 1977.

/13/ Буланин, Л.Л. Фонетика современного русского языка. Москва, 1970.

/14/ Георгиева, М. и М. Попова.Русская фонетика и интонация. 2 изд. София, 1982.

РЕАЛИЗАЦИЯ СОГЛАСНЫХ ПРИ САНДХИ (НА СТЫКЕ КЛИТИК И ЗНАМЕНАТЕЛЬНОГО СЛОВА) В РУССКОМ И БОЛГАРСКОМ ЯЗЫКАХ

ЛИЛИЯ НИКОЛОВА

Софийский университет
"Климент Охридски"
НР Болгария

1. Обоснование необходимости экспериментального исследования явлений на стыке на материале слова, фразы, сверхфразового единства.
2. Выводы о сходствах и различиях в реализации согласных на стыках слов знаменательных с проклитиками.
3. Данные фонетической интерференции в русской речи болгар при реализации проклитик и энклитик в потоке речи.

1.Понимая под сандхи позицию на стыке проклитики и знаменательного слова в рамках одного фонетического, само явление коартикуляции согласных мы рассматриваем как оглушение . озвончение согласных вследствие слитного произнесения слов в потоке речи.

Выделение в составе синтагмы фонетических слов производится не только на основе просодики (объединенность слогов вокруг одного ударного), но и по семантическому критерию. Так, проклитиками в русском и болгарском языках являются предлоги, реже составные частицы, объединяющиеся с последующей именной словоформой. Эта особенность предлогов в том и другом языках позволяет проследить изменения сегментных средств, которые становятся базой для слияния клитик с ударным словом в одну акцентно-ритмическую структуру.

Цельнооформленность фонетического слова позволяет на предварительном этапе провести внутриязыковое сопоставление таких предлогов как ПОД, НАД, ЧЕРЕЗ (русских) и соответствующих им приставок, иными словами, установить аналогию с межморфемным стыком. В целях выявления межъязыковых сходств и различий мы объединили проклитики, оканчивающиеся на согласные, и сопоставляли их по группам: предлоги на /д/-/т/ и частицы "вот","ведь";предлоги на /з/-/с/ где за исходные брались русские предлоги. Третья группа объединяет предлоги, существующие в двух вариантах в болгарском языке В и ВЪВ; С и СЪС; и предлог КЪМ в сопоставлении с русскими В;С, учитывая возможность их факультативной замены на ВО;СО;КО (1). Характеристика "наличие /отсутствие основного тона в позиции перед гласной сохраняется, и эту форму принимаем за исходную.

В зависимости от большей или меньшей слитности предлога со знаменательным словом, нами отмечены случаи полной и частичной ассимиляции, когда полоса основного тона распространяется на 2 и более согласные или же на графике отрезку повышенной интенсивности соответствует нулевой отрезок огибающей основного тона (на интограмме оглушение фиксируется как полное) равный по длине двум согласным на стыке.

Кроме этих двух случаев и нормативной реализации перед сонорными и носовыми /В/, нами прослеживаются и самые распространенные нелитературные варианты в болгарской речи.

Исследование динамического соотношения вариантов нормативной реализации и важных, с точки зрения сопоставления, отклонений от нее требовало различного подхода к материалу с учетом не только фонетического окружения согласного, но и места ударения, количество слогов в фонетическом слове и близость последнего к интонационному центру фразы.

Четырем видам выборок соответствуют 4 этапа их обработки, они в свою очередь подводят нас к исследованию интерференции в русской речи болгар.

1 этап) Сопоставление предложно-именных сочетаний в рамках отдельно взятой синтагмы - путем транскрибирования речи журналистов и дикторов Болгарского Радио и Телевидения и Центральной программы Советского Телевидения;исполнителей мастеров слова

2 этап) Сопоставление разных по коммуникативной направленности двух- и трехсинтагменных фраз, составленных на обоих языках и начитанных на пленку носителями языка (русских 3 мужчин; болгарских- 3 мужчин и 3 женщины).

3) Установление особенностей реализации в условиях, близких к спонтанной речи. Использовалось около пятидесяти диалогов и др.микротекстов, взятых из произведений современных авторов (ср.например 2).

4) В отношении стилистических особенностей использовались данные др.болгарских фонетистов (3;4). Исследовался официально/деловой стиль, представленный оригиналом и переводом на болгарский язык текста заключительной речи на 27 Съезде КПСС М.С.Горбачева. Предложенный для подготовленного чтения двум дикторам (мужчинам, нефилологам), текст несколько раз прослушивалса и анализировался аудиторами. Все записи производились в студии учебных записей на магнитофон "Revox" и затем подвергались аудитивному тесту, причем данные второго и третьего этапов анализировались и с помощью интонографа, а в 15 спорных случаях и на сонографе "КАУ".

"2"А"

Чередование /д/-/т/ в болгарских предлогах ЗАД; НАД; ПРЕД; ПОД; СРЕД; ОТ и в русских НАД;ПЕРЕД;ПОД;ОТ и частицы "ВОТ" и "ВЕДЬ". С первого же этапа выявилось отличие всех (сроме ОТ) проклитик на /Д/Т/. Случаям озвончения перед гласными, сонорными, шумными звонкими соответствует в болгарском реализация с глухим на конце. Интограммы (2ой и 3ий этап) подтвердили это, причем для болгарского материала немаловажно учитывать отклонение от нормы при стыке /от/+/в/ у $2^х$ дикторов наряду с нормативными встречались сочетания типа "о/д/ вашия" ;о/д/ малка. Для них характерен быстрый темп проговаривания, связанный с удаленностью этих фонетических слов от интонационного центра фразы.

Если в индивидуальном стиле говорящего такое озвончение происходит под влиянием диалектного произношения, то оно затрагивает и другие предлоги этой группы, а в русской речи болгар ожидается интерференция в области стыковых явлений и в частицах ВОТ - ВОД и ВЕДЬ /д'/.

В целом русским предлогам со звонким /д/ на конце соответствует глухой /т/ в болгарском. Это нашло подтверждение в реализации на уровне текста во фразах типа "Все зависит от‿нас, товарищи" - "Всичко зависи от‿нас, другари"(5)

"2"Б" Последовательное проведение трех этапов экспериментального анализа с попарным сравнением каждого сочетания, синтагмы фразы дало следующую картину изменений:

- Перед гласными, сонорными и /В/ русские и болгарские предлоги могут реализоваться сходным образом - с полным или частичным озвончением -, но для болгарского число таких случаев ограничено, такова реализация в сочетаниях, удаленных от центра. Например, "Пре/з'/‿изми́налото денонощие над по-голямата част от страната преваля и прегърмя". Или "Високите добиви се постигат чре/з/‿въве́ждане на нови технологии". Наиболее часты случаи непересечения между фонетическим обликом русского и формально (графически) соответствующего ему болгарского предлога на /З'С/ - русский предлог реализуется по типу соответствующей приставки - на конце озвончается под влиянием последующего звонкого или сонорного, или /В/. В болгарских же предлогах этой группы нормативно происходит оглушение на конце, а озвончение является лишь вариантом, например "чре/з/‿вди́гане на‿ръце́" или "бе/з/‿внима́ние". Эти случаи представляют комплексную артикуляцию трех согласных. Необходимо учитывать возможность побочного ударения на предлоге (6) "БЕЗ", что придает ему больший коммуникативный вес (ср.также частицу "бе́з‿да": по-русски смысл ее выражается деепричастием с НЕ). Для восприятия предлогов как более слитных с мменной формой, важную роль играет уподобление по месту и способу артикуляции."2"В" (7) Наибольшие различия в нормах наблюдаются в предлоге /В/ и соответствующем ему болгарском /Ф/-/В/: звонкий вариант, в отличие от русского предлога, возможен только перед звонким согласным. Характерное удвоение болгарского предлога происходит в речи:

- перед именными словоформами, начинающимися на /В/ или /Ф/, как этого требуют нормативные грамматики (3;8);
- в абсолютном начале фразы перед стечением из двух согласных;
- при акцентном выделении всего фонетического слова в целом, также при паузе хезитации; это придает речи характер непринужденности; во всех этих случаях реализуется предлог /ВЪФ/ с глухим согласным и часто с ъ-образным призвуком.

Аналогично и в предлоге /СЪС/ оглушение на конце сопровождается удлиненной окклюзией воспринимаемой как Ъ-образный призвук. Такой же призвук появляется в русской речи болгар при реализации всех трех русских предлогов В;С;К.

Итак, под влиянием частотных реализаций с глухим согласным на конце (по нормам болгарского языка) в русской речи учащихся наблюдаются следующие виды интерференции:

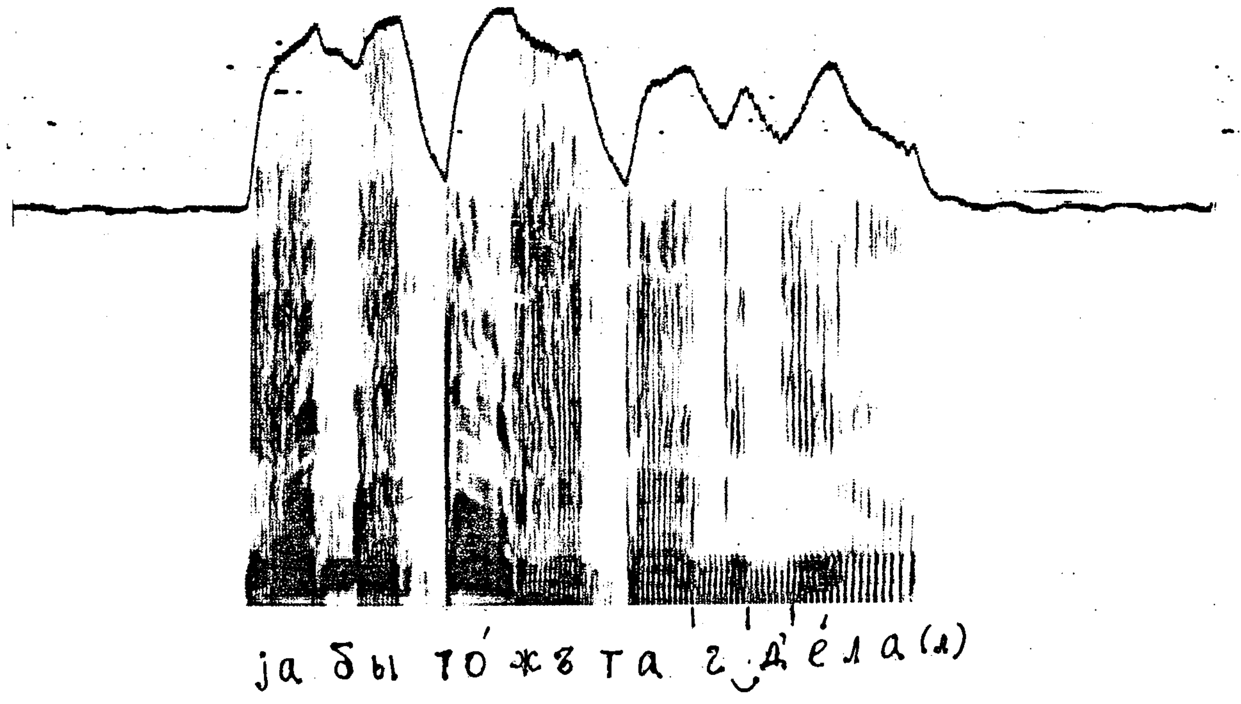
- оглушение конечных согласных в предлогах ПОД; НАД; ПЕРЕД;БЕЗ;ЧЕРЕЗ;ИЗ;СКВОЗЬ;БЛИЗ;
- озвончение под влиянием диалектного произношения /от/ и по аналогии с ним частиц ВОТ, ВЕДЬ и др. (9)
- недифференцирование вариантов русских предлогов В;С;К, неправильное озвончение С/з/‿ ва́с, /З/‿вино́м, под влиянием диалектной болгарской нормы и неправильное оглушение предлога /В/:/Ф/ начале; "име́ть /Ф/ ви́ду́".

Аудитивный эксперимент показал, что в записанной на пленку речи студентов русистов ошибки в реализации сандхи составляют около 52-55% всех ошибок и трудностей в усвоении русской фонетики и интонации. (10) Признак глухости/звонкости согласных при сандхи играет важную роль в восприятии односложных клитик, из которых при быстром произнесении могут выпасть гласные, в этом нас убеждает сопоставление двух сонограмм спонтанной речи чтеца-нефилолога; из-за технических трудностей, удалось записать лишь десятиминутный разговор в студии(См.) /участники не знали, что разговор записывается/.

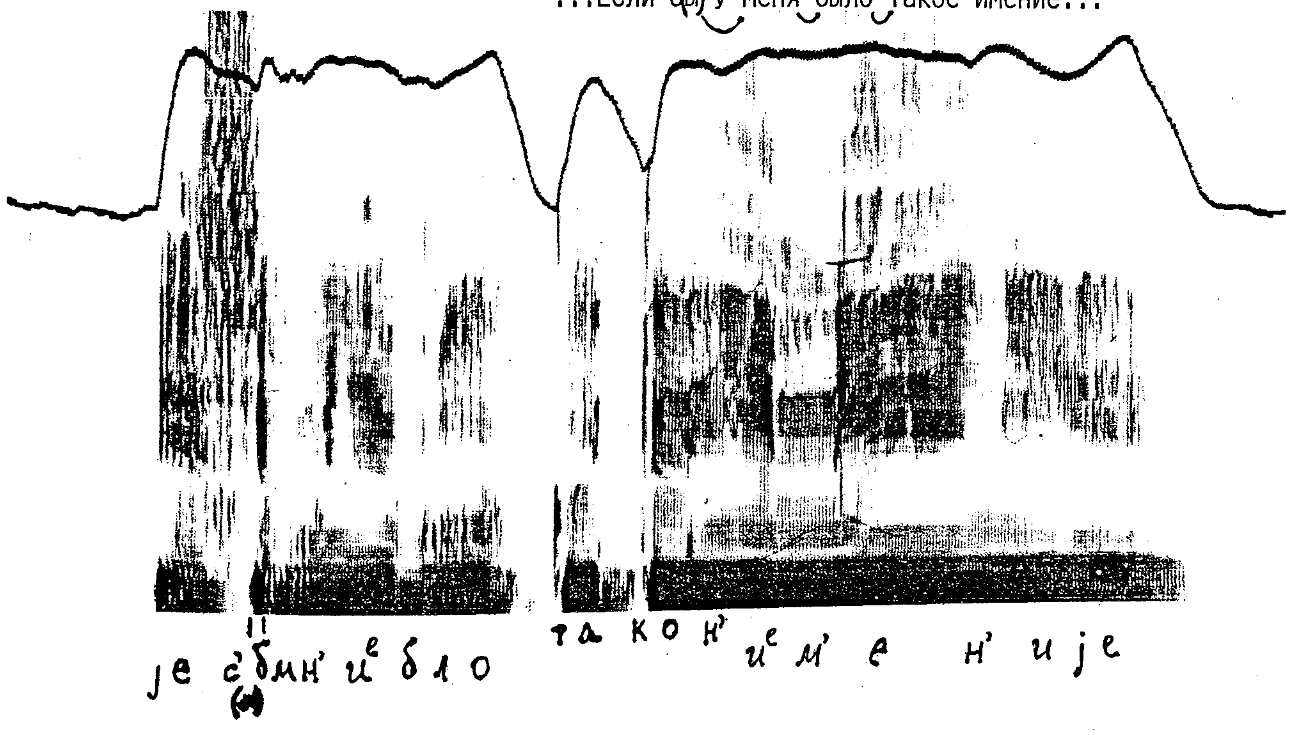
Выводы: при всех сходствах в реализации русских и болгарских клитик, необходимо дифференцировано подходить к их изучению в целях преодоления интерференции на уровне фонетической реализации синтагм и фраз. Основные отличия между русскими и болгарскими предлогами состоит в бо́льшем выделении последних путем оглушения конечного согласного, а для предлогов С и В путем их удвоения.

См. сонограммы разговора в студии:

Я бы тоже так делал



...Если бы у меня было тако́е име́ние...



(с.44, 63, 79, 197)

1) П.П.Рогожникова "Варианты слов в русском языке" М."Просвещение", 66 г. стр. 42-43; 48-50

2) Г.Данаилов "Деца играят вън" "Убийството на Моцарт" С. "Български писател" 86 г. стр.83,175 и др.Ю.Бондарев "Мгновения" (Знакомство в летний дождь) - Собрание сочинений т.6

3) Д.Тилков, Т.Бояджиев "Българска фонетика" 1977 стр.211-212

4) А.Слуцка "Асимилативни влияния и промени при съгласната /в/ в условията на междусловен допир в Български книжовен език"-ж."Език и литература" - С. 1986 г. №2 стр.71-72

5) М.С.Горбачов. Заключительная речь на XXVII съезде КПСС ж."Коммунист" №5/86 г; XXVII конгрес на КПСС. Партиздат, София 1986 г. (коллектив переводчиков) ред. Ст.Даковска

6) Т.М.Стоева "Явления сандхи в ритмомелодической организации синтагмы русского и болгарского языков" в кн. "Сборник доклади, Конференция по фонетика и фонология на славянските езици" 18-21 септември 1984, София

7. С.Н.Борунова "Сочетания /ш'ч'/ и /ш'/ на границах морфем" в сб. "Развитие фонетики современного русского языка" под ред. С.С.Высотского М."Наука", 66 г. стр.58; 67

8. Ст.Младенов, Ст.Василев "Граматика на българския език" С. 39 г. стр.100; Л.Андрейчин "Основна българска граматика" С.44 г. стр.50, второе изд. С. "Наука и изкуство" 1979 г. стр.47; П.Пашов, Хр.Първев "Правоговорен речник" С.1979 г. стр.9

9. Т.Стоева "Система суперсегментных фонологических средств русского и болгарского языков":Вопросы сопоставительного описания русского и болгарского языков" изд. "Наука и изкуство" С. 1982 г.

10. Ю.Т.Лебедева "Методика преподавания русского произношения болгарам" М.1973 г.

АРТИКУЛЯЦИОННАЯ БАЗА (АБ) РУССКОГО ЯЗЫКА В СОПОСТАВЛЕНИИ С АБ ХИНДИ И ЕЁ ТИПОЛОГИЧЕСКИЕ ОСОБЕННОСТИ.

ЧИНТА БАЛУПУРИ

Центр Русских Исследований
Университет им. Неру.
Нью Дели. Индия

На основе сопоставления АБ русского языка и хинди и других сопоставительных исследований, а также данных по общей фонетике выводятся типологические особенности русской АБ.

Для типологических исследований в широком смысле слова важно сопоставление разносистемных языков. Сопоставление языков приобретает особое значение при описании артикуляционной базы того иль иного языка, ибо "понятие артикуляционной базы может быть выведено только на почве сравнения артикуляторных навыков разных языков" /I/.

Фонетическая система русского языка сопоставлялась с фонетическими системами ряда языков и в зависимости от того, на фоне какого языка она рассматривалась, выявлялся определенный набор особенностей исследуемого языка, которые внесли несомненный вклад в определение АБ русского языка. Однако, специальных исследований, посвященных типологии русской АБ, за исключением отдельных работ, не проводилось. /2/

В данной работе сделана попытка на основе сопоставления АБ русского языка и хинди, а также других сопоставительных работ с привлечением данных общей фонетики описать АБ русского языка с точки зрения её типологических особенностей. Здесь также важно учесть тот факт, что типологическая характеристика предполагает не только межъязыковой, но и внутриязыковой анализ.

Под АБ языка понимается система привычных для данного языка положений и движений органов речевого аппарата, а также систему переключений с одного типа артикуляции на другой. АБ представляет собой стереотип, работы органов речи, необходимый для произношения звука, слова предложения. АБ определяется фонетической системой языка и её функционированием. В связи с этим АБ можно рассматривать в статическом и динамическом аспектах.

При сопоставлении рассматриваемых языков в статическом плане на уровне звуков учитывались особенности работы отдельных органов, сосредоточенность артикуляций и степень их выраженности и т.д.

Известно, что для образования гласных резонаторами служат полости рта и глотки. В отличие от русского языка,где отсутствуют носовые гласные, в хинди при образовании назализованных гласных в качестве резонатора подключается и носовая полость. Противопоставление в его фонетической системе носовых и неносовых гласных ([ā] - [ā̃],[ī] - [ī̃] и др.) активизирует работу мягкого нёба.

В хинди также более активны губы, т. к. в нем по сравнению с русским языком превышает количество лабиализованных гласных: [o] ,[ɔ],[u] , [ū] . Однако, при образовании русских лабиализованных [у], [о] губная артикуляция более напряженная, чем в хинди.

Артикуляция гласных предполагает движение языка по горизонтали и вертикали. Хотя и нельзя не согласиться с Л.Р. Зиндером в том, что "... гласные не могут быть точно локализованы в речевом аппарате ..." /3/, движение языка в хинди более дифференцировано, т.к. в нем по сравнению с русским языком наблюдается большая градация гласных по подъему и ряду. Ср: [i], [u] и др. открытые и [ī], [ū] и др. закрытые.

Кроме того, гласные в хинди различаются по долготе и краткости. Длительность артикуляции служит дифференциальным признаком: [kul] - [kūl], [sil] - [sīl]

В русском языке длительность артикуляции гласных обусловлена их расположением относительно ударного слога-удлинение ударного гласного и редукция безударных. Не будучи дифференциальным признаком длительность артикуляции гласного, однако, может быть использована для создания эмоциональных оттенков: "замечательно" и "замеча-ательно".

В образовании согласных звуков русского языка из четырех резонирующих полостей речевого аппарата участвуют только полость рта и носа. В хинди же, кроме полостей рта и носа, активнуюрроль играет полость глотки. Хотя в языке имеется единственный фарингальный звук [h], частотность его употребления в значительной степени активизирует работу мышц глотки. В хинди также встречаются звуки, образующиеся в полости гортани - [q], [ʁ], [x], однако они были заимствованы языком вместе с арабо-персидской лексикой и не получив статуса фонем, часто заменяются в речи собственно хиндийскими [k], [g], [k^h] соответственно, что снижает роль гортанной смычки.

В отличие от хинди в русском языке глоточных и гортанных звуков нет, однако в потоке речи при произношении сочетаний взрывного согласного с носовым вместо разрыва первого согласного нёбная занавеска отрывается от задней стенки носоглотки создавая "фаукальный взрыв" в полости глотки: [плóтный], [ʌбмáн] /4/.

Нагруженность губ произносительными работами при образовании согласных звуков характерна как для АБ русского языка, так и хинди: [б] - [б], [м] - [м], [п] - [п] и [b] - [b^h], [p] - [p^h], [ш]-[ш^h].

В обоих рассматриваемых языках большинство артикуляций согласных связано с работой языка. В русском языке особое место занимают переднеязычные артикуляции, которые составляют 3/4 всех язычных артикуляций, в то время как в хинди они составляют менее 2/3 всех язычных артикуляций. Обилие переднеязычных согласных, а также отсутствие глоточных и гортанных артикуляций в русском языке свидетельствует о том, что АБ русского языка в целом продвинута вперед по сравнению с АБ языка хинди.

При артикуляции переднеязычных согласных для русской АБ характерен общий дорсальный уклад языка ([т] , [д] , [з] и др.), в то время как в хинди наблюдается апикально-какуминальный уклад языка ([ṭ] , [ḍ] , [ṭ^h] и т.д.). Хотя в русском языке есть и какуминальные ([ж] , [р] и др.), и апикальные ([ч'], [ж̄'] и др.) согласные, выпуклось языка с опущенным кончиком является важнейшим укладом в АБ русского языка, т.к. он способствует палатализации согласных. На дорсальных артикуляциях базируется категория твердости/мягкости.

И для русской, и для хиндийской АБ характерны разные способы образования шума - смычные, щелевые, дрожащие и др. В русском языке количество щелевых преобладает по сравнению с хинди, в то время как АБ хинди богаче смычными артикуляциями.

В отличие от хинди, где фонологически противопоставлены придыхательные и

непридыхательные согласные, в русском языке выдох воздушной струи прекращается сразу после разрыва смычки. Однако в потоке речи возможно появление придыхательности в русских согласных, что не обусловлено ни фонетической системой языка, ни её функционированием, т.к. придыхательность - непридыхательность не является дифференциальным признаком, но с другой стороны, именно потому что аспирация не является различительным признаком возможно её появление: [вот^h]. Это же относится и к имплозивности русских согласных в потоке речи.

Как в русском языке, так и в АБ хинди важную роль играет работа голосовых связок. В обоих языках голосовые связки вибрируют при образовании гласных, сонорных и звонких согласных. В обоих языках голосовые связки вибрируют на протяжении всей артикуляции согласного.

При образовании глухих шумных голосовие связки пассивны. В обоих языках глухие согласные являются более напряженными, чем звонкие и имеют большую площадь соприкосновения активного и пассивного органов речи. Глухие согласные звучат дольше соответствующих звонких.

Механизм твердости-мягкости является универсальной чертой АБ русского языка, отличающая его от АБ хинди и целого ряда других языков. При продвижении всего тела языка вперед (палатализация) и его оттягиванием назад (веляризация) создается осооая тембровая окраска русских согласных. Эта типичная черта русской АБ охватывает всю консонантную систему языка , хотя не все согласные противопоставлены по твердости - мягкости. Так, С.А. Бaрановская отмечает, что непарные по твердости - мягкости согласные [ж, ш, ц] имеют подъем и напряженность задней части спинки языка, что свойственно механизму палатализации/веляризации /5/.

Механизм твердости - мягкости охватывает и динамический аспект русской АБ.

Перемещение языкового тела вперед и назад сопровождается подъемом и напряженностью мышц то средней, то задней части языка. Типичним для механизма палатализации - веляризации является его распространение на соседние гласные придавая им [и] - или [ы] - образную окраску: "сядь" - [с'^{и}а^{и}т'] и "сад" - [с^{ы}а^{ы}т]. Слабое примыкание, плавный переход от артикуляции согласного к артикуляции гласного обусловливают качественную неоднородность гласных в русском языке. Дифтонгоидность является характерной особенностью динамики русских гласных. Для русского слова также характерно переключение с переднеязычной артикуляции на зднеязычную и наоборот. Легкое движение языкового тела вперед и назад создает две области произносительных работ, связанных с тембровыми изменениями.

Особенность АБ русского слова заключается и в разнообразии чередований переднеязычных и заднеязычных щелевых согласных: восход - [сх].

Для обоих сопоставляемых языков свойственно лёгкое переключение с глухости на звонкость и наоборот, т.е. в пределах слова голосовые связки легко могут прекращать и возобновлять вибрацию: к рт'ин и [kəv əč].

Однако, законы реализации фонологической системы хинди не допускают таких явлений, как оглушение звонких согласных в конце слова, ассимиляция по глухости - звонкости и уподобление согласных по месту и способу образования, которые типичны для артикуляции русского слова.

АБ русского слова также отличается контрастным противопоставлением ударных и безударных слогов по длительности и напряженности артикуляции. Кроме названных признаков, ударный слог выделяется своим качеством - отчетливостью тембра, усилением мышечной напряженности. Безударные же слоги подвергаются количественной и качественной редукции. В русском языке количественная редукция усиливается в слогах, отстоящих от ударения.

В хинди же, где ударение выражено слабо, контрастно противопоставлены долгие и краткие гласные, причем краткие гласные могут быть ударными, а долгие-безударными. Длительность артикуляции гласных в хинди позиционно не обусловлена. В нем также не наблюдается редукции безударных слогов.

Характер словесного ударения, специфика выделения ударного слога, ритмическая организация слова, обусловленные фонетической позицией изменения гласных и согласных сказываются и на АБ русского предложения.

Как в русском языке, так и в хинди предложение может состоять из одного или нескольких слов. В последнем случае наиболее важное в смысловом отношении слово интонационно выделяется, что формирует интонационный центр предложения, который в обоих языках может находиться в начале, середине и конце слова, создавая различное соотношение предцентровой, центровой и постцентровой частей. Интонационный центр выделяется изменением компонентов интонации-тона, длительности, интенсивности и тембра, при этом наиболее значимые изменения в обоих языках происходят в области частоты основного тона. Интонационный центр в русском языке особенно выделяется отчетливостью тембра.

Для АБ русского предложения характерно переключение с большой частоты колебания голосовых связок на малую и наоборот, что создает контраст между тональными уровнями предцентра, центра и постцентра. Русский язык характеризуется широким диапазоном изменения частоты тона /6/.

АБ предложения в хинди в отличие от русского языка свойственен плавный переход от низкого тонального уровня к высокому и наоборот. Интонационный центр может распространяться на более, чем один слог, т.е. значимые изменения в частоте колебания голосовых связок могут охватывать часть предложения /7/.

Ритмика русского предложения определяется чередованием ударных/безударных слогов, в то время как в хинди она базируется на количественном противопоставлении долгих и кратких слогов.

Проделанный выше анализ даёт возможность выявить следующие особенности русской АБ: продвинутость вперед, дорсальный уклад языка, механизм твердости/мягкости и его распространение на соседние гласные - дифтонгоидность гласных, чередование переднеязычных и заднеязычных артикуляций, легкость переключения с твердых на мягкие согласные и наоборот, переключение с глухих на звонкие и наоборот, позиционное оглушение звонких согласных, ассимиляция по звонкости/глухости, контрастное выделение ударного слога и редукция безударных слогов. На уровне предложения АБ русского языка отличается контрастным выделением интонационного центра, широким диапазоном изменения частоты основного тона, резким переключением с большей частоты на малую и наоборот.

Данные сопоставлений фонетической и фонологической систем русского языка и других языков, а также данные общей фонетики дают основание отнести к типологическим особенностям русской АБ и следующие её черты /8/: отсутствие назальных гласных, т.е. относительно небольшая загруженность носовой полости, отсутствие глоточных и гортанных артикуляций, неразличение гласных по длительности их звучания, полнозвонкость согласных, т.е. вибрация голосовых связок на протяжении всей артикуляции звонких согласных, отсутствие противопоставления согласных по степени мускульного напряжения органов речи - сильных/слабых согласных; отсутствие имплозивности, т.е. русским согласным свойственны все три фазы артикуляции. На просодическом уровне АБ русского языка характеризуется легким переключением с большей длительности и мускульного напряжения на меньшую и наоборот. Просодия русского слова ввиду нетональности языка подчиняется интонации.

/I/ С.И. Бернштейн. Вопросы обучения произношению (применительно к преподаванию русского языка.иностранцам). - "Вопросы фонетики и обучение произношению". Под ред. А.А. Леонтьева и Н.И. Самуйловой. М., 1975, с. 22.

/2/ См. С.А. Барановская. Характерные особенности артикуляционной базы русского языка - "обучение русскому произношению студентов - иностранцев на начальном этапе". М. 1979.

Е.А. Брызгунова. Возможности речевого аппарата и артикуляционная база русского языка. - "Практическая фонетика и интонация русского языка". М., 1963.

/3/ Л.Р. Зиндер. Общая фонетика. Л., 1960, с. 206.

/4/ См. А.А. Акишина, С.А. Барановская. Русская фонетика М., 1980, с.с. 23-24.

/5/ См. С.А. Барановская. Соотношение артикуляционной базы языка с системой и нормой. - "Методика преподавания русского языка как иностранного". М., 1977.

/6/ См. Е.А. Брызгунова. Звуки и интонация русской речи. М., 1977.

/7/ См. Ч. Балупури, Ю. Ковалев. Фонетика русского языка. Нью Дели, 1984.

/8/ См. В.Н. Денисенко. Сопоставление гласных фонем русского и французского языков.- "Фонетические единицы речи". М., 1982.

Н.А. Любимова. Обучение русскому произношению. М., 1977.
В.И. Петрянкина. Функциональный аспект интонации и типология языков. - "Просодия
слога - слова - фразы." М., 1981.
Н.И. Самуйлова. К вопросу об акценте. - "Памяти В.В. Виноградова". М., 1971.
И.О. Прохорова. Особенности восприятия твердости-мягкости согласных русского я-
зыка латиноамериканцами. - "Проблемы теоретической и прикладной фонетики и обуче-
ние произношению". М., 1973.

A.R. Kelkar. Studies in Hindi - Urdu. Introduction and word phonology. Poona, 1968

M.G. Chaturvedi. A contrastive study of Hindi - Urdu phonology. Delhi, 1973.

T.R. Anderson. A case for contrastive phonology. IRAL, 1964, v II, No. 3.

GENERATIVE ACCOUNTS OF MISARTICULATIONS OF TWO JORDANIAN CHILDREN

FARES MITLEB

Department of English
Yarmouk University
Irbid, Jordan

ABSTRACT

This study describes the misarticulations of two Jordanian children with a generative framework. Accordingly, each child's phonological system is accounted for via some context-free inventory constraints and phonological rules. It is claimed that misarticulators possess differential knowledge which is not identical to the ambient system. Also, markedness violations are observed in each child's system. We furthermore putforward hypotheses with regard to the ease and difficulty of unlearning the deviant speech habits in favor of the normative data. We thus provide the necessary information for speech therapists to devise remedial programs for speech misarticulating children. This claim, however, could be tested clinically. [Research supported by Yarmouk University, Jordan].

INTRODUCTION

This paper examines the speech of two functionally misarticulating Jordanian children and illustrates the contribution of generative phonology to phonological descriptions. The term 'functional misarticulation' is typically used to describe the speech of speakers whose chronic articulatory errors cannot be attributed to any obvious organic disorders such as hearing impairment or cleft palate [7, 1]. The basic assumption of most of the work done on speech misarticulations is that children's knowledge is identical to that of the ambient speech community [2, 9]. Within this framework, misarticulators are viewed as a homogeneous group. However, the many diverse phonological rules that have been posited to change the underlying structure into misarticulators' surface structure makes it difficult to arrive at any but gross commonalities across functional misarticulation systems [7]. Any discrepancy between the misarticulators' system and the ambient system is described as a 'process' [12]. Such an assumption is a clear misrepresentation of the apparent differences across the misarticulation systems [6]. Recent development in the literature has shown an increasing interest in employing the generative framework to further characterize misarticulations in children [5, 8, 10]. Within this framework, misarticulators are classified into groups depending on the severity of the problem and the markedness violations [8]. Thus, phonemic inventory constraints are placed on children's productions and phonological rules are posited to convert misarticulators' underlying representations into their phonetic production. The purpose of this paper is to further support the claims of generative phonology to account for misarticulations.

METHOD

Two female Jordanian children, aged 7:2 (Child 1) and 7 (Child 2) years, served as subjects in the present study. Purely spontaneous speech samples were collected from the two children by eliciting certain alternations making use of picture naming, friends and family naming, and questioning the children. The two children were enrolled in regular schools in the second grade. They were referred to the research for speech remediation.

PHONOLOGICAL ANALYSIS

Child 1, age 7:2, produces 14 consonants of the 28 ambient language phonemes. Among the non-strident obstruents, she produces [b, t, d, ħ, ʕ, h]. The non-stridents [ṭ, ḍ, k, q, θ, ð, ð̣, x, ɣ] are never produced in any position, as can be gathered from the following forms:

Child			Ambient		
tawīl	ħatab	ħat	ṭawīl	ħaṭab	ħaṭ
dāl	mudal	ħad	dar	muḍar	ħad
tum	zuteyt	mafat	kum	ǰukeyt	mafak
dalam	?idlāmi	ħalad	qalam	?aqlām	ħalaq
tum	?ittīl	talāt	θum	?ikθīr	θalāθ
dāl	?adān	muʕad	ð̣āl	?aðān	muʕað̣
ħālid	munħal	muħ	xālid	munxar	mux
ħadah	?almahlib	dahdah	ɣadah	?almaɣlib	daɣdaɣ

These forms, support a claim that this child represents her morphemes underlyingly without non-anterior stops and low fricatives. This fact is described by a general context-free inventory constraint:

$$\begin{bmatrix}\text{-sonorant}\\ \\ \text{-strident}\end{bmatrix} \longrightarrow \left\{\begin{array}{l}\begin{bmatrix}\text{-continuant}\\ \text{+anterior}\end{bmatrix}\\ \text{-high}\end{array}\right\} \begin{array}{l}\text{a. [b, t, d]}\\ \\ \text{b. [ħ, ʕ, h]}\end{array}$$

(All non-stridents are either anterior stops or low fricatives)

As for strident obstruents, this child shows knowledge of anteriors [f, s, z] whereas the non-anteriors [š, ṣ] are never produced, as illustrated by the following forms:

Child			Ambient		
sāmi	?aslaf	?iflās	šāmi	?ašraf	?ifrāš
sūlah	?asfal	bas	ṣurah	?aṣfar	baṣ

To describe this fact, a second context-free inventory constraint is proposed:

$$\begin{bmatrix}\text{-sonorant}\\ \text{+strident}\end{bmatrix} \longrightarrow [\text{+anterior}]$$

(All stridents are anteriors, i.e. f, s, z)

With regard to liquids, Child 1 never produces [r]. Notice, for example, the following forms:

Child			Ambient		
lās	tundalah	dāl	rās	kundarah	dār
lam	bilāl	dal	lam	bilāl	dal

Thus, a third context-free inventory constaint is postulated to limit liquids to [l]:

$$\begin{bmatrix}\text{+consonantal}\\ \text{+sonorant}\\ \text{+continuant}\end{bmatrix} \longrightarrow [\text{+lateral}]$$

Child 1 sometimes deletes obstruents word-finally, as illustrated by the following:

Child			Ambient		
bawā	?izlāba	ħadā	bawāb	?iǰrabāt	ħadād
tufā	fara	ħalū	tufāħ	faras	xarūf

However, the child's speech shows obstruents in word-final position as in the following forms:

Child				Ambient			
?aħad	lās	luz	ħalaf	?axað̣	rās	ruz	xalaf

These forms suggest that the child's morphemes are represented underlying with obstruents in all positions. Thus, an optional rule that deletes obstruents word-finally is proposed:

$$[\text{-sonorant}] \longrightarrow \emptyset / \ \text{——}\ \# \quad \text{Opt.}$$

A second collapse of obstruent contrast is found in the speech of this child. Obstruents are optionally devoiced word-initially, as exemplified by the following forms:

Child		Ambient	
tub	subdih	dub	zubdih
tanab	talabni	ðanab	ḍarabni

However, voice contrast is observed in all positions. The following forms illustrate this fact:

Child		Ambient	
tam	sāl	kam	sār
dam	zāl	dam	zār

Therefore, the presence of voicing contrast in all positions motivates representing her morphemes underlyingly with voiced and voiceless obstruents. The devoicing process is accounted for by the following optional rule:

$$[\text{-continuant}] \longrightarrow [\text{-voice}] \ / \ \# \ \text{——}$$

To conclude, Child 1 has a limited knowledge of ambient phonemes indicated by the absence of /ṭ, ḍ, q, k, ǰ, θ, ð, š, ṣ, ð̣, x, ɣ, r/. To account for the absence of these consonants, three context-free inventory constraints are proposed to limit non-strident obstruents to anterior stops and low fricatives, stridents to anteriors, and liquids to the lateral [l]. Two phonological rules are motivated to optionally delete word-final obstruents and neutralize voice contrast word-initially. The second child, age 7 years, produces the obstruents [b, t, d, f, θ, ð, s, z, š, ṣ, ð̣, ħ, ʕ, h] and the sonorants [m, n, y, w]. The stops [ṭ, ḍ, q, k, ǰ] are never produced in any position, as can be seen in the following forms:

Child			Ambient		
tāwayih	batah	bat	ṭāwlih	baṭih	baṭ
difdāħ	madyab	baʕut	difdaʕ	maḍrab	baʕuḍ
dayam	?iwdiwiti	warat	qalam	?iwqwiti	waraq
tāsih	matatih	mafat	kasih	makātih	mafak
zamay	šazayih	ħas	jamal	šajarih	haǰ

Morphophonemic evidence given by these examples supports a claim that the child's underlying stop phonemes are limited to [b, t, d]. This knowledge is described by the following context-free inventory constraint:

$$[\text{-continuant}] \longrightarrow [\text{+anterior}]$$

i.e. b, t, d

As for continuants, the child does not show any knowledge of [x, ɣ] in any position, as illustrated by the following forms:

Child			Ambient		
ħatim	suħnih	wayħatmih	xatim	suxnih	walxaṭmih
ʕeym	duʕyi	fayiʕ	ɣeym	duɣri	fariɣ

Therefore, we postulate the following context-free inventory constraint that limits the non-anterior non-coronal fricatives to low ones:

$$\begin{bmatrix}\text{+continuant}\\ \text{-anterior}\\ \text{-coronal}\end{bmatrix} \longrightarrow [\text{-high}] \quad \text{i.e. ħ, ʕ}$$

With regard to sonorants, this child never produces [l, r]. Notice, for example, the following forms:

Child		Ambient	
yuʕbih	tayib	luʕbih	ṭalib
yamat	?izyiʕih	ramad	?izriʕih
musaziy	fay	musaǰil	far

A third context-free inventory constraint is proposed to account for the absence of liquids from the child's phonemic system:

$$\begin{bmatrix}+\text{sonorant}\\ -\text{nasal}\end{bmatrix} \longrightarrow \begin{bmatrix}-\text{syllabic}\\ -\text{consonantal}\end{bmatrix}$$

i.e. w, y

Thus far, we have established that child 2 exhibits three context-free inventory constraints that limit production of stops to anteriors, non-anterior non-coronal fricatives to low consonants, and prohibit production of liquids. To further characterize this child's phonological system, it is necessary to turn to some of the phonological processes motivated in her speech.

Child 2 devoices coronal obstruents word-finally, as illustrated by the following examples:

Child			Ambient		
?aswat	mos_	ħafiθ	?aswad	moz	ħafið
nafiθ	marit	——	nafið	marid	

Initial and medial voice contrasts are observed in the speech of this child as exemplified by the following forms:

Child			Ambient		
tum	zeyn	ða	kum	zeyn	ða
dam	sīn	θa	dam	sīn	θa
ħata	bizzi	?iððanab	ħata	bizzi	?iððanab
ħada	bissih	?iθθahlab	ħada	bissih	?iθθaʕlab

We, thus, propose that this child represents morphemes underlyingly with both voiced and voiceless coronal obstruents. A neutralization rule that devoices coronal obstruents word-finally is motivated:

$$\begin{bmatrix}-\text{sonorant}\\ +\text{coronal}\end{bmatrix} \longrightarrow [-\text{voiced}]\ /\ \_\_\ \#\ \text{Oblig.}$$

A second phonological process devoices the non-high back voiced pharengeal /ʕ/ post-vocalically, as exemplified by the following forms: maħāt sāħih maʕāk sāʕih

However, this consonant is observed in other positions:

ʕabdayah luʕbih ʕabdallah luʕbih

Morphophonemic evidence of this type supports a claim that /ʕ/ is posited underlyingly in all positions. Therefore, a neutralization rule that changes /ʕ/ into its voiceless counterpart [ħ] in post-vocalic position is proposed, specificallv

$$[+\text{low}] \longrightarrow -\text{voice}\ /\ V\ \_\_\ \text{Oblig.}$$

DISCUSSION

Examining the generative accounts of the phonological systems of the two children, we, first, realize an apparent violation of markedness. Thus, the two children produce the voiced alveolar stop /d/ for the voiceless uvular stop /q/. Child 1 devoices stops word-initiallv, but medial and final contrasts are preserved. Recall, Child 2 devoices the laryngeal fricative /ʕ/ word-medially whereas she uses /ʕ/ word-initiallv and-finallv. In both cases of devoicing, voiced segments are posited underlying. Violations of markedness seem to serve as one factor in characterizing the population of functionally misarticulators {8]. That is, phonological systems that evidence markedness violations are classified as "deviant" whereas other systems are said to be "delayed" [3]. The degree of deviancy is based on the severity of markedness violation.

These violations may question the assumptions of the typological-based implicational universals proposed by Dinnsen and Eckmen [4]. For example, implicational universals predict that the presence of voice contrast word-finallv implies the presence of this contrast word-mediallv and word-finallv. Child 1 represents a counter example to this prediction, Recall, Child 1 evidences stop voice-contrast word-medially and word-finallv, but not initiallv. Child 2 also violates the predictions of implicational universals; she produces /ʕ/ word-initially and-finallv but she changes /ʕ/ into its voiceless counterpart /ħ/ word-medially. To offer a generative account for each child's knowledge, non-ambient underlying representations are posited and inventory constraints are placed to restrict production of certain consonants that would be otherwise used by normal speakers. However, it could be argued that the two children possess ambient-like underlying representations, but production errors relative to the normal speakers' system are still present [8]. To account for such errors, we can entertain two factors [8]. First, the children have underlying structure identical to the ambient system in certain positions but not in others. Child 1 uses voiced and voiceless stops in word-medial and-final positons but not word-initially. Child 2 shows voice contrast of /ħ-ʕ/ in word-initial and-final but not intervocalically. The second factor is to assume that the child uses rules alternating ambient-like representations. Thus, Child 1 uses a phonological rule that optionally deletes obstruents word-finally. This implies that this child does not extend her knowledge of contrast to all morphemes. According to the approach that takes these two explanatory factors of speech misarticulation, we need to account for many processess of substitution and deletion which are found in the childs' phonological system in reference to the normative data. However,

it is a costlv account due to the manv rules and features needed to describe such processes. Also, it is not always easy to justify these processes within the framework of naturalness.

Within the framework of generative phonology, we can give specific characterization of the phonological systems of functional misarticulators by formulating context-free inventory constraints that specify each child's knowledge of the ambient-like system. Also, underlying representations specific to misarticulators can be posited to account for the phonological processes in each childs' system. The generative accounts enable speech therapists to design remediation programs based on the actual knowledge of misarticulators. Accordingly, it is hypothesized that misarticulating children will find it easier to learn the sounds and sound contrasts they have knowledge of than those of which they have no knowledge [5, 11]. Based on this assumption, it is expected that differences in knowledge among misarticulators result in differences in learning during speech therapy [8]. Taking the generative accounts into consideration, therefore, we could predict that Child 1 will find /b, t, d, ħ, ʕ, h, f, s, z, l/ easier to learn than /ǰ, k, q, θ, ð, ṭ, ḍ, ð̣, x, ɣ, r/ and the voice-contrast in stops word-medial and-final positions is easier than in initial-position, although medial and final contrasts are typologically more marked than initial ones. For Child 2, it could be hypothesized that fricatives are easier than the stops /ṭ, ḍ, q, k, ǰ / and the liquids /r, l/. Recall, stops are typologically less marked than fricatives. Also, this child is predicted to have no difficultv with voice-contrast word-initially and word-medially but would encounter diffibulty in the acquisition of coronal voice-contrast word-finallv which is more marked than other positions.

CONCLUSION

In conclusion, generative accounts are provided for the speech of two misarticulating children. These accounts contribute to a better understanding to the role of generative phonology to further characterize the knowledge of misarticulators of the ambient system. Descriptions of each child's phonological system are evaluated against the predictions of implicational universals. Thus the severity of the misarticulation problem can be measured according to the extent of markedness violations. Moreover, we point out that the generative framework furnishes speech therapists with the necessarv information to devise efficient remedial programs. However, the validity of this claim could be tested clinically.

REFERENCES

[1] Compton, A. <<Generative Studies of Children's Phonological Disorders>>, Journal of Speech and Hearing Disorders, 35, 315-339. (1970).

[2] __________ <<Generative Studies of Children's Phonological Disorders: A Strategy of Therapy>>, in S. Singh (ed.) Measurements in Hearing, Speech and Language. Baltimore; University Park Press. (1975).

[3] Connell, P. <<Markedness Differences in the Substitutions of Normal and Misarticulating Children>>, paper presented at the American Speech-Language-Hearing Association Convention. Toronto, Canada. (1982).

[4] Dinnsen, D.A. and Eckman, F. <<A Functional Explanation of Some Phonological Typologies>> in R. Grassman et al. (eds.) Functionalism. Chicago: Chicage Linguistics Society. (1975)

[5] Dinnsen, D.A. and Elbert, M. <<On the Relationship between Phonology and Learning>>, in M. Elberb, D.A. Dinnsen, and G. Weismer (eds.), Phonological Theory and the Misarticulating Child. American Speech-Language-Hearing Association Monograph. (1984).

[6] Dinnsen, D.A., Elbert, M. and Weismer, G. <<Some Typological Properties of Functional Misarticulation Systems>>, Phonologica. (1980).

[7] Ingram, D. <<Phonological Disability in Children>>, New York: Elsevier. (1976).

[8] Judith, A. <<A Further Characterization of Functionally Misarticulated Speech>>, in Robert Port (ed.) Research in Phonetics, 4, (1984).

[9] Lorentz, J.P. <<An Analysis of Some Deviant Phonological Rules>>, in B.M. Morehead and A.M. Morehead (eds.). Normal and Deficient Language. Baltimore: Univer. Part Press, 29-59, (1976).

[10] Mitleb, F. <<Generative Accounts of Misarticualted Speech>>, Arab Journal for the Humanities. Kuwait. To appear, (1986).

[11] __________ <<Speech Misarticulations>> A monograph to be published by Kuwait Societv for the Advancement of Arab Children, (1986).

[12] Shriberg, L.D. and Kwiatkowski, J. (1982). <<Phonological Disorders: A Diagonistic Classification System>> Journal of Speech and Hearing Disorders, 47, 226-241.

LATERAL RELEASE IN THE ARTICULATORY CLOSURE OF THE VELAR STOP SOUNDS [k, g]: AN ACOUSTIC STUDY

ZYUN'ICI SIMADA	MINAKO KOIKE	AKIKO OKAMOTO	SEIJI NIIMI
Dept. of Physiology School of Medicine Kitasato University	Dept. of Rehabilitation Kitasato University East Hospital	Dept. of ENT Kitasato University Hospital	Research Institute of Logopedics and Phoniatrics University of Tokyo

ABSTRACT

Lateral lisps involving the velar stop sounds [k, g] were examined acoustically. Lateralized and non-lateralized tokens consisting of the mora [ki], [ke] or [gi] were recorded from the same two adult females who had previously shown distorted pronunciations, but had been corrected to normal articulatory gestures with speech training. It was found that the r.m.s. amplitude of the burst of lateralized tokens was larger than that of non-lateralized tokens. In spectra, a local peak was noted consistently in the vicinity of 5.0 kHz for the lateralized tokens [ki] across both subjects.

INTRODUCTION

In speech clinics, therapists have noted that consonants such as [s], [k] and [tʃ] sometimes sound deviant before front vowels. This sound distortion has been observed typically among children with cleft palate, and is often referred to as a lateral lisp or, more specifically, as being lateralized. It is partly the turbulent outflow of air from the lateral paths of the consonantal constriction that causes the acoustic distortion detectable to speech therapists. However, the acoustic properties corresponding to lateral lisps are less clear, except in a very few reports [1]. In this study, we will focus on the acoustic differences between the normal and the lateral release of the closure of the velar stops [k, g] produced by speakers having previously shown lateral lisping in spite of no organic defects of the speech organs.

Theoretical considerations permit us to expect experimental results as follows. When a consonantal stoppage is released at the lateral side(s) of the tongue contact, the air needs to flow out through the narrow, constricted path(s). This probably results in turbulence noises with larger amplitudes compared to more open paths if we assume the air volume flowing into the constriction to be constant. Moreover, voicing may be delayed, since it takes a long time for the articulatory stoppage to achieve a complete opening. Accordingly, we will employ two measures to characterize velar stop bursts, namely, their duration and root-mean-square (r.m.s.) amplitude. Finally, a linear predictive coding (LPC) analysis will estimate spectral properties for the bursts due to distorted articulatory configurations.

METHODS

Subjects

Subjects were two adult females in their early twenties, S.A. and T.M., who had no apparent pathological defects of the speech organs. In their childhood, they had displayed the deviant type of articulation outlined in the introduction, but had been corrected to normal articulatory gestures with speech training. Subject S.A. was from Kobe, a port in the Kansai area of the Japanese mainland where a dialect different from Tokyo Japanese is spoken. Subject T.M. grew up in Sagamihara, a city on the outskirts of Tokyo.

Test Moras and Procedure

In both subjects' speech, distorted pronunciations had been salient in particular whenever velar stops occurred before the front vowel [i] or [e] and in palatalized moras like [kja]. The flap [r] of subject S.A. had also been deviant in the same phonetic environments. A total of 38 moras were chosen as test moras. These were the non-lateralized moras [ka, ko, ke, ku, ki, kja, kjo, kju, ga, go, ge, gu, gi, gja, gjo, gju, ra, ri, ru, re, ro, rja, rjo, rju] and the lateralized moras [Ke, Ki, Kja, Kjo, Kju, Ge, Gi, Gja, Gjo, Gju, Ri, Rja, Rjo, Rju] (for simplicity, lateralized moras will be represented by capital letters). The lateralized moras [Ke, Ge] and those containing the flap [r] were eliminated from the productions of subject T.M., since these were never lateralized.

The test moras were read in a sound-proof room and recorded with an audio tape recorder standing outside it and running at a speed of 38 cm/s. Both subjects were told to produce three repetitions of every mora in isolation with a short pause between them. Recordings were made of the non-lateralized series of tokens in the earlier stage of the recording sessions, and the lateralized series in the later stage. In the present study, we chose the moras [ki, ke, gi, Ki, Ke, Gi] of subject S.A. and [ki, Ki] of subject T.M. from the samples obtained in the recordings. Among the voiced tokens, only one [gi] and two samples of [Gi] could be analyzed, since the remaining tokens showed voicing simultaneously with the burst.

Acoustic Segmentation

The tokens of interest were transferred to a Hewlett-Packard 1000 computer for acoustic analysis. In digitization, the audio signals recorded were passed through a low-pass filter, with roll-off beginning at 8.5 kHz, and then sampled with a quantization level of 12 bits and at an interval of 50 µs. The waveform of every token was displayed on a graphics display terminal for the visual demarcation of the acoustic segments which were necessary in the acoustic analysis. The boundaries were determined by taking the coordinate values of the waveform in the vicinity of the point of interest with the aid of a movable cursor.

RESULTS

Figure 1 compares the waveforms of the non-lateralized token [ke] and the lateralized token [Ke] as produced by subject S.A. By inspection, we specified three landmarks on each waveform. The first was the instant of stop release--the point at which the signal exceeded the noise level of the baseline. The second was the onset of voicing in the vowel--the point at which the first glottal pulse emerged as a periodic pitch. Finally, the beginning of every pitch period that followed the first pulse was defined as the point at which the waveform curve crossed the baseline in an upward direction. The phrase "consonantal burst" in this study refers to the interval between the first and the second point.

Duration of the Burst

As is apparent in Figure 1, lateralized tokens tended to have a longer burst duration than non-lateralized tokens. The lateralized class had an average duration of 37.2 ms for [Ki], 58.9 ms for [Ke] and 13.6 ms for [Gi] in subject S.A., and 51.8 ms for [Ki] in subject T.M. The non-lateralized class had an average duration of 35.1 ms for [ki], 31.6 ms for [ke] and 30.3 ms for [gi] in S.A., and 44.0 ms for [ki] in T.M. Here, it is noteworthy that subject S.A. showed a shorter duration for the lateralized voiced tokens [Gi]. (One sample of [ki] and one sample of [Ke] produced by S.A. were eliminated from the averaging of the durations because of large deviations.)

R.M.S. Amplitude of the Burst

Figure 1 also illustrates the larger amplitudes of the lateralized token [Ke]. What time course did the r.m.s. amplitude take during the tokens? To answer this question, we made preliminary measurements of the short-term r.m.s. amplitudes. Figure 2 shows the curves obtained for the samples of Figure 1 over a period of 200 ms from the onset of speech, where a rectangular window was moved successively in steps of 5 ms during the burst, and placed pitch-synchronously over every glottal pulse during the vowel. As can be seen in Figure 2, the lateralized [Ke] had a temporal course at a higher level of the r.m.s. amplitude during the burst. Thus, again, we first measured the r.m.s. amplitude for every token of the same mora over the total length of the burst, and we then computed a ratio by dividing the average value of the lateralized tokens by the average value of the non-lateralized tokens. The ratios derived from subject S.A. were 2.42 (3.8 dB) for [Ki/ki], 1.68 (2.3 dB) for [Ke/ke] and 2.62 (4.2 dB) for [Gi/gi], and 2.66 (4.2 dB) for [Ki/ki] from subject T.M.

Above, we pointed out that the lateralized tokens had a larger r.m.s. amplitude value in the burst than the non-lateralized tokens. However, we could not directly evaluate the amplitude of the burst, since any change in the subjects' overall loudness would have had an influence on the absolute level. Accordingly, in order to evaluate the r.m.s. amplitude of the burst relative to the vowel portion, we adopted a slightly modified version of the energy ratio proposed by Jongman et al. [2]. Incidentally, in examining Figure 2 closer, we can note that the curves there show a rapid rise in amplitude and slow down some five glottal pulses from the onset of voicing, though the r.m.s. amplitude varies at a higher level for the token [Ke] than for the token [ke]. Therefore, we decided to average the amplitudes at the onset of the vowel over the period of its four complete pitches. The first glottal pulse was further added to the averaging step whenever it did not form a complete period. The size of the window for the vowel varied from 12.3 to 14.9 ms for S.A., and from 13.6 to 16.0 ms for T.M. In the last step, a ratio was computed between the r.m.s. value at the onset of the vowel and the r.m.s. value at the consonantal burst. The bigger the amplitude of the burst, the smaller the ratio value.

To sum up the energy ratios, we obtained the following average values for the lateralized tokens: 2.90 (4.6 dB) for [Ki], 4.10 (6.1 dB) for [Ke] and 1.69 (2.3 dB) for [Gi] from S.A., and 1.25 (1.0 dB) for [Ki] from T.M. For the non-lateralized tokens: 5.50 (7.4 dB) for [ki], 5.40 (7.3 dB) for [ke] and 2.90 (4.6 dB) for [gi] from subject S.A., and 4.63 (6.7 dB) for [ki] from subject T.M.

Spectral Properties

In the previous sections, we argued that the lateralized tokens were characterized by a relatively larger r.m.s. amplitude. If this property were ascribed to the peculiar configuration of the tongue, then we would expect, for example, divergences in the location of local peaks to be detectable in their spectral envelopes. A variety of research efforts so far have provided evidence that spectral peaks of short-term acoustic representation are important for the phonemic distinction of stop sounds [3, 4]. It is natural that we should want to take enough care in accurately estimating the spectral shapes of non-stationary waveforms, for instance, stop releases. Thus, in this report, we used an LPC analysis similar to the method of Stevens and Blumstein, despite its some shortcomings [3]. The mean value was first calculated for a truncated waveform and subtracted from the original signal. Then, the signal, after being multiplied by a half-Hamming window without high-frequency pre-emphasis, was submitted to a 24-term LPC computation using the autocorrelation method. The window was 25 ms long and placed at the start of the burst.

Samples of the spectra of the [ke] and [Ke] tokens, which correspond to the waveforms of Figure 1, are shown in Figure 3. In examining these spectra, we can note that both tokens have concentrations of energy in the regions below 0.5 kHz and from the mid-frequency 2.0 to 3.0 kHz. However, the component in the mid-frequency region was not very regular for the tokens [ke]; the concentration of energy was high and flat for one

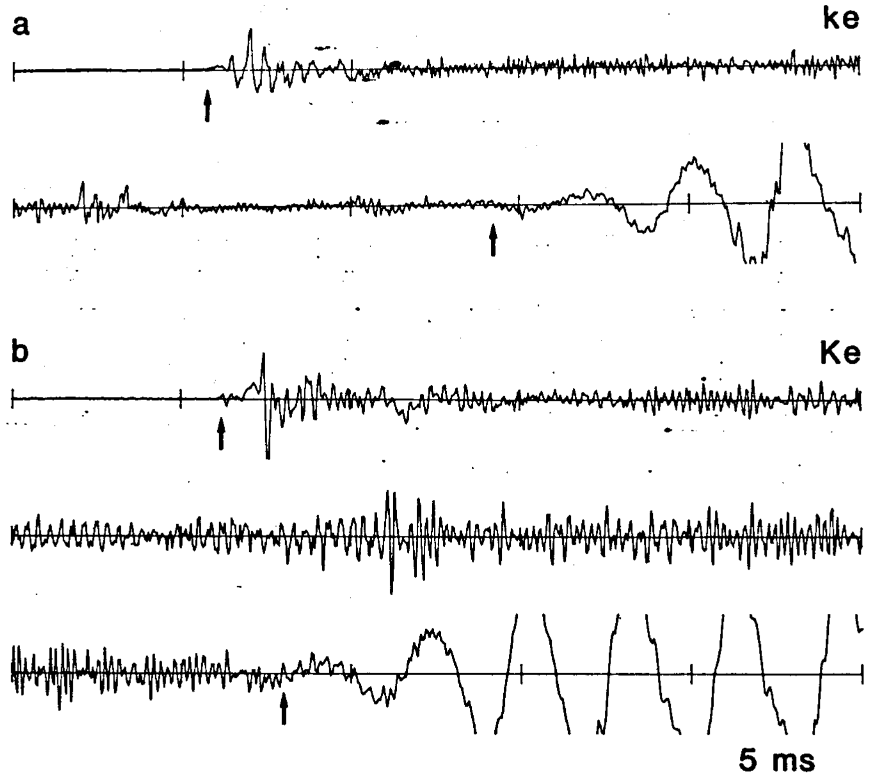


Figure 1. Waveform displays for the non-lateralized mora [ke] and lateralized mora [Ke] as produced by subject S.A. The arrows indicate the instant of the stop release and the onset of voicing in the vowel, respectively.

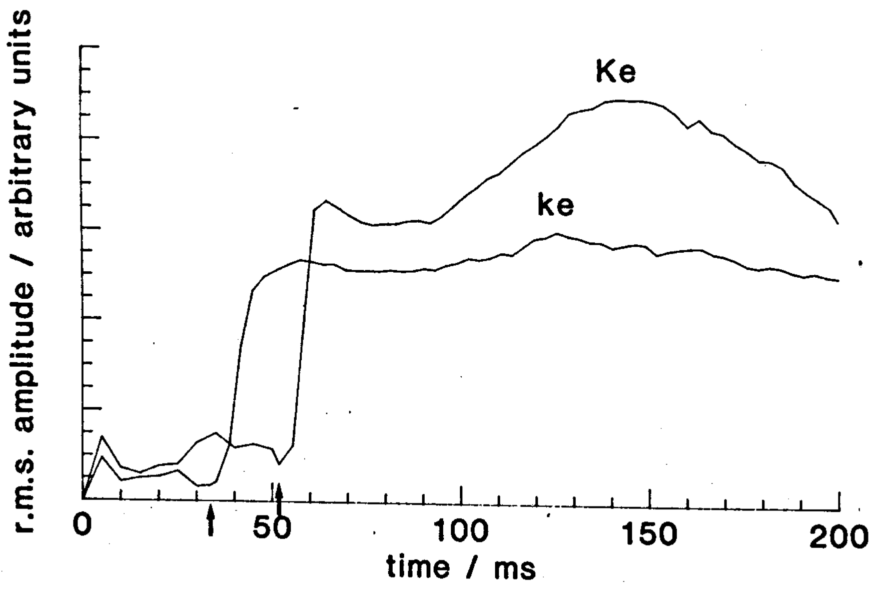


Figure 2. Time tracings for the short-term r.m.s. amplitudes corresponding to the samples in Figure 1. The arrows point to the onset of voicing in the vowels.

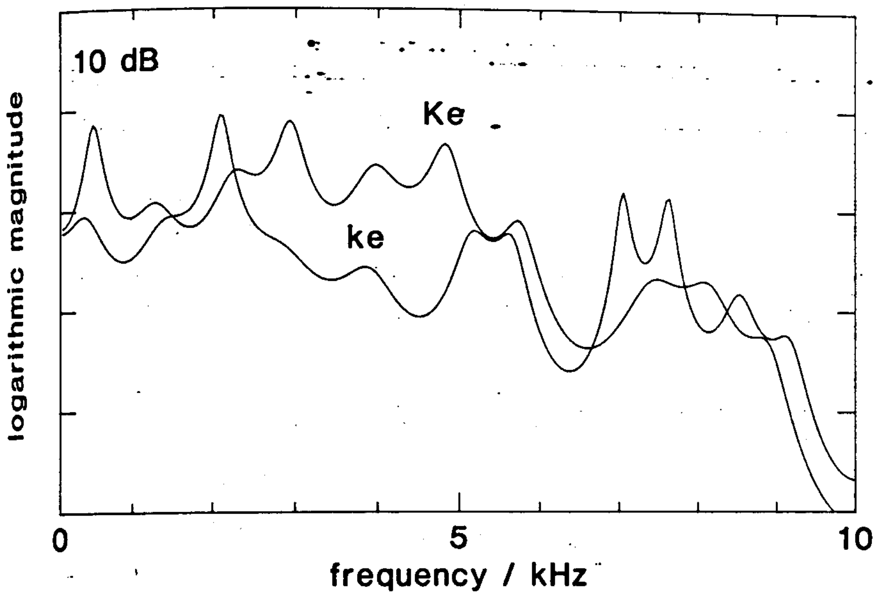


Figure 3. Spectra for the bursts obtained in the same samples as in Figure 1. A 25-ms half-Hamming window was used for the analysis.

sample, and had two separate peaks between 2.0 and 3.0 kHz in the other sample. On the other hand, for the lateralized [Ke] tokens, a single prominent peak was found consistently in the vicinity of 3.0 kHz. The non-lateralized [ke] tokens showed a significant drop in level in the region of 6.0 kHz.

In the spectra of [ki] and [Ki] from S.A., two local peaks were prominent in the vicinities of 3.0 kHz and 5.0 kHz. In particular, peaks at 5.0 kHz for the tokens [Ki] occurred rather consistently and showed as almost equal an amplitude as those at 3.0 kHz. In contrast, the corresponding peaks were irregular in shape and level for the non-lateralized tokens [ki].

In the spectra of the [gi] and [Gi] tokens, some samples showed a higher level between the first peak below 0.5 kHz and the second peak around 2.9 kHz. The spectra of [Gi] were similar in their envelopes to those of [Ki] in the range from 0 to 5.0 kHz.

For subject T.M., on the other hand, the spectra obtained showed consistent patterns in their envelopes for both classes of tokens. In the spectra of the non-lateralized [ki] tokens, a prominent peak was consistently found between 3.0 and 3.5 kHz. In the spectra of the lateralized [Ki] tokens, a peak was also seen in the same region, but it gradually shifted to a higher, sharper peak emerging in the region of 5.0 kHz. This was in contrast to the [ki] tokens, for which the spectra were rather flat in this region.

DISCUSSION

It was the aim of this study to examine how well the acoustic characteristics that seem to play a phonemically contrastive role distinguish such allophonic processes as non-lateralized versus lateralized articulatory gestures. According to Jongman et al. [2], the energy ratio of the burst is kept appreciably contrastive for alveolar and dental stops even in English, a language that has only alveolar stops, or Dutch, a language that has only dental stops. Our results also suggest that this metric of energy ratio is effective for separating non-lateralized and lateralized sounds. The ratio value of four, which indicates a difference of about 6.0 dB, seems to be adequate for this separation.

Finally, does the local peak found consistently around 5.0 kHz for the lateralized [Ki] tokens across both subjects contribute to the perceptual impression characteristic of these tokens? This question is interesting, since the role that local peaks play in the perception of phonetic variants of stops is not clear, and it needs to be resolved by perceptual experiments and other studies.

ACKNOWLEDGEMENTS

We are grateful to Y. Takeuchi for help with the audio recordings, to N. Takizawa for discussion on the issues of acoustic analysis, and to T. Katakura for technical assistance.

REFERENCES

[1] K. Takahashi, K. Michi, H. Hamada and T. Miura, "Acoustic characteristics of Japanese lateral misarticulation," Transactions of the Committee on Speech Research, The Acoustical Society of Japan, S85-92, pp. 719-725, 1986.

[2] A. Jongman, S. E. Blumstein and A. Lahiri, "Acoustic properties for dental and alveolar stop consonants: a cross-language study," J. Phonet., vol. 13, pp. 235-251, 1985.

[3] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am., vol. 64, pp. 1358-1368, 1978.

[4] D. Kewley-Port, "Time-varying features as correlates of place and articulation in stop consonants," J. Acoust. Soc. Am., vol. 73, pp. 322-335, 1983.

THE ARTICULATION DISORDERS IN MENTALLY RETARDED CHILDREN

Zbigniew Tarkowski
Institute of Pedagogy
M. Skłodowska University
Lublin

Andrzej Lewandowski
Departament of The Theory
of Teaching and Education
WSP Zielona Góra

The paper discusses the frequency of occurence, etiology and character of the articulation disorders in mentally retarded children and presents the specific character of the job of speech pathologist in special school.

1. The frequency of articulation diserders.

The mental retardation is most frequently accompanied by the speech disorders /1, 2, 3, 11/. Among these disorders the articulation disorders constitute the majority /above 81%/.

The frequency depends on the degree of mental retardation, its etiology, the child's age and the educational background.

It has been assumed /3, 4, 12/ that the deeper the mental retardation, the more retarded and disordered the development of articulation. It has been also estamated /8/ that the articulation disorders appear most frequently in the case of Down's syndrome /95%/, less often in the organic syndromes /84%/ and in the family mental retardation. It has been observed /5,10, 12/ that the number of the disorders decreases with the age of the children.

The articulation disorders seem to apper more frequently in children living in special care institution in comparison with the children living in their families.

2. The etiology of the articulation disorders.

In the case of the mentally retarded children one has to deal with the articulation disorders of the organic origins caused by the occlusion and dental defects, cleft palate, lips and tongue, dysarthia, aphasia, dyspraxia, impaired hearing, as well as with articulation disorders of functional origins caused by the lowered phonemic audition, accepting the incorrect pronunciation, emotional disorders. The articulation disorders of origins are assumed to occur more often in mentally retarded children than in children normally developed.

Dealing with the pathogenesis of mental retardation and the articulation disorders, it is possible to distinguish the following relationships between them:

a/ common etiology, e.g. infantile cerebral palsy

b/ different etiology, e.g. mental retardation and dysglossia

c/ some of the features of mental retardation /e.g. lower concentration of attention and low learnability/ constitute the cause of the delayed development of articulation.

3. The character of articulation disorders

A unified view on the character of development and disorders of articulation does not exist. On one hand, it has been noticed /4/ that the general pattern of articulation in underdeveloped reminds the articulation pattern in mentally retarded. It has been also assumed that this pattern is analogical in both pupils living in the special education institutions and pupils living with their families.

On this basis the development of articulation in children retarded and in children normally developed is considered to proceed according to the identical phases and rules, with the only difference that in the case of mentally retarded it proceeds slower and with greater number of errors. As far as the kind of errors is concerned, the children do not differ /1, 6/.

On the other hand, it is claimed that the distincive features of the mentally retarded are constituted by the phonetic deficiencies indicating the possibilities of appearance of the structural defects.

In our opinion, the characteristic phenomenon is very frequent occurence of complex speech disorders covering mostly dyslalia, voice disorders and stuttering /8/. The disorders, of course, posses different degree of intensity /1, 3, 5/.

4. Logopedical procedure.

The specific character of the logopedic therapy in special school results from the factors:

a/ the character of mental retardation

b/ complex speech disorders

c/ the parental attitude

d/ the organization of teaching

To the factors which hamper the therapy be long the following features of retardation poor verbal memory, low concentration, disordered visual and auditory perception, paliphrasia, limited transfer, low degree of self-criticism, slow pace of learning and low motivation in correcting the articulation.

The above characteristics accompany the complex speech disorders. The process of eliminating these disorders assumes the active role of parents, which is very often unsatisfactory.

Because of the lack of supervision over the articulation of the mentally retarded children its condition vecomes worse after the holidays.

Everything what has been said above makes the elimination of the speech disorders in special school difficult and long process aimed, what is worth emphasizing, at the realization of realistic goals. The goals are determined in an educational institution by the Individual Revalidation Teams, which members are, among others, the speech pathologists as well.

They are responsible for designing and putting into practice the programmes of therapy including:

- a/ the preliminary diagnosis
- b/ the individual programme of eliminating the speech disorders /including the articulation disorders/
- c/ diagnosis in the course of the therapy and the final diagnosis

The programme provides space for the co-operation between the speech pathologist, the psychologist,the teacher and the corrective physical exercises instructor.

So, for example, the therapy of a mentally retarded child with dyslalia, stuttering and low motor activity efficiencyconsists in the speech pathologist talking care of improving the articulation and the fluency of speaking, the psychologist reducing the muscular tone and the emotional tension by means of relayation, the corrective physical exercises instructor stimulating the development of the motoric skills and fiàlly the teacher fulfilling the prescriptions of the above to the best of his abilities.

The effects of the complex therapy depend on the accuracy of the diagnosis and the accepted programme of the therapy, the degree of complexity of the speech disorders the pace and motivation of learning, the depth of mental retardation,the co-operation of the speech therapist with the remeining members of the team, the involvement of parents.

REFERENCES: 1. Baun M.: Beeintrachtigungen der Sprache bei Geistigbehinderten, "Sonderpädagogik" 1978, 1, 15-25. 2. Blount W: Language and the more severely retarded: a review, "American Journal of Mental Deficiency" 1968, 73, 21-29. 3. Clarke A. M., Clarke A. D. B. /eds/: Mental Deficiency, London 1965. 4. Jordan T. E.: The Mentally Retarded, Ohio 1979. 5. Petrowa W. G.: Sprachentwicklung Schwachsinniger Kinder, "Die Sonderschule" 1979, 2 Beiheft. 6. Schiefelbusch R. L.: Language functions of retarded children, "Folia Phoniatrica", 1969, 21, 129-144. 7. Schiefelbusch R. L.: Speech, language and communication disorders of the multiply handicapped, "Folia Phoniatrica" 1984, 36 8-24, 8. Schlanger B. B., Gottsleben R.H. Analisis of speech defects among the institutionalized mentally retarded, "Journal of Speech and Hearing Disorders" 1957 22, 98-103, 9. Seeman M.: Sprachstörungen bein Kindern, Berlin 1974, 10. Sowczenko M. A., Noworowa R. A.: Niekotoryje osobàennosti naruszenij zwukoproiznoszenija u umstwienno otstałych szkolnikow, "Defektologija" 1976, 5, 16-20, 11. Spradlin J. E.: Language and communication of mental defectives, /In:/ Ellis E. R. /ed./:

Handbook of Mental Defectives, New York 1963. 12. Szuniewicz A.: Próba badań wad wymowy dzieci upośledzonych umysłowo w warszawskich szkołach specjalnych, "Logopedia" 1967, 7, 112-119.

ЗАВИСИМОСТЬ МЕТОДИКИ ОБУЧЕНИЯ ПИСЬМУ И ЧТЕНИЮ ОТ ОСОБЕННОСТЕЙ ЯЗЫКА

КАРЛ КАРЛЕП

Тартуский государственный университет
Тарту, Эстонская ССР, СССР, 202400

РЕЗЮМЕ

Часть операций действий письма и чтения, а также оперативные единицы при этом зависят от фонетико-фонематической системы и от правил графики конкретного языка. В эстонском языке специфическими операциями является определение группы звуков (гласные, смычные согласные, несмычные согласные) и их степени долготы. Звуковой анализ целесообразно проводить в составе слова, фонематический - в слове или в речевом такте. Наименьшей единицей синтеза является речевой такт. Для наглядного предъявления правил графики применяются составленные из фишек схемы слов, отражающие фонематические признаки звуков. В целях облегчения синтеза употребляются графические средства: выделение букв цветом или шрифтом, обозначение степени долготы слова знаком.

ИЗЛОЖЕНИЕ ТЕКСТА

В современной психологической и методической литературе письмо и чтение рассматриваются как сложные умственные действия, состоящие из нескольких групп операций (А.Р.Лурия, Д.Б.Эльконин и др.). Последние в том или ином языке определяются фонетикой, графикой и правилами графики языка. Формирование этих операций предполагает усвоение определенных учебных приемов, которые потом в автоматизированном виде войдут в состав основного действия. При этом объектом анализа и синтеза служит прежде всего звуковой состав слова, а основной трудностью - несовпадание оперативных единиц письменной и устной речи, т.е. несовпадание единиц произношения и графического изображения языка. Сказанное позволяет заключить, что теоретическое обоснование методики обучения чтению и письму опирается на языкознание, психолингвистику и педагогическую психологию.

Целью лингвистического обоснования методики обучения является определение тех языковых единиц, их признаков и отношений, анализом которых учащийся должен овладеть. Так как написанное слово представляет собой модель произносимого слова, в котором звуки (фонемы) символизированы буквами (графемами), то ученик должен прежде всего овладеть умением определить последовательность фонем в слове, а затем их алфавитным обозначением. В этих целях учащемуся необходимо ознакомиться со всеми сегментальными фонемами изучаемого языка, научиться различать их по слуху и в произношении как изолированно, так и в составе слова. Определение последовательности сегментальных фонем называется звуковым анализом. Необходимо учесть, что языки различаются по трудности анализа, так как системы фонем по дифференциальным признакам не совпадают. Методика должна учитывать именно то, какие признаки и их комбинации являются дифференциальными для данного языка. Например, в русском языке необходимо различать глухость-звонкость и мягкость-твердость звуков. В эстонском языке указанные дифференциации не требуются.

Если учащиеся кроме последовательности звуков определяют и важные для данного языка признаки и отношения фонем, то они проводят уже фонематический анализ. Именно последний лежит в основе применения правил графики при письме. Данное положение важно и для чтения, так как правила чтения являются обратными в отношении к правилам графики. Например, противопоставления по твердости-мягкости в русском языке определяют применение йотированных и нейотированных гласных букв после согласного (па-пя, ба-бя и т.д.).

Признаки, определяемые в результате фонематического анализа, могут быть свойственны сегментальным фонемам или быть супересегментального порядка. Во многих языках требуется, например, определение долготы звуков. Названный признак является относительной величиной и обнаруживается только в сравнении с долготой других звуков в какой-то единице более высокого уровня, чем фонемный, т.е. в слогах, речевых тактах или словах. В связи с этим необходимо подбирать и соответствующие приемы анализа.

В эстонском языке правила графики (правила применения букв алфавита при написании слов) опираются на два фонематических признака. Ими являются принадлежность фонемы к одному из трех групп звуков (гласные, смычные и несмычные согласные) и степень долготы звуков. Именно от комбинации названных признаков соседних звуков в слове зависит употребление одинарной или двойной буквы, а также выбор букв для обозначения смычных согласных.

Следовательно, сознательное усвоение правил графики и чтения возможно только с опорой на звуковой и фонематический анализ. Последние в свою очередь зависят от фонетико-фонематической и графической системы языка. Если определение последовательности сегментальных фонем является универсальной операцией для всех языков с фонетическим письмом (другое дело,если языки отличаются по системе сегментальных фонем), то операции фонематического анализа зависят от конкретного языка, т.е.от основных фонематических признаков и правил применения букв в данном языке. Последнее положение указывает и на то, что методику обучения чтению и письму нельзя прямо перенести с одного языка на другой.

Целью психолингвистического обоснования методики является определение оперативных единиц анализа и синтеза при чтении и пиcьме. Значимость проблемы заключается в том, что языковые единицы, принадлежащие символизации и моделированию, не совпадают с единицами произношения, и ребенку приходится постоянно переходить с одной системы на другую.

Центральной является проблема слога как минимальной единицы произношения. Для письма необходимо выделять звуки и определять фонемы, но в произношении слог является неделимой единицей, в составе которого неантагонистические артикуляционные движения соседних звуков реализуются частично одновременно. Этим объясняется и факт, что необученный ребенок при анализе слова выделяет слоги, а не звуки. В целях преодоления такой сегментации приходится усваивать особое произношение - позвуковое проговаривание слова. В результате каждый звук как бы образует отдельный слог. В таком случае ослабляется влияние коартикуляции, звук в слове приближается по своим акустическим и артикуляционным признакам к изолированному звуку. Если ребенок до этого в состоянии различать изолированные звуки, то он узнает их в результате проговаривания и в слове. Позвуковое проговаривание способствует также определению последовательности фонем.

Следовательно, слоговой анализ слова в целях обучения письму не требуется.Более того, если слогораздел происходит внутри звука (так разделяются, например, долгие и сверхдолгие согласные эстонского языка то такой анализ может вызывать даже недоразумение. Если анализ слова в эстонском языке провести сперва по слогам, а затем по звукам, может появляться"лишний" звук. Можно считать, что позвуковое проговаривание является одним из основных приемов перестройки фонематического слуха независимо от языка. Боязнь того, что слоговое произношение некоторых согласных является искусственным, не обоснована, так как на начальном этапе обучения чтению такое произношение (чтение по складам)свойственно всем детям.

Однако позвуковое проговаривание позволяет выяснить не все фонематические признаки фонемного уровня. Например. долгота звука как относительная величина может быть определена в какой-то более объемной единиице, чем отдельный звук. Исследования М.Хинта /1/ позволяют заключить, что наименьшей удобной единицей для определения степени долготы звуков в эстонском языке является речевой такт. Последний представляет собой единицу слогового уровня, начинается с ударного слога, продолжается до паузы или до следующего ударного (соударного) слога и состоит из 1-3 слогов. Характерной чертой речевого такта является то, что по своим фонетическим признакам он всегда может быть отдельным словом. Следовательно, в речевом такте сохраняются все долготные отношения звуков. В отдельном слоге эти отношения часто искажаются, так как по крайней мере один из звуков в таком случае должен быть сверхдолгим. Если в слове сверхдолгого звука нет, а имеются только короткие или долгие звуки, то речевой такт не может быть короче двух слогов.

Однако выделение, произношение или восприятие речевого такта ещё не обеспечивает определения долготных отношений звуков, так как эти отношения познаются только в сравнении. Обеспечивается возможность сознательного анализа в результате особого произношения - изменения долготы критических звуков.В этих целях учащихся обучают произношению критических звуков в речевом такте во всех возможных степенях долготы. Таких вариантов произношения в зависимости от речевого такта может быть от двух до шести. Есть основание предполагать, что применение более крупных.чем звук единиц анализа целесообразно не только для определения долготы звуков. К таким признакам относится, вероятно, и сингармонизм в тюрских языках, звонкость-глухость последнего согласно в закрытом слоге в русском языке и т.д.

При обучении чтению почти общепринято положение о том, что первичной единицей синтеза должен быть слог. Объясняется это тем,что слог - наименьшая единица произношения, в составе которого реализуются основные закономерности коартикуляции.Однако в языках, где суперсегментальные фонематические признаки имеют особое значение, слог может оказаться при синтезе неудобной единицей. Такое явление обнаруживается и в эстонском языке, так как при чтении по слогам искажаются долготные отношения звуков. Следовательно, единицей синтеза может быть такой речевой сегмент, где долготные отношения звуков возможно восстановить правильно.Этой единицей является речевой такт или слово. Так как речевой такт одновременно может совпадать и со словом и слогом (односложные слова), то на начальном этапе обучения чтению целесообразно применять именно односложные слова. В дальнейшем многосложные слова предъявляются для чтения по речевым тактам, а не по слогам.

Указанное положение подтвердилось и наблюдениями в школе. При требовании читать по слогам дети с эстонским языком обучения синтезируют сперва про себя слово целиком, а затем читают "для учителя" по слогам. Выяснилось также, что в 1-П классах вспомогательной школы дети читают предъявленный по слогам материал медленнее, чем текст без деления на слоги, и допускают больше ошибок.

Итак, психолингвистический анализ подтверждает также вывод о том, что обучение письму и чтению зависит во многом от системы языка. В частности в эстонском языке особое значение приобретает речевой такт как единица анализа долготы звуков и как единица синтеза при чтении.

Лингвистическое и психолингвистическое обоснование методики не самоцель, а должно служить построению системы упражнений при обучении и составлению учебной литературы. В этих целях прежде всего необходимо определить операциональный состав действий чтения и письма, а затем разработатьь методические приемы как для формирования каждой операции в отдельности, так и этих действий в целом. Конкретизируем указанные положения на примере эстонского языка.

В эстонском языке действие письма состоит из следующих операций:

- Звуковой анализ: отделение звуков в слове друг от друга, определение сегментальных фонем и их последовательности. Обеспечивается звуковой анализ такими приемами работы, как позвуковое проговаривание слов, сравнение слов по звуковому составу, классификация слов по наличию или позиции звуков, придумывание слов с данным звуком или звуками и т.д.
- Фонематический анализ: определение группы и степени долготы звуков. Применяются такие приемы работы, как нахождение критических звуков в слове и изменение их долготы, сравнение речевых тактов или слов по долготе звуков, определение степени долготы в результате сравнения, классификация слов по степени долготы критических звуков и т.д.
- Кодирование: выбор букв и их написание. Осуществляется выбор букв по правилам графики. Для наглядного предъявления этих правил целесообразно применить разнообразные схемы слов, составленные из фишек. При этом фишки обозначают группу звуков и их долготу, позволяют подбирать одинарную или двойную буквы.
- Самоконтроль: сличение написанного слова с оригиналом, повторный анализ и т.д.

В настоящее время считается, что предварительным условием обучения чтению является владение звуковым и фонематическим анализом. При недостаточном владении указанными операциями этап первоначального чтения протекает медленно, так как операции анализа приходится усвоить по ходу обучения.

Собственно операциями чтения являются следующие:

- Определение звуко-буквенного соотношения: различение фигур букв, узнавание графем (в том числе двойных букв) и соотношение их с сегментальными фонемами.
- Определение ориентиров для синтеза: узнавание гласных для восстановления коартикуляции в слогах, определение звуко-буквносителей долготы для восстановления ударно-ритмической структуры речевого такта (слова).
- Восстановление звуковой структуры слова: построение гипотез, произношение слова голосом или про себя.
- Самоконтроль по значению или по звуко-буквенному составу слова,

Необходимо отметить, что ориентиры для синтеза на начальном этапе обучения целесообразно предъявлять графическим способом: выделять соответствующие буквы цветом или шрифтом, обозначать ударно-ритмическую структуру слова знаком степени долготы.

Все указанные теоретические положения относятся к обучению всех категорий детей. Однако, их строгое соблюдение особенно важно при обучении аномальных учащихся: умственно отсталых детей, детей с нарушениями письма и чтения или с задержкой психического развития. Если нормальный ребенок обычно способен подобрать необходимые приемы анализа и синтеза языкового материала самостоятельно, то аномальные дети на такие открытия, как правило, не способны. Им свойственно также то, что сложные действия долго не автоматизируются, а некоторые операции из состава действия часто пропускаются. Поэтому требуется поэтапное формирование и автоматизация как отдельных приемов работы, так и действия в целом. Особое значение приобретает схематическое моделирование действий, в том числе и последовательности операций. В то же время значение отдельных учебных приемов не однозначно. Самыми важными необходимо считать позвуковое проговаривание в целях формирования звукового анализа и изменение степени долготы звуков для фонематического анализа.

/1/ M.Hint. Eesti keele sõnafonoloogia I. Tallinn, 1973: 253 lk.

INTONATION ET THEMATISATION EN RUSSE MODERNE

CHRISTINE BONNOT IRINA FOUGERON

U.E.R. de Slavistique
Université de Paris-Sorbonne
Centre Universitaire du Grand Palais, Cours la Reine, 75008, Paris, France.

RESUME

Pour éviter les ambiguïtés naissant de la définition traditionnelle du thème comme "ce qui est connu", on propose d'en donner une définition purement formelle, basée sur des critères prosodiques. Ayant mis en évidence trois intonations de mise en relief thématique qui permettent de préciser le caractère de la relation thème-rhème, on examine plus particulièrement l'une d'entre elles dont on montre qu'elle est destinée à signaler que la relation posée entre un thème T et un rhème R est fondée sur l'exclusion d'autres relations parallèles de type T - R' ou T' - R. L'étude est ensuite étendue au domaine des relations entre propositions dans le cadre de l'énoncé complexe.

0. Les chercheurs qui étudient la segmentation de l'énoncé en thème et rhème définissent généralement ces deux constituants sur la base de critères purement sémantiques : le thème serait ce qui est connu ou se déduit facilement du contexte antérieur, le rhème ce qui est nouveau, et l'énoncé irait de ce qui est connu vers ce qui est nouveau, c'est-à-dire du thème vers le rhème. Ces principes connaissent toutefois de nombreuses exceptions : comme le font remarquer la plupart des auteurs, il arrive (notamment en début de texte) que le thème soit constitué d'éléments entièrement nouveaux, ou qu'inversement tous les éléments du rhème soient déjà connus. D'autres fois, en particulier dans la langue parlée, les éléments qui apportent une information nouvelle précèdent ceux qui sont déjà connus : on considère alors généralement que le rhème précède le thème et que cette inversion de l'ordre "normal" remplit une fonction "expressive", sans que le terme d'expressivité soit nulle part clairement défini.

1.1. Une telle analyse nous paraît assez peu opératoire, ne serait-ce que parce qu'elle ne fournit pas de critères simples et univoques permettant de segmenter l'énoncé : comment reconnaître le thème si ce constituant n'a pas de place fixe dans l'énoncé et s'il peut comporter aussi bien des éléments nouveaux que des éléments connus ? Nous proposons donc d'abandonner les critères trop vagues de "connu" et de "nouveau" au profit de critères formels plus facilement contrôlables, tels que ceux que fournit l'analyse prosodique. Dans les énoncés assertifs, nous appellerons thème un segment, toujours en position initiale, qui est marqué par une montée du ton et peut éventuellement être séparé du reste de l'énoncé par une pause. Lorsqu'aucun élément dans l'énoncé ne présente ces caractéristiques, celui-ci est entièrement rhématique (c'est le cas en particulier des énoncés traditionnellement analysés comme présentant l'ordre "inversé" rhème-thème).

1.2. Dans certains énoncés complexes, ou lorsqu'il est introduit par une particule de thématisation (*a*, *že*, *-to*), le thème peut, tout en conservant ces trois caractéristiques fondamentales (position initiale, montée mélodique, possibilité d'une pause), présenter certaines particularités prosodiques qui assurent sa mise en relief. Nous distinguons trois types de mise en relief :

- la syllabe tonique du thème est affectée d'une montée en flèche du Fo, qui atteint ainsi la zone des fréquences élevées, puis redescend brusquement dès la première syllabe post-tonique. Cette rupture mélodique donne l'impression d'entendre une pause même lorsqu'il n'y en a pas. (Nous notons cette intonation : $\overset{\nearrow}{T}$ / $\overset{\blacktriangledown}{R}$.)
- la syllabe tonique du thème est encore affectée d'une forte montée mélodique, mais le Fo cette fois-ci se stabilise ensuite à un niveau élevé sur toute la partie post-tonique, pour ne redescendre brusquement que vers la première syllabe du rhème.

(Notation : $\overset{\curvearrowright}{T}$ / $\overset{\blacktriangledown}{R}$.)

- le Fo chute sur la syllabe tonique du thème, puis remonte sur la partie post-tonique jusqu'à la limite supérieure des fréquences moyennes. Autrement dit, la montée thématique se trouve déportée de la partie tonique vers la partie post-tonique.

(Notation : $\overset{\checkmark}{T}$ / $\overset{\blacktriangledown}{R}$.)

1.3. Ces trois intonations sont déjà connues des linguistes soviétiques. Il s'agit en effet des trois intonations de "non-finalité" que distingue E. A. Bryzgunova dans sa présentation des schémas intonatifs de base du russe (respectivement *IK-3*, *IK-6* et *IK-4*) /1/. Selon cet auteur, tout segment non final d'un énoncé, qu'il s'agisse d'un simple syntagme ou d'une proposition entière, pourrait être affecté d'une de ces intonations. Le choix de l'une ou l'autre d'entre elles dépendrait de critères avant tout stylistiques, la première étant caractéristique de la langue courante, la seconde d'un style plus "lyrique", émotionnellement chargé, et la troisième d'un style livresque et didactique.

Or l'étude d'énoncés insérés dans leur contexte montre que ces intonations ne peuvent pas affecter n'importe quel segment non final, et que leur apparition obéit à des contraintes très strictes. D'autre part, le choix de l'une ou l'autre d'entre elles semble dépendre de critères plus syntaxiques que stylistiques. Nous pensons donc qu'il ne s'agit pas de simples marques de non-finalité, mais d'intonations de thématisation destinées à préciser la nature de la relation que le locuteur établit entre le thème et le rhème.

1.4. Nous envisagerons plus particulièrement la première de ces intonations (*IK-3* dans la terminologie de E. A. Bry:gunova). Elle signifie que la relation établie entre un thème T et un rhème R exclut l'existence d'autres relations parallèles de type T - R' ou T' - R. Ainsi elle apparaît régulièrement :

- dans les énoncés adversatifs opposant deux thèmes par l'intermédiaire des rhèmes qui leur sont attribués :

1 — *Nu, ty končila ?*

— *Komnata / ubrana, a kuxnja / eščë grjaznaja.*

"— Bon, tu as fini ?
— La chambre est faite, mais la cuisine pas encore.

La mise en relief du thème *komnata* ("la chambre") (cf. figure 1) signifie que le rhème *ubrana*("est faite") ne peut être attribué à l'autre élément de l'ensemble de départ : *kuxnja* ("la cuisine"). Dans la première proposition, T_1- R exclut donc T_2 - R.

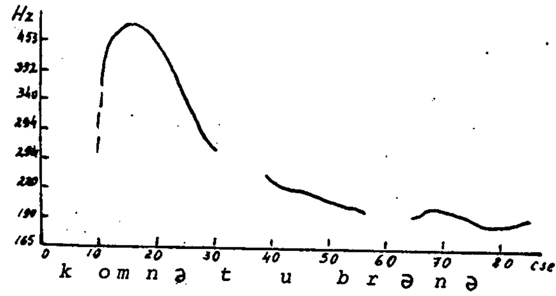


figure 1

Quant à la mise en relief du thème *kuxnja*, elle est rendue obligatoire par la présence de la particule *escë* ("encore"). Elle signifie en effet que le rhème attribué à *kuxnja* (*eščë grjaznaja* : littéralement "(est) encore sale") est autre que celui qui était visé : ici T_2 - R' exclut T_2 - R. Si l'on supprimait *escë*, l'idée que la cuisine doit être faite disparaîtrait, le rhème *grjaznaja* serait affirmé en lui-même et non pas opposé au rhème visé, et le thème *kuxnja* ne serait pas mis en relief.

On voit donc que, malgré les apparences, les deux parties de l'énoncé adversatif ne sont pas symétriques, puisque dans la première on exclut la possibilité d'avoir un autre thème, alors que dans la seconde on exclut celle d'avoir un autre rhème.

- dans les énumérations où différents thèmes, éléments d'un même ensemble, se voient attribuer chacun un rhème particulier /2/:

2 *Ja včera prišla s raboty — užin / gotov, stol / nakryt, igruški / ubrany — rebjata vsë sdelali.*

"Quand je suis rentrée de mon travail hier, le dîner était prêt, la table mise, les jouets rangés : les enfants avaient tout fait."

L'intonation T / R n'est possible ici que si la mère avait dit en partant qu'il fallait préparer le dîner, mettre la table et ranger les jouets (dans le cas contraire, on aurait des séquences entièrement rhématiques accentuées sur le premier terme.) Autrement dit, la mise en relief des différents thèmes T signifie que l'on vérifie pour chacun d'eux qu'il convient bien de lui attribuer le rhème R (le travail demandé a été fait) et non pas R' (le travail n'a pas été fait).

- dans les énoncés réfutant un présupposé :

3 — *Kak tixo ! Spit malyš ? Ukačala ego babuška ?*

— *Spit / babuška. Malyš molčit, no ne spit, igraet s pogremuškoj.*
— *Tak on babušku razbudit.*

"— Quel silence ! Le petit dort ? Grand-mère a réussi à l'endormir ?
— C'est grand-mère qui dort. Le petit est tranquille, mais il ne dort pas, il joue avec son hochet.
— Mais alors il va réveiller grand-mère !"

Au thème *spit* ("dort") est attribué un rhème (*babuska* : "grand-mère") différent de celui qu'on aurait pu attendre : T - R exclut T - R'.

2.1. Nos hypothèses peuvent être étendues aux énoncés complexes. En effet, chacune des trois intonations de mise en relief que nous avons distinguées au début de cette étude peut caractériser la première proposition d'un énoncé complexe. Nous pensons qu'on peut alors considérer que cette première proposition joue le rôle de thème, tandis que la seconde joue celui de rhème, et que, là encore, la mise en relief du thème permet de préciser la nature de la relation qui l'unit au rhème.

2.2. Ainsi, si l'on prend l'exemple des énoncés coordonnés de la forme "p et q", on voit que la proposition p est prononcée avec la première intonation de thématisation analysée ci-dessus lorsqu'il convient de souligner que p est associé à q et non à q', ou que q est associé à p et non à p' /3/:

4 — *A gde že Andrej ? Ja dumala, on u tebja poživët !*

— *Da net ! Èkzameny končilis', i Andrej uexal.*

"— Mais où est donc André ? Je pensais qu'il allait rester quelques jours chez toi !
— Penses-tu ! Les examens terminés, il est parti." (littéralement : "les examens sont terminés et il est parti.")

Contrairement à ce que pensait l'interlocuteur, André n'est plus là, car la fin des examens ne pouvait avoir d'autre conséquence que son départ : p ("les examens sont terminés") exclut q' ("André est encore là"), ce qui entraîne sa mise en relief (cf. figure 2).

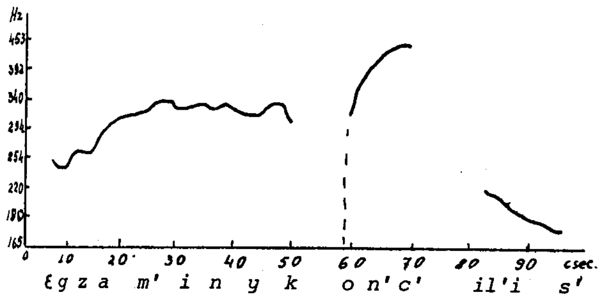


figure 2

5 *K večeru u Kirki podnjalas' temperatura. Do soroka. Ja legla s nim i spala kak u raskalënnoj pečki. A utrom smotrju : on krasnyj kak rak, a temperatura vrode men'še. Ponimaeš', pojavilas' syp', i temperatura spala.*

"Le soir, la température de Kirka est montée jusqu'à 40. Je me suis couchée avec lui et j'ai eu l'impression de dormir à côté d'une fournaise. Le lendemain matin, il était rouge comme une écrevisse, alors que la température avait l'air d'avoir baissé. Tu comprends, l'éruption apparue, la fièvre était retombée." (littéralement : "était apparue l'éruption et la température était tombée.")

Contrairement à ce que l'on aurait pu penser, l'éruption s'était accompagnée non pas d'une augmentation, mais d'une diminution de la température : on a p et q là où on attendait p et q'.

6 *My v ètom godu turpoxod na Bajkal zatejali. Uže vsë gotovo dlja putešestvija. Ostalis' tol'ko Irkiny èkzameny. Ira sdast, i možno exat'.*

"Cette année nous allons camper sur les bords du lac Baïkal. Tout est prêt pour le voyage. Il ne reste plus que les examens d'Irka. Dès qu'elle les a passés, on peut partir." (littéralement : "Ira (les) passera et on peut partir.")

L'événement q ("nous partons en vacances") ne peut avoir lieu tant que p ("Ira passe ses examens") n'est pas réalisé. Autrement dit, q est incompatible avec p' ("Ira n'a pas fini ses examens").

3.1. Les données que nous avons rassemblées indiquent qu'il est possible de donner une interprétation similaire des deux autres intonations de mise en relief thématique. Ainsi, dans une communication récente /4/, nous montrons que l'intonation réalisée par un mouvement descendant-montant (T / R, *IK-4* dans la terminologie de E. A. Bryzgunova) apparaît lorsque le caractère exclusif de la relation thème-rhème se complique d'un élément de pluralité:

- pluralité au niveau du rhème : Un thème T se voit attribuer un rhème R décomposé en plusieurs éléments r_1, r_2, r_3... En même temps, T - R exclut T - R'.

7 (Un jeune auteur dramatique lit sa pièce à un directeur de théâtre, Ivan Vassiliévitch.)

Ivan Vasil'evič sidel soveršenno nepodvižno i smotrel na menja v lornet, ne otryvajas'. Smutilo menja črezvyčajno to obstojatel'stvo, čto on ni razu ne ulybnulsja, xotja uže v pervoj kartine byli smešnye mesta. Aktëry očen' smejalis', slyša ix na čtenii, a odin rassmejalsja do slëz.

Ivan že Vasil'evič / ne tol'ko ne smejalsja, no daže perestal krjakat'. I vsjakij raz, kak ja podnimal na nego vzor, videl odno i to že : ustavivšijsja na menja zolotoj lornet i v nëm nemigajuščie glaza.

(M. Bulgakov, *Teatral'nyj roman*)

"Ivan Vassiliévitch était absolument immobile et me regardait fixement de son lorgnon. Je fus extrêmement troublé par le fait qu'il ne sourit pas une seule fois, bien que dès le premier tableau il y eût des endroits drôles. Les acteurs avaient beaucoup ri en les entendant à la lecture, et l'un d'eux en avait même pleuré de rire.

Ivan Vassiliévitch, lui, non seulement ne riait pas, mais il avait même cessé de glousser. Et chaque fois que je levais les yeux sur lui, je voyais la même chose : un lorgnon en or braqué sur moi et dedans des yeux qui ne cillaient pas."

Au thème *Ivan že Vasil'evič* est attribué un rhème en deux parties : r_1 (*ne tol'ko ne smejalsja* : "non seulement ne riait pas") et r_2 (*no daže perestal krjakat'* : "mais avait même cessé de glousser"), qui permettent d'opposer la réaction du directeur du théâtre à celle qu'avait eue les acteurs. Si le rhème était composé d'un seul élément, l'intonation de mise en relief thématique serait modifiée : le mouvement descendant-montant (cf. figure 3) serait remplacé par une montée en flèche sur la syllabe tonique, la relation T - R excluant T - R'. (Cette modification entraînerait parallèlement le remplacement de la particule de thématisation *ze* "quant à" par *a*.) Cf. 7a :

7a *A Ivan Vasil'evič / daže ne ulybnulsja. I vsjakij raz, kak ja podnimal na nego vzor, ja videl odno i to že...*

"Ivan Vassiliévitch, lui, n'eut même pas un sourire. Et chaque fois que je levais les yeux sur lui, je voyais la même chose..."

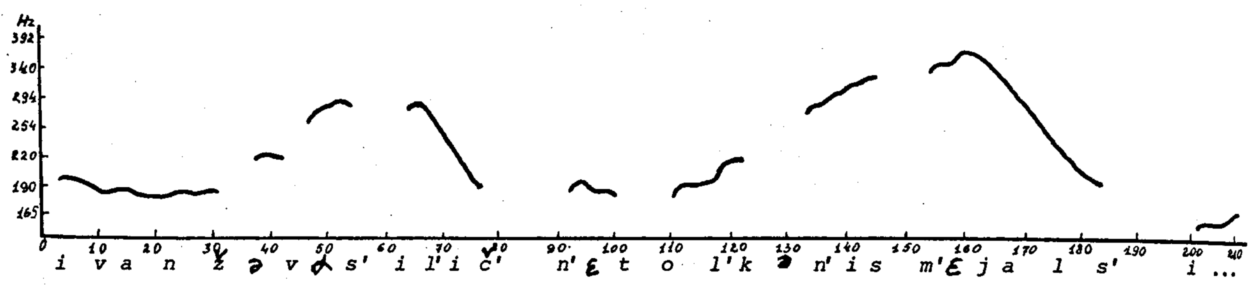


figure 3

- pluralité au niveau du thème : un rhème R est attribué à un thème T décomposé en plusieurs éléments t_1, t_2, t_3... Là encore, T - R exclut T - R'.

8 (A propos de la location d'une datcha)

— *A čto ty razdumyvaeš' ? Učastok / xorošij, dom / brevenčatyj, les / čerez dorogu — nu počemu ne snjat' ?!*

"— Qu'est-ce que tu as encore à hésiter ? Le terrain est bien, la maison solide (littéralement : "en rondins"), la forêt, il n'y a qu'à traverser la route... Pourquoi est-ce qu'on ne louerait pas ?"

La division en thème et rhème se fait ici à l'intérieur de chaque proposition et à l'intérieur de l'énoncé complexe :

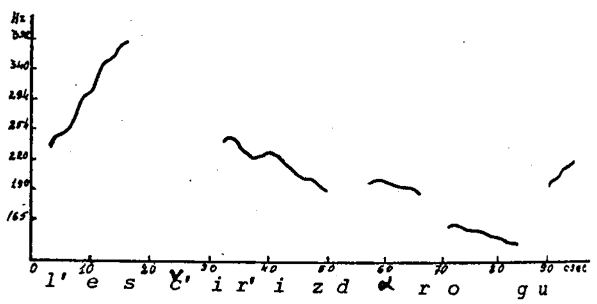
- à l'intérieur de chaque proposition : les thèmes *učastok* ("le terrain"), *dom* ("la maison"), *les* ("la forêt") sont chacun mis en relief par une montée en flèche du ton sur la syllabe tonique, car on vérifie qu'à chacun d'eux il convient bien d'attribuer le rhème R ("T est conforme à ce que l'on désirait") et non R'. Sur ce plan, l'analyse de l'exemple 8 est identique à celle de l'exemple 2 donné plus haut.

- à l'intérieur de l'énoncé complexe : chacune des propositions p_1 (*učastok xorošij* : "l'endroit est bien"), p_2 (*dom brevenčatyj* : "la maison est en rondins") et p_3 (*les cerez dorogu* : "la forêt, il n'y a qu'à traverser la route") fonctionne comme thème vis-à-vis de la conclusion q (*nu počemu ne snjat'* : "pourquoi est-ce qu'on ne louerait pas ?"). Chacune d'elles est un argument destiné à faire changer d'avis l'interlocuteur qui paraît, lui, adhérer à la proposition contraire q' ("il ne faut pas louer"). L'énumération p_1, p_2, p_3 est présentée comme incompatible avec la proposition q', ce qui entraîne la mise en relief de chacune des propositions qui la composent par un mouvement descendant-montant sur leurs prédicats respectifs (cf. figure 4).

- pluralité au niveau des occurrences : chaque fois que T se répète, il est associé à R et exclut R'.

9 (A un bachelier qui, avant de s'inscrire à l'université, est allé demander des renseignements à des étudiants plus âgés.)

— *Nu kak, ty našël kogo iskal ?*



fugure 4

— *Da net.*
— *Nu konečno, a čto ty udivljaeš'sja? Èkzameny / končilis', i studenty raz"exalis'.*

"— Eh bien, tu as trouvé qui tu cherchais ?
— Ben non !
— Evidemment, ça n'a rien d'étonnant. Les examens terminés, les étudiants sont tous partis (littéralement : "les examens sont terminés et les étudiants se sont dispersés.").

Cet exemple diffère de l'exemple 4 donné plus haut en ce que cette fois l'énoncé p et q n'est pas destiné à rejeter un présupposé q' de l'interlocuteur. Celui-ci a en effet déjà pu se convaincre par lui-même que les étudiants n'étaient plus là. Le but de l'énoncé est ici d'indiquer que l'événement q n'a rien d'étonnant, puisqu'il n'est que la conséquence d'une loi générale : chaque année, la fin des examens est toujours immédiatement suivie du départ des étudiants. Là encore, la division en thème et rhème doit être étudiée à deux niveaux :

- à l'intérieur de la proposition p : le thème *èkzameny* ("les examens") est mis en relief par une montée en flèche du ton car le rhème R qui lui est attribué (*končilis'* : "sont terminés") est opposé au rhème inverse R' (*ne končilis'* : "ne sont pas terminés"). Ce n'est en effet que lorsqu'on a R et non R' que l'événement q ("le départ des étudiants") peut prendre place.

- à l'intérieur de l'énoncé complexe p et q : la première proposition fonctionne comme thème vis-à-vis de la seconde. L'événement p est présenté comme excluant q' toutes les fois qu'il se représente, ce qui entraîne la mise en relief du prédicat *končilis'* par un mouvement descendant-montant du ton. On remarque que cette référence à une loi générale donne à la réplique un ton légèrement sentencieux. Nous pensons que c'est peut-être à cause d'exemples

de ce type que les linguistes considèrent généralement que l'intonation descendante-montante est caractéristique du style didactique.

3.2. Quant à l'intonation réalisée par une montée du ton suivie d'une stabilisation (T / R, *IK-6* dans la terminologie de E. A. Bryzgunova), nous n'en avons pas encore achevé l'étude, mais il semblerait qu'elle apparaisse préférentiellement lorsque la relation entre le thème et le rhème est présentée comme allant de soi (il s'agit souvent d'une relation préexistante).

4. Les trois intonations de mise en relief thématique que nous avons distinguées au début de cette étude, loin d'être interchangeables, ont donc chacune une signification précise, et leur apparition dépend de critères non pas stylistiques, mais strictement syntaxiques. Cela confirme, s'il en était encore besoin, que l'intonation ne sert pas seulement à exprimer les affects, mais qu'elle joue également un rôle syntaxique important, puisque ce sont les liens que l'énoncé entretient avec le contexte qui déterminent le choix entre les différents contours prosodiques que le locuteur a à sa disposition.

BIBLIOGRAPHIE

/1/ E. A. Bryzgunova, *"Zvuki i intonacija russkoj reči"*, Moscou, 1977.
/2/ C. Bonnot, I. Fougeron, "Enumérations en russe moderne : étude prosodique et syntaxique", in "Bulletin de la Société de Linguistique de Paris", t. LXXIX, f. 1, 1984.
/3/ C. Bonnot, I. Fougeron, "Intonation de 'non-finalité' dans les énoncés coordonnés en russe moderne", in "Les particules énonciatives en russe contemporain", Institut d'études slaves, Paris, 1986.
/4/ C. Bonnot, I. Fougeron, "Différents types de relations exclusives entre thème et rhème (sur la base d'une analyse prosodique)", à paraître dans les Actes du V^e colloque de linguistique russe, Poitiers 14-16 mai 1987, Institut d'études slaves, Paris, 1987.

SEMANTIC DIVERSITY IN INTONATION

Ivan FONAGY

Centre National de la Recherche
Scientifique Paris (France)

1. Even if *intonation* is taken in the restricted sense (as a synonym of 'tune' or 'speech melody'), what the term covers is highly heterogenous, both at the level of content and expression, as well as in the relation between expression and the content expressed.
Lecturers in French as a second language are compelled to teach such intonation patterns as the triangular pattern (rise of an augmented fifth, fall of a major third) of questions expressing doubt and disbelief; or that of sudden rise expressing evidence in face of doubt, in contrast to an apparently quite similar rising pattern, that of neutral *Yes/No* questions. Hardly any teacher would, however, feel obliged to teach how to express anger or tenderness in French by vocal means. It seems advisable to distinguish between *primary emotions* [29], *social attitudes* and *modalities (moods)*, both as the level of expression and content.

2. Primary emotions, such as anger, hatred, joy, fear or tenderness, are *reflected simultaneously at all levels* of the vocal apparatus : at the respiratory and glottal level as well as at the pharyngeal and oral level. One type of anger (angry quarrel) is reflected in English, French and Hungarian by features such as : forceful expiration; imperfect voicing; higher tension in the articulatory organs; increase of the maxillary angle; withdrawal of the mandible; withdrawal of the tongue in vocalic segments etc. [19]; at the same time, angry quarrel is characterized by at the prosodic level by a rigid metrical pattern with equally distributed heavy stresses; a rigid melodic base-line interrupted by sudden rises, resulting in a peaking melodic contour [14]. The relevance of the prosodic patterns could be confirmed by means of semantic tests based on laryngographic recordings [12] and synthesized variants [17]. Anger and tenderness are easily identified by the listeners even in the absence of voice production. (Angry whisper is rarely confounded with tender whispering.)

3. The vocal expression of *social attitudes*, as opposed to primary emotions is clearly *confined to the glottal level*, and can easily dispense with oral mimetics. Attitudinal intonation patterns are at the same time more language dependent than emotional vocal displays [8] [14]. Parallel tests with 20 French and 22 Hungarian informants show that both French and Hungarian subjects clearly distinguish angry and tender variants of *Viens ici, regarde un peu ce qu'ils font !* At the same time, Hungarian informants, in contrast to French subjects, were unable to identify correctly the triangular echo question expressing disbelief, *Il était là ?!* [= 'You pretend he was there. I don't believe it'], or to distinguish the indignant exclamation *Il était là !* [= 'He was there ! why do you pretend he was not ?' from the neutral question *Il est là ?* Clearly, language dependent (unpredictable) features seem to play a more important role in the case of signalling definite attitudes than in that of primary emotions. Such melodic clichés are strictly patterned. Thus, the tune of indignant exclamation contrasts with the interrogative pattern essentially by the steeper rise : quantity turns into quality : two patterns differing in grade function as different configurations. According to semantic tests presented in a previous study [15] the quotient *rise in semi-tones/duration in csec* is more than 1 in sentences perceived as exclamations, and less than 0.5 in those perceived as Yes/No questions.
A similar precision is required in the case of some twenty melodic clichés in contemporary Parisian French [16]. E.g. the cliché expressing, among others, a gentle approach requires a *descent in quarter-tones*. According to synthesized variants presented to French listeners, a descent in halftones proved to be 'unacceptable'. The compelling force of the clichés clearly appears in the course of mimicking experiments. The average deviation between the individual reproductions of the stimulus was 0.82 quarter-tone in the case of the cliché *Oh qu'il est mignon !*, and of 2.37 in that of a neutral statement. Regular interval simply intrasyllabic regularity of laryngeal vibrations, perceived as chanting [26: 65], as stylized speech [25: 169-196]. To account for differences in intrasyllabic regularity we have to stipulate a third dimension of speech melody, characterized by different degrees of perceived *melodicity* and we also need appropriate measures [20: 39-41]. The degree of intrasyllabic regularity approximated by means of different measures proved to be significantly higher for melodic clichés than for plain discourse [16].
No wonder that foreign speakers often fail to reproduce such melodic patterns accurately. Melodic divergences are perceived by native French speakers as a kind of 'melodic accent'. Semantic tests based on the presentation of laryngographic recordings of French sentences spoken by Chilian, Hungarian [18], Japanese [27] show that French listeners detect the 'foreign accent' most easily if the speakers improperly reproduce melodic clichés. Paradoxically, a French speaker may be taken for foreign, if he does not make use of melodic clichés at all [18].

4. The high degree in tonal precision corresponds to a higher degree of semantic organization. The expressed attitudes are definitely hearer-oriented, and linked with typical social situations. They refer to these situations without providing a further, conceptual analysis of the situations they are pointing at. They have no descriptive function, in contrast to lexical tones. However, they pave the way for conceptual analysis. We made an attempt to define the semantic field of French melodic clichés (a) by grouping the situations which frequently elicit them; and (b) by means of semantic tests where the informants were invited to associate sentences with the wordless 'utterances', i.e. with the clichés presented in a filtered version (recorded with a laryngograph). The informants assign to the *slightly descending* cliché the following meanings: (a) tender approach, (b) gentle warning, (c) complaint, (d) joyful surprise, (e) longing, (f) invitation, in form of a conditional phrase *(Si on allait boire un pot ?)*, (g) Yes/No question introduced by an interrogative morpheme, (h) allusive, elliptical questions *(Et l'année prochaine ?)*, (i) disbelief, refusal *(Mais qu'est-ce que tu racontes ?)* [16]. We could hardly trace back all the different uses of *slight descent* to a basic meaning, in the same way as Robert Ladd attempted to interpret the divers uses of the English falling-rising tune as derivatives of the basic meaning 'focus is in the given set' [25: 52-162]. 'Damping, softening' (sordino) seems to be a semantic distinctive feature common to (a), (b), (e), possibly to (c) and (d). In other cases we cannot dispense with the concept of 'melodic homonymy' [30: 137-146].
Let us add that meaning (i) of the smoothly descending cliché is, in fact, more elaborate than the label seems to suggest. The cliché is elicited by a specific situation. The interlocutor *I* makes a remark or a suggestion. The locutor *L* refutes the statement or rejects the proposal, giving expression to his surprise ('How could you say/propose such a nonsense ?').
It is a significant that in a number of cases the attitudes expressed by such intonation patterns can be also conveyed by means of lexical or grammatical morphemes, conjuctions [1] or adverbs ('modals') without any semantic loss [32] [25: 121-123]. In Hungarian the word *hiszen*, a derivative of *hiszem* 'I believe it' can be considered as a translation of the melodic cliché consisting of *a fall of a fourth followed by a rise of a flat third*. If this pattern is assigned to the sentence *Megmondtam* 'I told it', the utterance implies : 'You pretend that I didn't tell *p*.. This is not true : I did tell *p*. I am really surprised that you pretend I did not'. The sentence *Hiszen megmondtam* would have exactly the same implication. This is certainly not true for 'super-sentences' such as 'I tell you in anger that...' which could hardly be considered as an equivalent of the vocal expression of anger, as proposed by Yorio [34].

4. Modal intonation patterns represent the highest level of semantic organization that can be reached by prosodic means. We could even be tempted to attribute a referential function (Darstellungs-Funktion) to modal intonation. Roman Jakobson is, however, probably right in rejecting such a claim : "The interrogative sentence is not a reference but only a kind of appeal for reference" [23 : 281]. Intonation had to cover, nontheless, a long distance in semantic space to become from a mere reflection of emotional states a mark of a modal category.
It is not easy to draw a demarcation line between moods (modal categories) and attitudes. Modal categories correspond to the most essential, the most general attitudes. Verbal communication could not do without them. Attitudinal intonation patterns are always felt as stylistically marked. Modal intonation patterns may be neutral, stylistically unmarked. Stylistic markedness is, however, an elusive feature. The most satisfactory way of tracing the demarcation line between attitudes and modalities is offered by the grammar itself. Nonmarkedness and generality is acknowledged by the grammar's providing grammatical morphemes in order to distinguish different moods. This implies that languages might widely differ in this respect. Most languages have non-prosodic (segmental) markers for Yes/No questions. Few languages have, however, grammatical markers for such moods as *probabilitive, necessitive, precative, pejorative* mood that are inherent features of the verbal system in Vogul [24]. Consequently, we will have to consider 'imploring' as an emotive attitude for Indo-European languages, and as a grammatical mood as far as the Vogul is concerned.

5. Even if clear cut demarcation lines could be traced between basic emotions, social attitudes and moods, this would certainly not prevent melodic configuration switching unperceived from one category to the other [9: 55]. Thus, attitudinal melodic patterns may metamorphose into neutral indicators of mood. In English, German or Hungarian, the steep rise and sudden fall characterize in English, German or Hungarian control-questions expressing mistrust or irony in connection with a previous statement (e.g. *Really ?*, Hung. *Jo ?* 'Well ?'), as it appears from the listener's reaction in the face of synthesized variants [12]. The speaker seems to echo the partner's categorical statement ironically exaggerating its melodic profile. This stylistically marked melodic form became the dominant, unmarked intonation pattern of Russian Yes/No questions probably in the first half of this century [3] [4]. Similarly, during the last decades, an *intonation metaphor* - the transfer of interrogative melody to imperative sentences - yielded in Hungarian a new category of mood, that of (gentle) invitation vs. (categoric) order [10]. According to Dwight Bolinger the transfer of (interrogative) final rise to assertive sentences "will be probably taken in some sense of *Why do you ask ?*" [2].

6. The three kinds of melodic patterning represent *different levels of human signalling behaviour*, and mental elaboration. The vocal expression of primary emotions can be considered as a reduced reproduction of some fundamental ancestral activities [7]. As formulated by G.W. Crile [5] : anger is a phylogenetic product of fight, fear reproduces flight. Similarly, Plutchik retraces primary emotions to prototypic adaptive patterns : anger to destruction,

fear to protection against threat, joy to reproduction, disgust to rejection [29: 160]. (For a detailed and pertinent analysis of the theories of emotion and their vocal expression, see Scherer [31]. Social constraints reduced real activity to an acting-out at the level of the respiratory, laryngeal and oral level. I attempted in previous publications to draw a preliminary sketch of the vocal encoding of emotions [14]. I should lay emphasis on its high complexity and diversity. Thus, we have to distinguish a primary glottal gesturing and a secondary vocal mimicking. The expression of hatred by means of *a strangled voice* - due to an excessive innervation of the constrictors, a reduced and quite harmless form of strangling the partner or a third person - could exemplify a direct acting-out of emotion at the glottal level. Secondary tonal mimicking is based on the perception of pitch as spatial movement. The rigid melodic base line interrupted by sudden rises in angry arguments, and sweetly undulating melodic line in tender speech are such forms of projective tonal gesturing.

Direct acting-out of emotions at the different levels of the speech apparatus plays a secondary role in the expression of emotive attitudes. The messages are conveyed essentially by melodic movements. Speech melody is the only prosodic vehicle of grammatical moods. Emotive vocal patterns are *ex-pressions* in the literal sense of the word. As far as we can dissociate the form and content intermingled in the acting out of emotions, the content is directly present at the level of expression. Despite the partial overlapping of expression and content, the forms of melodic expression of primary emotions are language dependent. Intonation patterns at all the three levels are motivated (iconic) conventional signs. They differ, however, in the degree of motivation (iconicity). The motivation is more subtle and complex in the case of attitudinal intonation patterns than in the expression of primary emotions. The isomorphism between melodic movement and semantic content may be concealed by previous transfers in modal intonation patterns.

The *polysemy of intonational patterns* is either due to the genuine polyvalence of melodic gestures, or to the metaphoric use of a pattern. Genuine polyvalence may be the source of the polysemy of 'open' forms highlighted by Alen Cruttenden [6]. The multiple meaning of the Hungarian rising-falling pattern, conveying polite solicitation as well as interrogation, results from recent melodic transfers. The same melodic pattern may express different attitudes in function of its 'phonetic context'. According to semantic tests based on the presentation of 46 variants of the pseudo-Hungarian 'sentence' /'kisera 'mera 'ba: vatag/ to 25 university students, falling-rising melodic pattern tends to be interpreted as the expression of *coquetry* if a swift final rise is palliated by a drop of intensity; as a *menace* if the falling-rising melody is accompanied by a parallel fall and rise in intensity [14].

Let us add, that recurrent (typical) confusions of vocally expressed attitudes, "remarkably stable" errors [8: 144 f.], may *reveal* more or less *hidden analogies* between attitudes. Thus, in semantic tests based on laryngographic recordings [12] 'hatred' was frequently interpreted as 'disdain' or 'reproach'; all these affects share an element of aggression. The semantic tests on the emotive version of /'kiser 'me:ra 'ba:avatag/ offer some other revealling surprises : 'jubilation' is confused with 'anger' (Davitz refers to the same confusion o.c. 144); this points at a possible common feature : violent emotion and its sudden discharge. Fight and flight are both emergency responses in face of a conflictual situation, involving the activation of the sympathetic autonomous nervous system. This could account for the confusion of anger and fear [33], corresponding to fight and flight, in the framework of the Darwinian theory of emotions. The most unexpected typical error committed by the foreign informants of the /'kisera/ test was the confusion of 'anger', 'menace' and 'argumentation'; and the confusion of 'logical deduction' and 'command' or 'peremptoriness'. The confusion seem to corroborate hypotheses regarding the aggressive instinctual basis of logical reasoning, as formulated by I. Hermann [22]. Metaphoric expression such as *sharp intellect, scientific rigor,* the French nominal phrase *esprit tranchant*, or the recent semantic development of the word *argument* might reflect our preconscious knowledge of such parallels. Such cases of typical errors can be best interpreted is the frame-work of Klaus Scherer's component patterning theory, a biologically based distinctive feature analysis of emotional processes [31: 215-222].

In view of the fundamental diversity of the encoding of primary emotion, on one hand, and social attitudes and modalities on the other, it is most unlikely that all kinds of intonation should have the same neuro-physiological underpinning. The production and analysis of highly conventional distinctions, such as sharp and slow rise in French assertive exclamation vs. question, incomprehensible without the mastery of French, could hardly be produced and analysed without the participation of the left hemisphere.

The *superposition and integration* of two of three different melodic patterns is a further source of semantic complexity. The components of the complex pattern can be dissociated by means of mimicking tests. It characterizes artistic vocal performances [13]. The integration of two different melodic configurations may give rise to a new melodic pattern. Thus, in Hungarian, the superposition of the rising-falling question melody and the falling tune of surprise resulted in a straight melodic line at mid-high level with a final fall in the last syllable, and is the recurrent, conventional melodic expression of surprised questions. The integration of accent, tone and intonation in a complex curve is probably the general way to convey multiple prosodic information [28] [21]. The integration of different attitudinal and modal patterns may announce a new phase in the evolution of melodic encoding.

REFERENCES

[1] Ch. Bally, "Intonation et syntaxe", Cahiers Ferdinand de Saussure 1 (1941), 33-42.

[2] D. Bolinger, "Intonational signals of subordination", in: Proc. 11th Annual Meeting of the Berkeley Lingu. Soc. 1984, 401-414.

[3] S.C. Boyanus, "The main types of Russian intonation", in: Proc. 2nd Int. Congress of Phonetic Sciences, Cambridge, 1936, 100-113.

[4] B.A. Bryzgunova, "Prakticesjkaja fonetika i intonacija Russkogo jazyka", Moscow-Moscow University Publ. 1963.

[5] G.W. Crile, "The origin and nature of emotions", Philadelphia-Saunders, 1915.

[6] A. Cruttenden, "Falls and rises : meaning of universals", Journal of Linguistics 17 (1980) 77-91.

[7] Ch. Darwin, "The expression of emotions in man and animals", London-Murray 1872.

[8] J.R. Davitz, L.J. Davitz, "The communication of feelings by content-free speech", in: J.R. Davitz ed. The communication of emotional meaning, New York-McGraw-Hill, 1964, 143-156.

[9] O. von Essen, "Grundzüge der Hochdeutschen Satzintonation". Rattingen-Henn, 1956.

[10] J. Fónagy, "Métaphores d'intonation et changements d'intonation". Bull. Soc. Lingu. Paris 64/1 (1969), 22-42.

[11] I. Fónagy, "Synthèse de l'ironie", Phonetica 23 (1971), 42-51.

[12] I. Fónagy, "A new method of investigating the perception of prosodic features", Language and Speech 21 (1978), 34-49.

[13] I. Fónagy, "Artistic vocal communication at the prosodic level", in: Current Issues in the Phon. Sci. Amsterdam-Benjamins, 1979, 235-244.

[14] I. Fónagy, "Emotions, voice and music", in: Research aspects on singing. Stockholm-Royal Swedish Academy 1981, 51-79.

[15] I. Fónagy, E. Bérard, "Questions totales simples et implicatives", Studia Phonetica 8 (1973), 53-97.

[16] I. Fónagy, E. Bérard, J. Fónagy, "Clichés mélodiques du français parisien", Folia Linguistica 17 (1983), 153-185.

[17] I. Fónagy, J. Fónagy, J. Sap, "A la recherche des traits pertinents prosodiques : hypothèses et synthèses", Phonetica 36 (1979), 1-20.

[18] I. Fónagy, M. Guzman, E. Bérard, "Comment mesurer l'accent d'intonation ?" Travaux de l'Inst. Lingu. et Phonet. 2 (1976), 41-61.

[19] I. Fónagy, M.H. Han, P. Simon, "Oral gesturing in two unrelated languages". Quantitative Linguistics 19, Bochum-Brockmeyer 1983, 103-123.

[20] I. Fónagy, K. Magdics, "Das Paradoxon der Sprechmelodie", Ural-Altaische Jahrbücher 35 (1963), 1-55.

[21] E. Gårding, "The relation between sentence and word prosody", Proc. 9th Int. Congress Phon. Sci. vol. 2, 1979, 375-379.

[22] I. Hermann, "Psychoanalyse und Logik", Leipzig-Psychoan Verlag 1924.

[23] R. Jakobson, "Zur Struktur des Phonems", in : Selected Writings 1, The Hague-Mouton 1971, 280-310.

[24] B. Kálmán, "Vogul chrestomathy", Bloomington-Indiana Press, 1965.

[25] D.R. Ladd, "The structure of intonational meaning", Bloomington-Indiana University Press, 1980.

[26] M.Y. Liberman, "The intonational system of English", Bloomington-Indiana University Linguistic Club, 1978.

[27] S. Nakamura, "Contribution à l'étude des interférences prosodiques", Ph.D. Thesis, Paris-University of Paris III, 1978.

[28] S. Öhman, "A model of word and sentence intonation", Reports of the 6th Int. Congress of Acoustics 2 B 54, 1968, 163-166.

[29] R. Plutchik, "Emotion : a psychoevolutionary synthesis", New York-Harper & Row, 1980.

[30] M. Romportl, "Studies in phonetics", Prague-Academia, 1973.

[31] K. Scherer, "Vocal affect signalling", in J.S. Rosenblatt et al. eds. Advances in the study of behavior vol. 15, New York-Academic Press, 1985, 189-244.

[32] M. Schubiger, "English intonation and German modal particles", Phonetica 12 (1965), 65-83.

[33] C.E. Williams, K.N. Stevens, "Vocal correlates of emotional states", in: J.K. Darby ed. Speech evolution in psychiatry, New York-Grune & Stratton, 1981, 221-240.

[34] C.A. Yorio, "The generative process of intonation", Linguistics 97 (1973), 111-123.

PROPERTIES AND FUNCTIONS OF THE PROSODIC PHENOMENA IN LANGUAGES

Pavle Ivić

Serbian Academy of Sciences, Belgrade, Yugoslavia

ABSTRACT

Prosodic distinctions are quantitative and relational. Their perception implies comparisons on a syntagmatic axis, which involve the element of time. All these properties are corrolaries of the basic fact that sound intensity, pitch and duration are dimensions of the sound signal.

The functions of the prosodic phenomena in languages are determined by their nature.

PROPERTIES OF THE PROSODIC PHENOMENA

The term "prosodic" is used in this paper to denote all linguistic phenomena based on sound intensity, pitch or duration. These phenomena are interconnected by frequent co-variation (for instance, in many languages the prominence of the accented vowel is implemented simultaneously by greater intensity, increased duration and higher pitch) or by implication and incompatibility rules (e.g., in Ancient Greek tone contrasts imply both accentedness and length, so that unaccentedness and shortness are incompatible with distinctive tones). However, such cases still do not entitle us to group intensity, pitch and duration together. This can be done only if they have at least one common property.

Various authors mention such properties, or hint at them, but I do not know of a scholarly work where an exhaustive list of such properties would be presented. Let us try to enumerate them now.

(1) Distinctions in the domains of intensity, duration and pitch have a quantitative character. They do not imply the presence or absence of a phenomenon (cf., for a different situation, nasality, voicing, rounding, aspiration etc. in the realm of the inherent features), but a variation as to the amount of a phenomenon, always along a continuous scale.

(2) All three distinctions considered here are relational. Since pitch, intensity and duration of sounds vary greatly depending on individual properties of speakers and on communication circumstances, it is usually impossible to give absolute numerical values for these phenomena, both in general and in a particular language. Thus, in order to determine the linguistically relevant prosodic characteristics of a sound, we have to compare it to some other sound(s) within the same spoken chain.

(3) This comparison takes place on a syntagmatic axis, in contradistinction to the mere paradigmatic comparison which suffices for the perception of inherent features. The perception of stress and of intersyllabic tone involves a comparison between syllables, and the perception of intrasyllabic tone ("Tonverlauf") is based on the comparison between various points within the given syllable (or, more exactly, the given syllable nucleus). Such syntagmatic comparisons must precede the comparison to the other member of the phonological paradigm. Quantity contrasts imply even three relations:

a) the duration of the given vowel, i.e., the time distance between its beginning and its end;

b) the ratio of this duration to the duration of other sounds within the given spoken chain;

c) the paradigmatic contrast between that ratio and a different ratio in the word which constitutes the other member of the contrast.

(4) The syntagmatic comparison involves the element of time: it is necessary to compare points at a certain time distance within the same utterance.

Our enumeration of the properties of prosodic phenomena should not be concluded

here. Another fact, usually unmentioned in the existing scholarly literature, also deserves our attention, and I would like to insist on this fact.

(5) All prosodic phenomena are based on variation in the dimensions of the acoustic signal. Dimensions are properties of magnitude which can be mapped on a continuous scale, so that the reduction of this property to zero implies the disappearance of the object characterized by the property. Sound signals, being vibrations of material particles, have three dimensions: amplitude, which roughly corresponds to sound intensity, frequency, which is the basis of pitch, and duration. Thus, our three categories of prosodic phenomena cover the whole range of dimensions of the acoustic signal. These entities inevitably exist in all utterances, and even in all speech sounds (a certain reservation must be added in connection with pitch in some consonants, where the picture is more complex). In contradistinction to this, an inherent phenomenon may be present or absent in a sound; if it were present in all sounds, it would be unable to accomplish its distinctive function.

Now, the question arises: are all five properties enumerated independent from each other? Or does there exist a logical conditioning among them? And if so, which is the fundamental one?

I submit the following answer to these questions: the crucial fact is that prosodic distinctions are based on the dimensions of the sound signal, i.e. on elements which are ubiquitous in language. The other four facts are corrolaries to this one and constitute a logical chain. Since we have to do with phenomena which are present anyhow, and which are dimensions, absolute (yes or no) distinctions are precluded, so that contrasts are necessarily quantitative. And since the given elements also vary depending on communication circumstances and peculiarities of individual human voices, distinctions are not based on absolute numbers, but on relations - obviously within the same spoken chain, which means in the speech of the same speaker and under the same communication circumstances. Given the linear character of speech, relations within the spoken chain are always relations between different points in time. Thus we can conclude that the logical and the perceptual particularities of prosodic phenomena are conditioned by their physical nature.

In the opinion of many authors the fundamental characteristic of the prosodic phenomena is that they are connected with the concept of syllable. In fact, what matters here is not the syllable, but the vowel. The physical and perceptional properties of vowels render them the most appropriate domain for distinctions based on the dimensions of the sound signal. Only in exceptional cases do consonants carry contrasts based on pitch or intensity, and consonantal contrasts as to duration are less frequent than corresponding contrasts in vocalism. Since a syllable usually contains a vowel, a relationship between prosodic phenomena and syllables arises.

FUNCTIONS OF THE PROSODIC PHENOMENA

The linguistic functions of the prosodic phenomena are threefold:

(1) They play a decisive role in what is known as sentence intonation.

(2) They serve to divide utterances into words or syntactic groups, thus facilitating the act of communication.

(3) They may be distinctive on the level of word phonology.

Note that in written texts the first function is performed by punctuation signs, and the second function basically by blanks between words. The third function of the prosodic phenomena is shown only by certain writing systems (e.g., Greek, Czech, Hungarian), whereas other alphabets fail to furnish this kind of data (Russian, Lithuanian, Serbocroatian, English, Rumanian) or provide only partial information (Italian, Spanish, German).

The first function (sentence intonation) is an exclusive domain of prosodic phenomena. True, some morphemes such as interrogative pronouns or particles can assume certain functions belonging otherwise to sentence intonation, but we have to do here with morphosyntactic units, and not with inherent sound features. In the second, configurational function prosodic phenomena are much more common than the inherent ones. However, the latter features also can serve as boundary signals (by the, pauses, which so often appear in this function, also belong to the prosodic domain, since they are definable as a reduction of sound intensity to zero). As to the third function (word phonology), it constitutes the principal domain of inherent sound features. In all languages of the world those features serve to distinguish phonemes, which implies that they distinguish words, whereas in only a part of the languages - a very large part, to be sure - is that function performed also by prosodic features.

Looking at the same facts from another viewpoint, we may state the following: prosodic phenomena in the first function are universal, in the second function they occur in almost all languages, and in the third function in the great majority of languages. Thus the range of functions of prosodic features across languages includes all functions which can be accomplished by any sound element. This range embraces the whole field covered by the inherent features plus a vast domain particular to the prosodic phenomena.

The functional delimitation between prosodic and inherent phenomena is governed by their physical nature. The so often recurring types of semiotic information conveyed by sentence intonation require vehicles that are ever-present. This is why only variation in the dimensions of the speech sounds is suitable for this role. It is even hard to imagine how inherent features could accomplish such a function. Should segmental features be added or subtracted, or should this happen to entire phonemes? Anything of that kind would largely increase the complexity of linguistic patterns and possibly also interfere with the lexical message of the utterance. Basically the same is valid for the configurational, i.e., the culminative, demarcative and interrogative functions, which are so important for the understanding of the verbal message. True, inherent features of segmental phonemes, too, play sometimes a role in these domains, but such situations are marginal and incidental. Even in such instances the primary function of the segmental entities is to denote lexical (or morphological) meanings, but their distributional characteristics cause their presence to be interpreted by the hearer also as a signal helping him to determine the boundaries of a word or of a syntactic group.

The exclusive or predominant use of prosodic phenomena on the levels of sentence intonation and word configuration makes it possible to avoid an interference of signals pertaining to these domains with the segmental composition of the word. However, the question arises: what happens when sentence intonation patterns get superimposed to the prosodic characteristics of words in languages where these characteristics are relevant in word phonology? The answer to this question is very instructive. A coexistence is perfectly possible in spite of the circumstance that both systems of signs utilize the same physical substance, i.e., the variation in the dimensions of the sound signal. What gets realized is a vector-type compromise between the two systems. The prosodic shape of the word is materialized as a modification of the pertinent segment of the sentence intonation pattern, - or the sentence intonation is manifested as a modification of the prosodic shape of the word. This is possible because we have to do here with quantitative and relational contrasts; such compromises would be unimaginable in the realm of inherent phonological phenomena with their "yes or no" contrasts.

The vast field of lexical meanings necessitates the use of many more distinctions than can be supplied by the limited set of prosodic distinctive features. Only the more numerous and more variegated inherent sound phenomena can accomplish this role. In this case the position of the prosodic phenomena is peripheral. They are not indispensable and may be absent. True, in many languages they belong to the inventory of DF in word phonology, but even there the preponderance of inherent phonological phenomena is incontestable.

SOME OTHER CORROLARIES TO THE NATURE OF THE PROSODIC PHENOMENA

The abstract, relational character of the prosodic phenomena is responsible for the fact that speakers are less frequently aware of their presence, let alone of their nature. Segments defined by inherent sound features are more tangible. As a rule even the linguistically untrained speaker is able to describe, if asked, the segmental composition of

a word. As to forms differentiated only by their prosodic characteristics, the same speaker often "feels" that they are different, but usually he is unable to identify the units which carry the difference. This explains why so many writing systems do not note the prosodic phenomena, although they normally show all distinctions based on inherent sound features.

Also the difficulties experienced by persons trying to master a prosodic pattern in the process of foreign language learning are attributable to the abstract nature of the phenomena in this domain. The same circumstance influences the diachronic fate of prosodic contrasts in word phonology, which show much less stability than distinctions based on inherent features. To be sure, in the past of various languages we find many instances where an inherent feature, such as voicing or consonantal palatalization or aspiration, disappeared from the system, or was introduced into the system. But this always concerns only a fraction of the inherent features operative in the given linguistic pattern, whereas prosodic distinctive features can all disappear from the system. This happened for instance in Polish, in Upper and Lower Lusatian, in many Western Macedonian dialects, etc. Standard French tends towards the same goal by eliminating the last traces of vocalic quantity. Structural differences in the prosodic domain play an extremely important role in languages with complex prosodic patterns, such as Chinese, Thai, Japanese, Swedish, Norwegian, Serbo-Croatian, Slovenian or Latvian. Apparently the average lifetime of a prosodic distinction is shorter than that of an inherent distinction. The topic deserves further study. It would be a worthwhile task to calculate, on the basis of language histories known to us, approximate life-expectancy indexes for each DF, inherent or prosodic, or, for that matter, of various phonemes or categories of phonemes. The results would furnish a basis for further investigations seeking to establish the causes of the different fate of various entities involved.

We can conclude that the understanding of the true nature of the prosodic phenomena contributes to the understanding of their behavior in many respects.

ТОНАЛЬНАЯ ЭМФАЗА В РУССКОМ ЯЗЫКЕ

С.В.КОДЗАСОВ

Кафедра общего, сравнительно-исторического и прикладного
языкознания, филологический ф-т, МГУ, Москва II9899

РЕЗЮМЕ

Одна из функций интонации - маркировка элементов сообщения, нарушающих презумпции ("ожидания") коммуникантов. Возможные изменения "системы ожиданий" довольно разнообразны, однако все они кодируются общим просодическим средством, которое мы называем "тональной эмфазой".

ИНТОНАЦИОННЫЕ СРЕДСТВА РУССКОГО ЯЗЫКА

I. Тональная эмфаза - некоторая стандартная модификация обычных (неэмфатических) интонационных средств. Поскольку наша система интерпретации и записи этих средств существенно отличается от общепринятых [1], [2], необходимо дать ее краткое изложение (см. также[3]).

Интонация предложения определяется двумя компонентами: уровнями базового тона и акцентами. Базовый тон (БТ) может иметь 4 уровня: 0 - низкий, I - средний (нейтральный), 2 - средне-высокий, 3 - высокий. Изменения БТ не связаны со словесными ударениями и происходят на границах семантико-синтаксических составляющих. Нейтральный уровень БТ в записи не обозначается, прочие уровни обозначаются надстрочной цифрой при круглой скобке, которая указывает конец группы, имеющей маркированный БТ (см. примеры (I8), (I9) и др.).

Имеется два типа акцентов: скользящие и уровневые. Основные скользящие тоны: ´- восходящий и ` -нисходящий. Они используются в "общих" высказываниях, прежде всего, в общих вопросах и ответах:

(I) А. - Он верну́лся из отпуска?
Б. - Верну̀лся.

Падающий тон встречается и в спонтанных утверждениях - при подтверждении осуществления предполагаемого факта:

(2) - Ваня уже приѐхал.

За подъемом тона на ударном слоге обычно следует возвращение к исходному уровню на заударном, а падению тона на ударном предшествует подъем на предударном. Такие автоматические движения, обеспечивающие поддержание общего нейтрального уровня БТ, в транскрипции не отмечаются.

Для уровневых акцентов важно не скольжение тона на протяжении ударного гласного, а наличие "скачка" с БТ на другой уровень. В случае положительного акцента происходит скачок вверх-вниз (знак ˈ), в случае отрицательного - скачок вниз-вверх (знак ↓). При нейтральном БТ ˈ означает фигуру I-2-I, при средне-высоком - фигуру 2-3-2. Аналогично, ↓ может означать фигуры I-0-I и 2-I-2. Обычно смещение и возврат к БТ происходит в пределах ударного слога, однако при положительном акценте подъем возможен и на предударном начальном слоге - с возвратом на ударном (в начальной его части).

Положительный акцент - наиболее частый из всех акцентов. Он используется в "специальных" вопросах и ответах на них, а также в спонтанных специальных утверждениях:

(3) А. - Ты когда̍ к нам придешь?
Б. - В суббо̍ту.

(4) - Ваня приезжает в суббо̍ту.

Отрицательный акцент также используется в "специальных" высказываниях, обычно он придает сообщению дополнительный смысл очевидности ("само собой"):

(5) А. - Ты когда к нам придешь?
Б. - В суббо̍ту, как обы↓чно.

Уровневые акценты могут объединяться в фигуру, которую мы называем "композицией" и обозначаем "углами":

(6) - В како⌜м ваго⌝не вы едете?

Эта запись означает, что после подъема на средне-высокий уровень в начале слога ком тон не падает вплоть до начала слога го. То есть, вместо последовательности ˈ...ˈ (акустически /\.../\ ) мы имеем ⌜...⌝(акустически /‾...‾\ ). Отрицательные акценты также могут композиционно объединяться:

(7) - Но са⌞м такой автор вообще ничего
не понимает.

В композиции возможен паузальный разрыв, в этом случае первый из акцентов становится скользящим:

(8) - Это живо⌜тное относится к хи⌝щникам.

(9) - Это живо╭тное - хи⌝щник.

Акцент ╭ отличается от ⌜ иным характером восходящего движения тона, а от ´ - отсутствием заударного падения тона.

Акценты вида ╭ и ⌜ используются (как показатели "незавершенности") и вне композиции, в частности, в сочинительных конструкциях:

(I0) - Он работает во вто╭рник, пя╭тницу
и суббо̍ту.

(II) - При каких обстоя⌜тельствах, когда⌜,
ке̍м допущена ошибка?

Удобно иметь для всех акцентов, при которых нет возврата к БТ, общее название; подходящим представляется слово "полуакцент": ⌜- положительный полуакцент, ╭ - восходящий полуакцент, ⌞ - отрицательный полуакцент.

Всякий акцент в общем случае относится к семантико-синтаксической составляющей, которая может быть выражена как одним, так и несколькими словами (в том числе и целым предложением). Если акцент относится к неоднословной составляющей, то все слова внутри нее произносятся быстро и слитно. Это отмечается в записи заключением группы в скобки, которые повторно помечаются знаком акцента:

(I2) - Это кто̍ был? (Твой ленинградский
знако́мый)´?/ (Тот парень, о кото-
ром ты мне говори́л)´?

(I3) - (Выста̀вка достижений народного хо-
зя̍йства)ˈ - откры́та сегодня?

Если имеет место пословное акцентирование всей фразы, то используются обозначения Aˈ или A↓ при закрывающей скобке:

(I4)-Хочу тебе напо̍мнить: (ты завтра
в пять идешь к зубному) Aˈ.

(I5) - Вообще↓: (всякое дело требует
внимания) A↓.

2. Тональная эмфаза выступает в двух вариантах. Если ей подвергается однословный член, то она обычно реализуется удвоением интервала тонального изменения, что в записи выражается удвоением знака акцента: ˝, ˵ и ˶ вместо ´, ` и ˈ, соответственно. Сравните:

(I6а) А. - Вы встретили его ве́чером?
Б.-Да̀, вѐчером.

(I6б) А. - Вы встретили его именно
ве̋чером?
Б. - Да, именно ве̏чером.

(I7а) - Он сделал это ве̍чером.

(I7б) - Ока̍зывается, он сделал это
еще ве̎чером.

Отметим, что удвоение интервала падающего тона достигается за счет повышения начальной точки нисходящего скольжения до уровня 3, а не за счет снижения конечной точки до уровня 0 (3→I, а не 2→0).

Если эмфаза относится к неоднословной составляющей (это может быть и целое предложение), то повышается (до средне-высокого) уровень БТ всей группы, при этом акценты сохраняют одинарный интервал:

(I8) - Ока̍зывается, он сделал это еще
(вчера ве̍чером)^2.

(I9) А. - Это случилось но̍чью.
Б. - Непра̍вда: Горе̍ть начало (часов
в де↓сять)^2.

Эмфаза однословных членов не обязательно выражается удвоением интервала акцента. В случае ´, ` и ˈ возможно использование схемы, типичной для неоднословных составляющих:

(20) - Ока̍зывается, он сделал это еще
(ве̍чером)^2.

При отрицательном акценте на однословном члене эта схема представляется преобладающей:

(2I) - Ты ошиба̍ешься. Все ви̍дели, что
он вернулся (ве↓чером)^2.

Огласовка ве⇊чером (при нейтральном БТ) менее вероятна. Это предпочтение объясняется, очевидно, стремлением говорящего избежать выхода за пределы "нормального" диапазона тона (ниже уровня 0).

Эмфатическое удвоение интервалов наблюдается и в композициях акцентов. Примеры будут приведены ниже, при обсуждении семантики эмфазы. Пока укажем на необходимость отличать средне-высокий (эмфатический) уровень БТ от фонетически сходного повышения уровня тона в послеакцентной группе при инверсии порядка слов:

(22) - С вами говорит (Ивано⌜в Иван
Петрович)⌜.

Такое поддержание уровня тона после повышения служит для маркировки "сферы действия" акцента. Опасность смешения с эмфазой особенно велика, когда полуакцент приходится на начало группы:

(23) А. - А где Ма⌜ша? Б. - (О⌜кна моет)⌜.
Сравните:

(24) А. - А где Ма⌜ша? Б. - Моет о̍кна.

В заключение отметим, что предлагаемая система интонационных единиц основана преимущественно на слуховом анализе речи. Строгое описание акустических коррелятов воспринимаемых элементов - отдельная задача.

ЭМФАЗА РЕМЫ

I. Рассмотрим семантику типичных случаев эмфазы ремы. Начнем с общих вопросов и ответов на них. Сравним два примера:

(25) А. - Ну, как ты, купил велосипед?
Б. - Да, купил. /Нет, не стал.

(26) А. - Ты что - купил этот велосипед?
Б. - Да, купил. /Нет, (взял напрокат)^2

В (25) речь идет об ожидаемом факте покупки велосипеда. Вопрос касается заполнения экзистенциального оператора "да"/"нет" при заданном факте. Пространство альтернатив тут фиксировано, и изменения презумптивного знания в результате ответа не предполагается. В (26) проверяется некоторая гипотеза, т.е. новый факт, который может быть включен в систему знаний. Пространство альтернатив здесь не фиксировано, и предполагается изменение презумптивного знания.

Различие вопросов к оператору и к факту можно проиллюстрировать также дизъюнктивными конструкциями:

(27) А. - Он уехал или нет? Б. - Уехал.

(28) А.-Он уехал или (ушел пешком)^2?
Б. - Уехал.

К "фактическим" вопросам примыкают уточняющие реплики типа:

(29) А. - Самолет из Уфы - прибыл?
Б. - Утренний?

Здесь эмфаза указывает релевантную информацию, ранее отсутствовавшую в презумциях вопроса коммуниканта А.

Фактические вопросы часто имеют отрицательную форму. Этим говорящий указывает на гипотетичность, малую вероятность факта, не позволяющую включить его в "систему ожиданий" до ответа. Этой же цели служит слово случайно:

(30) А. - Он случайно не во вторник приехал? Б. - Нет, в среду.

(3I) А. - Ты ее случайно (не Петру Ивановичу)^2 отдал? Б. - Да, (Петру Ивановичу)^2.

Отметим различие места скользящего акцента в отрицательной и положительной группах.

В вопросах с ли восходящий тон заменяется на положительный акцент с сохранением двойного интервала подъема; в положительном ответе сохраняется скользящий падающий тон:

(32) А. - Не во вторник ли это произошло?
Б. - Во вторник.

Способ оформления ли-вопроса к группе виден из такого примера:

(33) А. - (Не во вторник ли вечером)^2 это произошло?
Б. - Нет, (в среду утром)^2.

Фактические вопросы - главный, но не единственный класс случаев эмфазы ремы в общих вопросах. Второй класс составляют случаи "контрастивности" (в широком смысле). Здесь типично противопоставление верифицируемой гипотезы всем прочим альтернативам:

(34) А. - Только Иванов остался?
Б. - Да, только Иванов.

Возможны и аналогичные общие утверждения:

(35) - Именно Иванов должен нести ответственность.

Эмфаза и здесь указывает на смену установки: множество возможностей заменяется единственным фактом.

Еще один типичный случай эмфазы общего утверждения - контрастивный императив, предполагающий замену поведенческой установки:

(36) - Ты лучше на такси поезжай.

Другой случай контрастивности - несоответствие гипотетического заполнения верифицируемой позиции заполнению других позиций:

(37) - Ваш ребенок в морозы - гуляет?

Гуляние ребенка в морозы - маркированный факт, противоречащий стандартной "системе ожиданий", отсюда эмфаза.

2. Эмфаза рематических членов в специальных утверждениях указывает на то, что данное заполнение семантической позиции было неожиданным для говорящего, и предполагает сходную ситуацию со слушателем. Часто идея неожиданности выражена одновременно и лексическими средствами, особенно характерно употребление слова оказывается (это аналог слова случайно в общем вопросе). Примеры:

(38) - Он, оказывается, на истфаке учится.

(39) - Можешь себе представить, они уже (телефон получили)^2.

Подчеркнем, что заполнение нулевой презумпции эмфазу не вызывает:

(40) А. - Он где учится? Б. - На истфаке.

Это означает, что новое на пустом месте не есть изменение прежней "системы ожиданий".

В то же время как перестройка системы знаний трактуется установление некоторых семантических связей между событиями и объектами (причинные связи, взаимные отношения и проч.):

(4I) А. - Зачем ты ее привела?
Б. Да ей одной скучно было.

(42) А. - Почему ты вернулся?
Б. - (Забыл кошелек)^2.

(43) Все эти люди - родственники.

Второй класс случаев эмфазы ремы в специальном утверждении - контраст того или иного вида. Чаще всего это замена одного знания или намерения другим:

(44) - Последняя цифра не пять, а шесть.

(45) А. - А теперь мы идем в магазин.
Б. - В магазин мы пойдем завтра.

Другой тип контраста - несоответствие разных характеристик ситуации:

(46) - Уже десять, а еще светло.

Как мы видели, в случае общих высказываний эмфазе гораздо чаще подвергается рема вопроса, чем утверждения. В специальных высказываниях обратная картина: рема вопроса несет эмфазу очень редко. Обычно она выражает контраст позиций, подлежащих заполнению:

(47) - Меня не интересует, когда ты их отдала. Кому ты их отдала?

Специфический тип эмфазы - "экзаменационный" вопрос. Если ударение в вопросительном слове приходится на I-ый слог, то реализация положительного акцента обычна:

(48) - Кто руководил этим движением?

Если ударение падает на последующие слоги, то двойной акцент реализуется фигурой " " - на I-м слоге происходит двухинтервальный подъем тона, а на ударном - двухинтервальное падение:

(49) - В каком году это произошло?

Эта фигура представляет собой как бы эмфатический вариант композиции положительных акцентов, хотя реализует один акцент. Встречается, однако, и эмфаза нормальной двухакцентной композиции:

(50) - Он полную сумку белых набрал!

Возможны также двойные полуакценты:

(5I) - Откуда это взялось...?

3. Рассмотрим теперь случаи, когда эмфазе подвергается целое предложение. Это, прежде всего, может быть "нерасчлененное" предложение:

(52) - (Гроза начинается)^2!

(53) - (Пароход идет)^2!

Такие восклицания произносятся, конечно, и с повышенной громкостью. Соответствующие неэмфатические фразы огласуются по образцу предложений с инверсией:

(54) - (Гроза начинается).

(55) - (Пароход идет).

Эмфатический вариант нерасчлененных предложений используется для сообщения о событиях, имеющих высокую значимость и нередко требующих поведенческой мобилизации коммуникантов.

Сходную семантику имеет эмфаза и в случае предложений, неэмфатические корреляты которых являются расчлененными:

(56а) - Теперь заболел (Иван Иванович)^1.

(56б) - (Иван Иванович заболел)^2!

(57а) - На Кузнецком открылась (выставка Шилова)^1.

(57б) - (На Кузнецком выставка Шилова открылась)^2!

Как видно из примеров, при эмфазе происходит изменение порядка слов и утрата всех акцентов, кроме акцента на подлежащной составляющей. Однако, если подлежащее введено в предшествующей части дискурса, эмфатическое сообщение расчленяется:

(58) А. - Иванов, видимо, опаздывает.
Б. - (Иванов заболел)^2.

Сравните нерасчлененный вариант:

(59) А. - Что случилось?
Б. - (Иванов заболел)^2!

Заметим, что предложения с пословным акцентированием сохраняют все акценты при эмфазе:

(60) - Ты представляешь? (Иванов женился на Вале Петровой)^2 А^1.

Эмфаза вопросительной фразы обычно используется при введении гипотетической причины некоторого события:

(6I) - Ты почему вернулся? (Дождь пошел)^2?

(62) - Ты что такой грустный? (Опять поссорился с женой)^2?

Интересно, что далеко не всякое вопросительное предложение с такой семантикой способно подвергнуться эмфазе:

(63) - Ты что такой грустный? *(Тебя опять обругал учитель)^2?

ЭМФАЗА ТЕМЫ и РЕМЫ

Эмфаза темы, как кажется, обязательно требует эмфазы ремы. Сравним пары предложений:

(64а) - Импортные куртки бывают у вас?

(64б) - А импортные куртки - бывают у вас?

(65а) - В субботу можно к вам придти?

(65б) - В пятницу я не смогу придти.
А в субботу - можно?

Эмфаза ремы в (64б) и (65б) явно не имеет собственной семантической мотивации и обусловлена контрастивностью темы.

Вместе с тем в большинстве случаев

мы имеем параллельную семантическую контрастивность как темы, так и ремы:

(66) – Другие могут поступать как хотят. Мой сын останется дома.

(67) А. – Я не буду ждать. Б. (прочим присутствующим). – А кто будет ждать?

Типичный пример двойной эмфазы – контрастивные сопоставительные предложения:

(68) – Петя работает, а Ваня – учится.

(69) – Петя гуляет, а Ваня – (сдает экзамены)^2.

Фигура ⁄‾...‾╗ , очевидно, является эмфатическим вариантом разорванной композиции ⁄‾...‾┐. Встречаются и аналогичные отрицательные композиции:

(70) А. – В десять я уже пришел.

Б. – (Но в одиннадцать вас не было) .

Отметим, что при наличии двойных отрицательных акцентов средне-высокий уровень БТ, очевидно, обязателен.

Типичный случай использования отрицательной композиции – сопоставительные вопросы:

(71) А. – В театре мы уже были.

Б. – А на выставке – были?

Пример эмфазы такой композиции при контрасте:

(72) А. – На выставке Иванова мы уже были.

Б. – (А на выставке Петрова – были)^2?

При эллиптическом опущении ремы вопрос из (71) приобретает форму:

(73) – А на выставке?

Здесь используется одна из фигур незавершенности. Сравните:

(74) – Мы были в театре, на выставке и в кино.

Совершенно аналогично меняется при опущении ремы контрастная отрицательная композиция:

(75) – (А на выставке Петрова)^2?

ГИПЕРЭМФАЗА и АНТИЭМФАЗА

1. Резкое нарушение системы ожиданий приводит коммуниканта в аффективное состояние – удивление. Вызывающие удивление смысловые элементы кодируются увеличением тех тональных сдвигов, которые характерны для обычной эмфазы: вместо средне-высокого уровня используется высокий, вместо двойных акцентов – тройные.

Приведем примеры гиперэмфазы положительных акцентов:

(76) – Она такое вытворяла!

(77) – Он как прыгнет!

(78) – (Иванов выиграл машину)^3 А' !

Примеры восходящих акцентов:

(79) – Она совсем вернулась?

(80) – Ты (видел Олю)^3?

2. Те элементы сообщения, которые нарушают установки коммуникантов, несут бóльшую информацию, а тем самым имеют бóльшую значимость, чем элементы, не выходящие за пределы наличной системы ожиданий. Увеличение интервала акцентов, повышение БТ – это в конечном счете средства кодирования высокой информационной значимости данного фрагмента смысла.

Картина будет, однако, не полна, если мы не укажем в заключение на наличие средства кодирования тех смысловых элементов, которые, напротив, малоинформативны (избыточны, факультативны) в сообщении. Таким средством является низкий уровень БТ. Огласовку семантико-синтаксических составляющих низким уровнем можно назвать "антиэмфазой".

Низкоуровневую огласовку обычно получают составляющие высоких синтаксических рангов:

(81) – Тебе Оля – (я ее вчера видел)^0 – привет передавала.

(82) – "Садитесь!" – (сказал участковый)^0.

Однако антиэмфазе могут подвергаться и составляющие низких рангов:

(83) – Вам звонит Иванов (из отдела снабжения)^0.

(84) – Я тебе напишу (когда-нибудь)^0.

Какие элементы интерпретируются как малоинформативные и произносятся на низком уровне – задача отдельного исследования.

Литература

1. Брызгунова Е.А. Звуки и интонация русской речи. М., 1977.
2. Николаева Т.М. Фразовая интонация славянских языков. М., "Наука", 1977.
3. Кодзасов С.В. Интонация вопросительных предложений: форма и функции. В сб.: "Диалоговое взаимодействие и представление знаний". Новосибирск, 1985.

INTONATIONAL PHRASING AND ITS ROLE IN SPEECH COMMUNICATION

Olga Krivnova

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

The theoretical claim of this study is that intonational phrasing /IP/ functions in speech communication as one of language orientation means.

The paper presents the results of a psycholinguistic experiment, whose aim was to examine sense segmentation devices in two different situations: I/ in a written text having no formal signs of segmentation, whatsoever, 2/ in an oral text based on normally organized written speech sample.

INTRODUCTION

The cognitive background of modern linguistics presupposes the view that language as a means of communication can be adequately analysed only with references to the general principles of human speech behaviour.

It should be noted in this connection that the treatment of IP /i.e. division of an utterance stretch into phonosyntagmas, phrases and more lengthy speech units/ in Russian and Soviet linguistics emphasizes the direct relationship between this phenomenon and human speaking activity, viewing the former as inseparable from speech production and comprehension process.

This view is based on the works of the prominent Russian scholar - academician L.V.Scherba, who defined the minimal IP unit - phonosyntagma - as a phonetic unit conveying a semantic whole in the speaking-thinking process. Recent studies on IP are aimed at providing further evidence for the general idea of IP's relationship with speech activity. It seems that problems concerned with IP can't be solved without investigating the general principles of processing and verbalization of conceptual information as they carried out by the speaker. Therefore we think it useful to consider some ideas emerging in the framework of current theories of speech acts.

One of the most widely-known attempts to formulate the general rules governing speech behaviour belongs to the American scientist H.P. Grice /I/. He advanced an idea that speech communication is based on the Cooperative Principle in accordance with which participants of the communicative process interested in its successful realization, perform cooperative efforts to achieve understanding at minimal conceptual and linguistic expense.

The Cooperative Principle implies that the act of verbalization as any other component of the speech act must be oriented towards the recipient in advance. In the process of language coding the speaker must provide the addressee with the effective cues for semantic text interpretation and thus facilitate the assimilation of its content by the listener.

It is supposed that any language has special means which function as orientation devices in the process of verbal decoding. These devices may not be relevant for expressing the informational content as such, but serve to organize it properly, giving it point and precision and thus contributing to its adequate understanding.

In comparison with other language functions the function of orientation has at least two peculiarities. The first one derives from the fact that the linguistic form of a speech utterance depends on the general communicative intention of the speaker and the speech situation.

As early as in I920-30-s some Soviet scholars pointed out, in particular, that monologue as a form of speech communication presupposing more or less prolonged speech impact on the listener typically conveys complicated conceptual information. As a result, monologue is bound to set greater demands to the act of verbalization than dialogue. In connection with this a well-known Soviet scientist L.P.Jakubinsky wrote: "Монолог не только подразумевает адекватность выражающих средств данному психологическому состоянию, но выдвигает как нечто самостоятельное именно расположение, компонирование речевых единиц ... . Здесь сами речевые отношения становятся ... источниками появляющихся в сознании по поводу них переживаний... . На этой же почве возникают всевозможные явления синтаксического параллелизма и симметрии, т.к. сложность естественно вызывает какую-то организацию, построение" /2, p. 37/.

The second peculiarity consists in the fact that the use of language orientation means depends on the speaker's cooperative effort and on his evaluation of the intelligibility of his own utterance. As a consequence, orientation means are not subject to grammaticalization to the same extent as the means expressing the propositional context of a speech message.

To sum up, orientation means are related to such characteristics of the speech act as the complexity of information transmitted, time of speech impact, the listener's cognitive and cultural background and the current state of his mind, the degree of the speaker's cooperative effort and so on. All of these features point to a clearly pragmatic character of language orientation means.

Let's now turn to intonational phrasing. It is commonly agreed that the role of IP is more significant in monologue than in conversation form of speech. It is most actively used in texts that are complicated from the intellectual point of view and well-formed linguistically. In the following parts of this paper we are going to consider the ways in which IP can be used as an orientation device in the process of speech decoding. We'll also discuss the results of a psychological experiment that was undertaken with the purpose of illuminating of IP and revealing its functions in speech communication.

INTONATIONAL PHRASING AS A MEANS OF SENSE PROCESSING CONTROL IN SPEECH DECODING

It is a known fact that in a well-formed text IP correlates with its syntactic /in a wide sense/ structure. This is why IP presents a special interest for modeling speech decoding process. At the same time it's evident that IP is not the only possible means of transmitting information about the syntactic /deep and surface/ structure of a speech message. Syntactic relations between words, their correspondence to the general communicative intention of a speaker can be identified through morpho-syntactic cues /word order, form-words and morphemes/, lexical cues, verbal and situational context, the person's cognitive background and so on.

On account of such plurality of language means it seems justified to evaluate the contribution of each separate cue in the general procedure of sense decoding depending on whether understanding is possible in the absence of this particular cue. It is evident that IP /at least in Russian/ doesn't occupy a predominant position in the set of linguistic cues for decoding of speech message. It can be proved by various facts among which the following should be mentioned as the most significant:

I/ in speech reality utterances whose ambiguity is solved by IP solely are quite rare /as in "казнить нельзя помиловать"/;

2/ written texts with no formal segmentation marks can be perfectly understood: for example, Old Church-Slavonic and Old Russian manuscripts which are known to have neither punctuation marks, nor word gaps or capitals;

3/ unnatural word-by-word text pronunciation makes its understanding more difficult but not impossible.

From what has been stated above it follows that IP reduplicates the function of the lexical-syntactical cues that are constantly present in any written and oral utterance. IP can be regarded then as one of non-obligatory, redundant linguistic devices serving to increase the reliability of sense decoding in oral communication. This conclusion, however, doesn't take account of the most important characteristics of IP, and namely, the fact that IP separates speech currently and at the same time takes part in forming it as rhythmically organized time process.

Following many Soviet linguists and scholars in literature we consider the speech to be rhythmically organized if it is divided into the subjectively isochronic speech stretches with utterly or partly reproducible phonetic structure. IP's correlation with speech rhythm is recognized by the majority of Soviet phoneticians and there is no necessity to prove it here. At the same time we'd like to discuss some peculiarities of IP in its relation to rhythm from the point of view of speech analysis-synthesis process.

Considering speech generating process it is usually noted that rhythm as one of IP's forming factors is functioning under constant sense control /or corrected intellectually/. This idea is often accompanied with that of considering IP as some means of packaging the lexical content of an utterance into some rhythmically organized linguistic form. This is hardly true. "A cognitive draft" /a well turned expression of L.P.Jakubinsky/ becomes the verbal text on account of the simultaneous acting of semantic, pragmatic and phonorhythmic factors. Thanks to this the generated text meets both the needs of adequate transmission of information and needs of comfortable pronouncing and perception.

From the point of view of speech perception some ideas of contemporary cognitive psychology are very interesting. In particular it advances an assumption that solution of different intellectual tasks, speech recognition among them, is performed by a listener on the continuously acting rhythmical background which regulates his attention, memory and other cognitive devices. It is possible that in the case of speech perception rather quick and adequate detection of IP's units is provided not only by listener's knowledge of the general rhythmical principle of speech patterning but also due to his own rhythmical activity getting him into the state of pre-expectation or preparedness for detection of certain fragments in rhythmically organized text he hears. In this case the effort which the listener expends on sense decoding depends on the extent to which the speaker coordinates information structure of the message with its intonational phrasing. Distinctness of IP's phonetical realization and correlation of the speaker-listener rhythmical activity are also important.

The rhythm-forming character of IP permits us to speak of its role in the process of speech decoding in the following way. On the one hand, in view of IP's correlation with syntactic structure of an utterance we can consider the current detection of rhythmo-intonational units to be the first step of sense decoding proper. On the other hand, IP is set by the speaker and we can see that by generating it the speaker is governing the procedure of listener's speech perception. The prosodic means with the help of

which IP is realized can be regarded then as special devices whose function is to orientate the listener in sense space of the text developing in time. Everything that has been told above let us consider IP to be a direct result of the cooperative principle in speech communication.

EXPERIMENT

Design

The aforesaid considerations must be undoubtedly supported by some experimental ground. To our regret there are practically no concrete data about the ways of sense decoding and we do not know how IP is used in this process. It seems reasonable that in such situation first of all it is necessary to reveal the most striking features of sense processing of a text which can have two different forms of presentation: I/ with no formal segmentation marks, 2/ naturally pronounced. We have conducted the pilot study in this direction and we'll outline the results below.

For the experiment we have chosen a sample of a scientific text in Russian. The length of the text was about 300 words.

Procedure

The study was divided into two stages. On the first stage the subjects /I2 persons/ were given the text in the written form, printed in one line without word gaps, punctuation marks and capitals. We'll call it further "written text" /WT/. On the second stage /about a month later/ the same subjects were presented the same text but naturally pronounced. We'll call it further "naturally taken text" /NTT/.

On both stages the subjects must have carried out the same task: to rewrite the analysing text without word alteration in the most convenient and usual for them form. The subjects were informed beforehand that the text was rather lengthy and they have to copy it piece-by-piece.

Some restrictions on the procedure of selecting the fragments were fixed:

I/ In selecting a fragment next in turn the subjects were allowed to read or to listen the text only once starting from the point which was defined by the endpoint of the preceding fragment.

2/ The subjects were not permitted to read the text aloud /or to repronounce it/. It is of interest to note that, as the subjects admitted themselves later, they nevertheless articulated the WT silently.

3/ The choice of a rewritten fragment was claimed to be done under sense control, that is a chosen fragment is supposed to stand out as some sense unit including a word or a succession of words. Besides these there were no strict restrictions regarding the fragment length and the concrete character of sense unity. In accordance with the general principle of sense control the subjects were free to choose any fragment best suitable for them to carry out the task set.

The goal of the experiment reported here was to compare the subjects' behaviour in analysing the WT and that of NTT. For this purpose some formal characteristics have been fixed for every fragment chosen by each subject: a) time of selection /in sec/, b) length /in words/, c) the presence of the so-called "looking ahead" or a situation when the chosen fragment was smaller than the text piece that had been read or listened for its selection, d) necessity of correcting the place of the preceding fragment's endpoint or a situation when a subject selecting a fragment coming in turn discovers that he was wrong in the choice of the endpoint of the preceding fragment, e) the presence of word-alterations in the fragment written down /errors in word identification, omissions, substitutions/.

In addition to the above mentioned characteristics we have also fixed some syntactic properties of the fragments written down by every subject. These properties can be called indicators of the fragments' coherence, completeness and autonomy.

The features in question can be analysed and formally determined only in case of fixed representation of syntactic structure of an utterance. The method of representation applied in our study is known in syntax /3/ as the linearized syntactic dependency tree /LSDT/. This kind of representation takes account of both syntactic relations between words and the word order.

Briefly, these features are as follows:

- coherence. A fragment is considered to be coherent if all the words within it are linked, directly or indirectly, and only for one word the head is out of the given fragment. Otherwise the analysing fragment is considered to be incoherent. Formally coherence means that the only one external arrow of LSDT can enter the fragment.
- Completeness. A fragment is considered to be complete if all the words within it have no subordinates out of the fragment. A fragment is considered to be partly incomplete if the only word in it which has subordinates out of the given fragment is the main predicate of an utterance. And finally, a fragment is considered to be incomplete if there is at least one word being not the main predicate of the utterance has a subordinate out of the given fragment. Formally completeness means that there is no external arrow coming out of the fragment.
- Autonomy. A fragment is considered to be autonomic if it corresponds to a simple sentence or, in other words, if it is equal to the entire argument-predicate structure. Otherwise the analysing fragment is considered to be non-autonomic. Formally autonomy means that the only external arrow which enters the given fragment is that to the symbol S of LSDT. The latter occurs only in cases of syntactically compound utterances.

For the experimental goal it was interesting to study the behaviour of different subjects writing down the text presented in the same form /WT or NTT/ and behaviour of the same subject writing down the text presented in different forms /WT and NTT/. To estimate the homogeneous character of subjects' behaviour we have calculated the relative number of coinciding fragments for each pair of experimental text copies. It seemed also useful to value the stability of segmentation locations /i.e. boundaries of selected fragments/. Numerically it can be expressed as the relative amount of the subjects which have distinguished the given boundary. The frequency distributions of boundaries with different stability can serve as a measure of behavioral homogeneity of the whole group of subjects having analysed the same text.

Results and analysis

The conducted experiment has exposed many interesting peculiarities in the subjects' behaviour. Here we have to discuss them in a brief and conspective form.

First of all it is necessary to note that the rewriting of the WT /text with no formal segmentation marks/ was admitted by all the subjects as a rather complicated task which required a great intellectual and psychological effort. According to the strategy of text rewriting all the subjects can be subdivided into two groups. The main distinctive feature of this division is the preferable length and syntactic properties of the fragments having been singled out. It seems that the basis of this distinction is the difference in the volume of subjects' short-term memory. It should be noted that all the subjects were approximately of the same age /25-35 years old/ and the same social rank /collaborators of the Moscow University philological department/. From the whole group /I2 persons/ the majority of subjects /7/ prefered to select rather short fragments, 3 persons, on the contrary, singled out rather long stretches, and the behaviour of 2 persons was inconsistent. The results of the latter group were not considered later.

We'll denote further the first group of subjects as I-G and the second one as II-G.

The ordered set of fragments, i.e. the copied text written down by every subject, was analysed from the point of view of the formal characteristics described above. Table I presents the results of this analysis.

Table I shows that the change of the text presentation form from WT to NTT provokes the following modifications of SF characteristics:

I. The length of SF becomes at an average one and a half longer, what causes decreasing the total number of SF in subjects' NTT copies on 3I,4%.

2. The time expended on fragment selection shortens twice.

3. The number of fragments singled out with the "looking ahead" strategy substantially diminishes; it is seen that for the subjects having used this strategy in writing down the WT the number of SF /with "looking ahead"/ makes 45% of the total. In writing down the NTT the same subjects used this strategy only in 2-5% of all the cases.

4. The need of current correction of SF boundaries drops utterly.

5. The number of SF with word alteration is rather small in both cases /6-7% for I-G and I6-23% for II-G, the greater value for the latter clearly correlates with more lengthy SF/.

6. As for the syntactic indicators the autonomy modificates most of all: in NTT copies the number of autonomic SF increases essentially /at an average I5-I7%, for some subjects up to 50%/. The coherence and completeness indicators change insignificantly.

Table I. Characteristics of selected fragments /SF/. Mean values obtained for all the subjects of the same group. Values in % are given in relation to the total number of SF.

	WT		NTT	
	I-G	II-G	I-G	II-G
Length of SF	3,I	5,2	4,6	7,7
Time of selection	9,2	I2,5	4,3	6,4
The total SF number	86	5I	59	35
Number of F selected with "looking ahead"	8,9%	44,8%	6,5%	2,I%
Number of SF with correction	9,0%	6,I%	-	-
Number of SF with word alteration	6,4%	I6,0%	7,I%	22,7%
Number of incoherent SF	8,2%	9,7%	8,5%	2,6%
Number of complete SF	44,5%	5I,6%	49,4%	56,7%
Number of partly incomplete SF	I4,2%	8,8%	I2,9%	I,9%
Number of incomplete SF	4I,3%	39,3%	37,6%	40,8%
Number of autonomic SF	7,0%	30,4%	22,3%	46,9%

The character of the syntactic indicators' modifications and also that of the SF length allows us to conclude that the diversity in the same subject's behaviour as to WT and NTT cases is not based on differences in text comprehension by the speaker and by the subject himself. Most likely, the truth is that the speaker in his producing the text sets the higher level of syntactic structure of an utterance, i.e. the level being nearer to the entire sense frame of a message.

Let's now discuss the homogeneitty of subjects' behaviour. As an illustration, the data concerning the relative number of coinciding fragments in text copies for some subjects are presented in table 2.

Table 2. The number of coinciding fragments in text copies of different subjects /in % of the total number of F in each pair of copies/

SUBJ.			WT			
			I-G			II-G
			I	2	3	I
NTT	I-G	I	38,6	56,8	54,7	26,3
		2	63,9	28,0	55,7	24,6
		3	70,I	64,8	29,0	26,I
	II-G	I	68,3	6I,3	68,5	67,9

In table 2 the diagonal corresponds to the same subject in case of WT and NTT. Above the diagonal there are results for the different subjects in WT case, under it - the results in NTT case.

Table 2 shows that in NTT case tne homogeneity of selecting the fragments increases significantly. It is expressed by the difference obliteration between two groups of subjects and especially between various subjects in the same group. Thus, for example, the analysis reveals that in I-G the relative number of coinciding fragments is about 55% for a pair of subjects in WT case, but about 72% in NTT case. At the same time in different text copies /WT and NTT/ written down by the same subject there are only 37,5% of coinciding fragments.

As to the stability of segmentation locations the data concerning this problem are represented on fig. I.

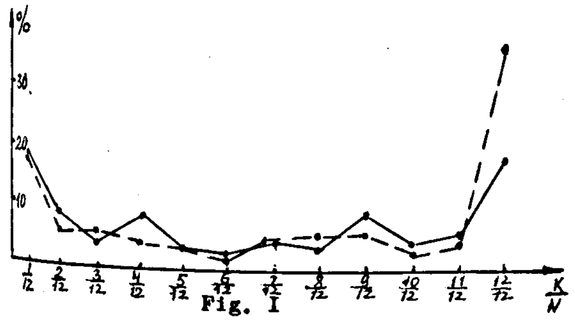


Fig. I

Frequency /in % / distribution of boundaries with different stability. Continuous and dotted lines correspond to WT and NTT cases accordingly. $\frac{K}{N}$ - the relative number of subjects having distinguished the given boundary.

Our study reveals that in NTT case the total number of boundaries distinguished by the whole group of subjects decreases greatly in comparison with WT case /from I34 to 85/. At the same time there is the increasing of the amount of boundaries distinguished by all the subjects, i.e. the most stable ones /from I8% to 35%/.

CONCLUSION

In this study we proceeded from the following assumptions:

I. The process of piece-by-piece rewriting of a text realizing under sense control reflects to some extent the real process of sense decoding.

2. A "written" text with no segmentation marks is equivalent to its-pronounced copy from the point of view of sound information included and the only difference between the copies is the absence of prosodic cues in WT.

If these assumptions are reasonable to some degree, then the obtained results permit us to make the following conclusion: IP, being the main determining factor of prosodic characteristics of the analysed text, controls the strategy of sense processing, unifies it and makes it more optimal from its aim point - decoding of the entire sense frame of a message. It is scarcely probable that the diversity of subjects' behaviour in WT and NTT cases is explained only by different perception modes and in no connection with IP cues.

It stands to reason that further research is necessary. In particular, it is interesting to compare our results with the analysis of subjects' behaviour in case of both a normally organized written text and a text pronounced monotonically word-by-word.

REFERENCES

/I/ Grice H.P. Logic and conversation. - In: Syntax and Semantics, v.3. N.Y. Ac.Press, I975, p. 4I-88.

/2/ Якубинский Л.П. Избранные работы. Язык и его функционирование. М., I986.

/3/ Падучева Е.В. О способах представления синтаксической структуры предложения. - ВЯ, I964, № 2.

THE INTONOLOGY OF THE 80-ES

TATJANA NIKOLAYEVA

Institute of Slavistics and Balkanistics;
Academy of Sciences of the U.S.S.R.
Moscow,USSR, 125040

The report sets out a brief review of main trends in intonology of the latest decade, reveals contradictions in methods of one and same field of science, and a great need in common metalanguage. Two types of sentence accent are put forward: neutral sentence stress which is obligatory and has communicative delimitative function and accents with special communicative and prosodical characteristics.

The idea of a prosodic dictionary is put forward.The problems of colloquial speech intonation and the intonation of a text as a whole are raised.

1. We would like to summarize some new features in the intonology of the late decade and to find something in common in the seemingly uncoordinated innovations (and to try also to determine the problems of future investigations.

2. The melody is the main parameter for intonologists as before. Yet the approach to the description of the melodic contour has changed. The contour was described before as a set of levels, configurations of levels or a global line corresponding to a certain sentence type or as a binomial phenomenon with only one relevant part (intonation centre).

The distinction of two frequency movements in the melodic contour is new. These are the " main, or the baseline" and something like a wavy like pattern superimposed on the base line,or canvas. They are called the "baseline" and the "peakline" /1/, the " nesyscij ton"/2/, the main line and the line of frequency peaks /3/,etc The new base notion of "declination" is connected with these distinctions.It turns out that the declination notion is perceptually conditioned.Thus of the two peaks that have an equal range, the second one is perceived as intoned higher/4/,because the hearer is expecting the descent.

3.But the problem of the semantic relevance in this double contour is still open and is still discussed.For some investigators the "accent focus zones" are of importance/5/,for others it is the terminal zone /6/.Some investigators want to see the whole configuration /7/, others - some minimal units. On the acknowledgement of these different conceptions depends in fact the intonology as science. For example,I.Fónagy thinks that there is a limited set of " cliches mélodiques"/8/. On the contrary, D.Bolinger considers the number of melodic contours to be an open set, because contours are combined by minimal unites - "profiles" in practically enumerable ways / 9/.

I would have like to draw your attention to N.D.Svetozarova´s monography /10/ , which to a great extent shows the Soviet intonology of the present day. The Russian intonation is described in this book as a given list of contours.But in the language these contours are given differently and this fact is very essential. They include both " loose" declarative sentences and high-clichés questions.Experimental data proved that the contour as such is important for the perception /11/.Thus the degree of the grammaticalization of contours is different and is functionally conditioned.All this doesn´t annul the value of the intonation centre, i.e. the part of the contour which is usually singled out by acoustic parameters.

4.All said above refers to the plan of expression of sentence intonation.The problem of the so-called "sentence stress" has become a debating question in the intonology of the 80-ies.This problem is divided into a number of sub-problems,such as: "contrastive" versus "normal" stress, nuclear sentence stress","accent of power" and "accent of interest"," default accent",etc.Let us analyse them in a more detailed way. Only two decades ago there prevailed only one leading theory. This theory was so unshakable that its validity seemed to be undoubtful.Namely,

1/ In any sentence is one main stressed element.
2/This element is usually located at the end of the sentence.
3/ It can be located in some other place. It is then an index of some special prominence,underlining.
4/ The underlined elements are loud.
5/ Loudness is emphasis.
6/ The underlined element usually means contrast.
7/ Rheme is usually located at the sentence end.
8/ That´s why Rheme is connected with the sentence stress.
9/ If the stress is removed,the Rheme is removed too.
10/ Where something is loud,it is Rheme.
11/ Rheme is connected with something new, indefinite and important.
12/ The important and indefinite is always accented and the old and Thematic is unaccented /12/

The first blow was delivered to the idea of indispensable contrastivity of prominence. It turned out that there were sentences which had no contrastive elements bur only an accent of prominence.Here are some Russian examples: Mne očen´ponravilas´Vaša žena (not - ne-očen); Tiše, babuška spit (not ne-babuška);Nado poiti v izdatel´stvo - Ja i chodil; Čital li Pet´a roman Aitmatova? - Net, on ne čitajet setjoznych knig, etc.It is quite evident, that the stressed element in these utterances is not new, important or indefinite.

We would now like to argue the main thesis of this theory which says that the sentence has only one stressed element, or only one stress focus.We think that there are two principially different phenomena. First, we have sentence stress,which has delimitative function,it makes a sentence complete and thus separates sentences one from the other. This stress is usually located at the sentence end and is not perceptually audible as a loud one.For example, Eto novyj dom → Eto novyj dom brata → Eto novyj dom brata muža ,etc.These sentence stresses are organized in a set of communicatives types: declaratives, questions,imperatives and so on. Many utterances have only this type of stress.In Russian,for example,these utterances are usually descriptive and correlated with the event as such /13/.

The second type of stress is parallel and independent of the first.We called it accent of prominence. The element under accent of prominence is to be prosodically emphasised ,so it is audible,i.e. perceptually active. For example,Eto moja žena and Eto moja žena.In the last case we all hear the prominence of eto.Thus we are of the opinion that accent of prominence and sentence stress are parallel prosodic phenomena. How can we proved it? We have to evidently adopt some postulates We consider these phenomena as different ones:1/ if they are perceptually distinguished,2/if they can coexist (i.e. are not in complementary distribution),3/if we can formulate its communicative specificity.We can conclude then that the first two conditions correlate with the expression plan, and the last one with the content plan.

Indeed, these phenomena are perceptually distinguishable.The sentence stress is unmarked,the accent of prominence is marked.Segodn´a cholodno has no prominence;Segodn´a cholodno - segodn´a is emphasised. Moreover,they can coexist: the accent of prominence does not liquidate the sentence stress (see,for example, the fundamental study of T.M.Nadeina- /14/). However,they can overlap,if the end is underlined: Ja prošu pal´to — Ja prošu pal´to. At last the third condition-they must have communicative differences. We can propose two specific features of prominence accent for discussion:
1/It is always connected with "extranormal" situations and qualifications. For example, On smog eto sdelat´(i.e. it was very difficult); Groza načalas´(the thunder was unexpected),etc.;2/ It always creates a certain supplementary semantic aura (certain presuppositions) around the sentence. For example, Tol´ko on ne vernuls´a -- in means that 1/ There were more than one;2/ Others returned. Now compare, Otec skoro pon´al, v čem delo -- no supplementary semantic aura.
Thus we have:

	Sentence stress	Accent of prominence
1.Is obligatory?	+	-
2.Is perceptually perceived?	-	+
3.Connected with norm?	+	-
4.Does create the communicative supplemantary aura?	-	+

Both stresses can coexist.

Above accent of prominence was analysed as if it were only one type of phenomenon. In reality, there exist several types of it with different functional status.Here is a list of some of them and the reader is free to decide if the distinctive features mentioned above are really present.
1/ The contrast is the simplest kind of prominence.Ja ne teb´a zvala.
2/ The accent of appraisal.Mne očen´ponravils´a film;Ty absol´utno neprav, etc. Sometimes only the fact of appraisal accent helps us to understand ,if the described subject is big or small, frequent or rare. On jezdit tuda každuju nedel´u (i.e. very often); Ja zaplatila p´atnadcat´ rublej za eti perčatki (i.e. they cost too much).And now we can see that there are also extra-normal situations and supplementary communicative aura.

3/ The accent of result. I resul´taty jest´. ;Jezdil, jezdil ,a uvidel krasavicu; Ja podumal - i rešil zadaču. In such cases the accent is usually located in the verb.

4/The accent of negation.Dajte mne tort bez krema; Svobodnych mest net!; Ne bylo ni vežlivosti,ni uvaženija,etc.Here we also can see the deviation from pressuposed normal situation.

5/The accent of disagreement.Počemu on ne izbavitsa ot svoich knig?- On čitaet ich; Vy očen rezko jemu otvetili.- Mne nravit s a byt rezkoj.

We suppose that the widely discussed "default accent" is also the accent of dis agreement.Čital li Pet´a "Burannyj polustanok"?- Net, on ne l´ubit serjoznych knig. We have a disagreement with presupposed Pet´a´s image and at the same time we expressed our opinion on high level of Aitmatov´s novel.

Let us see the "default accent" examples from papers of R.Ladd,A.Fuchs,A.Cutler et al./15/:

A.Has John read Slaughterhouse -Five?
B. No,John doesn´t read books.

The disagreement here is doubtless. Thus we demonstrate quite another intonological credo:we ought to search for the "de-accenting" reasons not in deaccenting of words, but in the semantics of a sentence as a whole (that´s what to my mind S.Schmerling thinks - /16/).So,"default accent" has always deep disagreement (what I call polemics) because it contradicts a global situation and does not contradict just one single notion within the sentence.In this case the verb is accented not because of de-accenting of terminal words, but because the verb has the property to be a situational centre. It is typologically remarkable that the English examples of "default accent" discussed in literature may be adequately translated into Russian though intonational systems of both languages are very different /17/.In the case of "accent of polemics" the semantics of identification may be.For example, A.U men´a bolit zub, ničego ne pomogaet.- B.Tak u nas chorošaja poliklinika naprotiv.-A.Ja i chodil tuda. Here we have a case of disagreement (with an idea of not going to see a dentist) and identification (the speaker was already in this clinic). I would to like to say a few words about the function of the Russian particle i(and,even,too). It is ,so to speak,the definite article for verbs in Russian: Vam nužno pozanimat´sa japonskim.- Ja i uču jego,etc. To sum it up,we can say: Sentence stress and accents of prominence are parallel and co-existing.That´s why in the sentence John doesn´t read books we have both accents: the first on the read (accent of prominence) and the second - on the books (sentence stress).

That is why, in our opinion,there is no "de-accenting" at all.

6/We called the sixth type of prominence "an extraordinary introduction to the situation".Here we have another typologically remarkable coincidence of utterances in intonationally different languages. Cf.:Tiše! Papa spit!;Paul ruft!; Truman died!,etc.We think that there an inversion in the normal situation expressed.Yet this type of prominence is close to the disagreement accent (See the paper of Chr.Bonnot - I.Fougeron:Podéžjaem k Krymu - sneg ležit;Chotela uchodit´ - telefon zazvonil /18/).The prominence of sneg,telefon does not mean an opposition to something (not-sneg,not-telefon), but here situation as such is contriversial to the normal situation, the one that we expect.

Thus, the typological feature of intonology in the 80-es is the interest to non-standard (non-trivial) prominences. Connected with this are some problems of expression form.Namely,

1/ How many prominences can there be in one sentence? What do they depend on?

2/What namely is accented in the prominence cases? What is the extent of accented part? Does it depend on the language or only the sense?

Some intonologists suppose that the stressed syllable of the main word is only accented /19/.A more wide-spread idea is that the word is accented.There is a new hypothesis proposed by S.V.Kodzasov /20/ ,that the whole constituent may be accented.This constituent may consist of one word or of a few words which are accented altogether.

It is to propose that in "default accent" cases the whole verb-group is accented.

I would now like to ask English speaking linguists.How should one distinguish terminilogically and conceptually two main notions: "accent" and "stress" (in a sentence)? I have tried to compare all the definitions in intonation papers in English.Sometimes they are quasi-synonymous and are used in one context. Cf.: R.Ladd:"Contrastive stress" also corresponds to the default accent...

Sometimes the "stress" is considered as a word property, and "accent" as a sentence property(A.Cutler).(In Russian we have a slightly different intonational terminology).It also becomes clear that "stress" and "accent" are distinct by acoustic properties.It is naive to think so to-day.

I invite you to discuss my suspicion that these two terms reflect two different oppositions (which are not contradictory):

1/"Stress" is more abstract, metalinguistic notion, and "accent" is more concrete."Stress" is more "morphonological", so to speak.

2/"Stress" designates the neutral sentence stress,which is obligatory for each

utterance; as to "accent" it designates underlining.However our colleagues do not overcome the psychology of one prominent zone routine.E.Selkirk ,to my mind,is more close to this double definition /21/. I would like to hear your opinion,if I am right.

4.So the debates around prominence is one of the most important features in intonology of the 80-es.With this is connected another feature: the absence of common metalanguage for the description of intonational systems. This fact explains why the intonology to-day gives so little to syntax and comparative intonology (accentology).

R.Ladd has compared types of intonation description with the description of an elephant by the blind people who can feel only one part of the elephant /22/.If it were so, we could simply summarize intonational descriptions.We have metalinguistic differences and theoretical differences which generate differences in the description, and not vice versa. How can we compare "intonational constructions" / IK/ introduced by E.A.Bryzgunova, by-contour description by N.D.Svetozarova, "metrical grids" by Liberman-Prince (and E.Selkirk), low-high peak combinations by J.Pierrehumbert,"cliches melodiques" by I.Fonagy and D.Bolinger´s "profiles"? To my mind, only a long co-laboration and discussion of this problem will permit to elaborate a common metalanguage and a common theory.

I would like to pay attention to the distinction of acoustic parameters in this sense. All emotional and conceptual divergencies are connected mainly with melodic (and less - with dynamic) parameter of intonation complex.Meanwhile we have rich data of temporal characteristics (Lehiste, Nooteboom,Krivnova et al.).The complement each other in a very interesting way and do not call for discussion.That is why we do not analyse the temporal aspect otherwise our paper from the problemsearching will be reduced a review.But this difference of attitudes to different acoustic parameters has given me an idea of the existence of one more opposition in intonology.We have actually two types of orientation: essence and description.My opinion may be opposed that a successful description is always inseparable from the essence of phenomena. Yet we should draw one´s attention to the tendency of linguistics to-day : to distinct the language facts and the facts of linguistics. Simple language phenomena may be complicated from the point of view of their linguistic status.Temporal characteristics seem to be closer to language, as melody descriptions - they are the domain of linguistics. Consequently there is much in common while describing duration, timing ,ets, while there´s more divergency in melody descriptions.

But I want to say that in our linguistics little attention is paid to the formal possibilities of intonation description and phrase stresses gradualness.That is why "metrical grid" investigation ought to be popularized.

6.J.Pierrehumbert´s works constitute,in my opinion, a significant phenomenon in the latest decade intonology /23/.It gives one also an idea as to how to study the transition from lexical tonal accents of archaic prosody to its amalgamation into the present specific "grammaticalized" melodic contours.Thus we suppose that supersegmental sentence phonetics has its autonomous history (evolution) as segmental phonetics has.And thus melodic contours grammaticalization is the fact of language evolution (See my section paper on the Congress).Going back to I.Fonagy´s ideas of "cliches melodiques", I would like to emphasize that there are two kinds of intonation cliches:

1/Grammaticalized intonation contours like in alternative questions, repeated questions,echo-questions,replikas and so on.

2/Intonation "idioms".They occur in Russian when the whole sentence meaning do not correspond to its lexeme sum, but may even contradict to it.In that case the true sense of utterance can be recognised only through the special "idiomatic"intonation. For example,Nužna mne Vaša kniga! (i.e. I don´t need your book);Tak ja Vam i otvetila! (i.e. I will not answer),etc.If we replace "normal" intonation for this specific one, we will have an absolutely different sentence meaning, or the sentence will be senseless. Nu,ljudi! - sounds as an appraisal.Nu,ljudi. has no sense. Raboty ! (i.e. there is much work).Raboty... has no sense.These observations were obtained from Russian data.It is interesting to know of other languages facts.

6.We have now many affirmations in intonological papers that intonation do not coincide with syntax and may even contradict it.(D.Bolinger speaks in such situations about "grammar"/24/). Now,in the 80-es the question arises: what syntax really is? It is not so far clear from intonological works.Moreover,I sometimes have an impression that syntax for intonologists is more archaic notion thah for syntaxicists themselves. There are some remarkable and important even to-day ideas expressed our academician L.V.Shcherba. Shcherba´s syntagm is not only a prosodic unit,but also a sense unit.That is why the utterance can be divided into syntagms only in accordance with the speaker´s intention.Owing to this division we can understand the communicative intentions of the speaker (Kaznit´ nel´za/pomilovat´ -- Kaznit´ / nel´za pomilovat´).These rules of division into syntagms are equal for any language speaker. In fact,it is not necessary to be a "mind reader" in order to understand a sentence-division correctly. It is quite enough to know the language. Moreover,Shcherba´s ideas are also significant because they bring to the conception of the "intonation syntax" (not only "syntax intonation").In other words the arrangement of syntagmas ,the length of the pauses, the melodic types of syntagms,etc are meaningful themselves/25/.The punctuation is to be to a certain extent parallel to it, however punctuation systems are so different in different countries not for prosodic reasons /26/.In Russian syntax we have many constructions without conjunctions.It is only a "intonation syntax" that determines the sense.For example,Les rub at,shchepki let at (the enumeration); Les rub at - shchepki let at (the condition); Les rub at: shchepki let at(the explanation).In all such cases correlations of the acoustic parameters are regular and well described. I suppose that this approach to "intonation syntax" can be very interesting for intonologists even to day.That is why,in my opinion,it is senseless to say that intonation is independent of syntax, or to say that syntax is independent of intonation. They cannot be separated being both products of communication activity with their own sense units.They may "work" in different directions but they can disperse too.It is out of date now to treat syntax as IC-structure or as a set of sentence types.

7.One of the interesting features of in intonology to-day is the conversion to the intonation of colloquial speech. At present we have enough data /27/.The remarkable feature is a striking similarity of colloquial speech intonation in different languages (Russian and English,for example).The investigators have proved that colloquial speech intonation has a specific organization such as short syntagms, indistinctness,"slurring" of melodic contour structure, an abundance of prominent words,a frequent prominence of the beginning (even if it contain auxiliary words) absence of regular meaningful pause duration we have discussed above.The common colloquial syntax data in comparison with literary syntax codified : an iconic word order, ill-formed structures,etc. draw us to the intersting hypothesis put forward by T.Givón and his colleagues /28/ about the existence of the primary "pragmatic code" and its later "syntactization". In this relation colloquial speech data are much closer to this primary code" than literary data.The proximity of colloquial speech intonation models is understandable; according to Givon, at the "pragmatic code" stages languages are closer to each other. But, in a paradoxical way, it is the colloquial speech and its intonation that represents a storehouse of relicts and an innovation arena at the same time.

8.The new feature in intonology is the introduction of lexics into the intonological analysis. At present two main sets of facts are outlined.First of all, it turns out that certain lexems (or lexeme classes) are inclined to attract the accent of prominence.It concerns some types of adjectives /29/, intensifying adverbs like očen´, absolutno,nevožmozno,etc.There is another class of lexeme that are inclined to be prosodically weak,not-prominent,their segmental structure according to these prosodic reasons may be modified.

In the second place words may have distinct prosodic coloring.The distinction is connected with words pragmatic connotations.For example, words with semantics of "big" have a different intonation than words with semantics of "small"/30/.

The appraisal component of word is important too: words with the so-called "bad" semantics have a different pronunciation from those with "good" semantics. Thus there arises a new idea: to compose a dictionary of prosodical potentiality of lexemes (their prominence and phonation possibility).It is very important to know the degree of universality in such dictionaries.

9.The last point of my paper is devoted to the text intonation (of course, the list of problems is still open).We know now that the connected and integral text has its own intonation structure /31/. This intonological sphere has much in common with phonosylistics and communication units syntax (see I.G.Torsuyeva paper in this Symposium).

References

/1/DGibbon, "Intonation as an adaptive process",Intonation,Accent and Rhythm, B.-N.Y.,1984,p.172.

/2/O.F.Krivnova."Sostavljajuščaja nesuščego tona v strukture melodičeskoj krivoj",Issledovanija po strukturnoj i prikladnoj linguistike,Publikacija OSIPL MGU,vyp.7,M.,1975.

/3/ A.Cutler,"Stress and accent in language production and understanding",Intonation,Accent and Rhythm.

/4/ A.Di Cristo,"De la microprosodie à l´intonosyntaxe",t.2?Aix-en-Provence, 1985,p.537.

/5/ E,Selkirk,"Phonology and syntax",Cambr Mass.-L.,1984.

/6/ D.Bolinger,"Intonations and its parts. Melody in spoken English",Standford, California, 1986,p.25.

/7/ N.D.Svetozarova,"Intonacionnaja sistema russkogo jazyka",L.,1982.

/8/ I,Fonagy,E.Bérard,J.Fonagy,"Clichés melodiques",Folia linguistica,1983, t.17.

/9/ D.Bolinger,Op.cit.

/10/N.D.Svetozarova,Op.cit.

/11/M.G.Radijevskaja,N.D.Svetozarova,"O strukture intonacionnogo´kontura v russkom jazyke",Experimental´no-

fonetičeskij analis reči,L.,1984,p.148.

/12/See in detail: T.M.Nikolayeva,"Sémantika akcentnogo vydelenija",M.,1982.

/13/T.M.Nikolayeva,"Kategorial´no-grammatičeskaja cel´nost vyskazyvanja i jego pragmatičeskij aspekt",Izvestija AN SSSR,Serija literatury i jazyka, 1981,t.40, N 1.

/14/T.M.Nadeina,"Akcentnaja struktura vyskazyvanija v russkom jazyke",Avtoreferat kand.diss.,M.,1986.

/15/R.D.Ladd,"Light and shadow.A study of the Syntax and Semantics of Sentence Accent in English",Contributions to grammatical studies,Leiden,1979; A.Fuchs,"Deaccenting" and "default accent",Intonation,Accent and Rhythm.

/16/S.F.Schmerling,"Aspects of English sentence stress",Austin,L.,1976.

/17/ It is intersting to analyse these accent similarities in detail.

/18/Chr.Bonnot,I.Fougeron,"Accent de phrase non final et relations internonciatives en Russe moderne",Révue des etudes slaves,1983,t.LV, N 4.

/19/D.Bolinger,"Intonation and its parts, p.10.

/20/S.V.Kodzasov,"Tonal´naja emfaza"(see his paper on this Symposium).

/21/E.Selkirk,"Phonology and syntax" (NSR-Nuclear Stress Rule,PAR - Pitch Accent prominence Rule).

/22/R.D.Ladd,"Light and shadow",p.102.

/23/J.Breckenbridge-Pierrehumbert,"The phonology and phonetics of English intonation".D.of Ph.Dissertation,MIT, 1980.

/24/D.Bolinger,Op.cit. and other works.

/25/See:T.M.Nikolayeva,"Semantika akcentnogo vydelenija".

/26/See:I. et J.Fonagy,"L´intonation et l´organisation de discours",Bulletin de la Societé de linguistique de Paris,t.LXXVIII, fasc.1,1983.

/27/ See books and papers of "Razgovornaja rec" group in Institute of Russian language of Academy of Sciences USSR.The recent work is also from LGU group:N.I.Geilman,"Razgovornaja proiznositel´naja norma i var´irovanije fonetičeskich oboloček slov", Voprosy normy i realizacii jazykovych sredstv, Gor´kij,1984.

/28/.T.Givón,"Universals of discourse strucure",Language universals and second language acquisition,Amsterdam-Philadelphia,1984.

/29/T.P.Skorikova, Funkciolal´nyje vozmožnosti intonacionnogo oformlenija slovosočetanija v potoke reči,Avoreferat kand.diss.M.,1982.

/30/S.V.Kodzasov,O.F.Krivnova,"Fonetičeskije vozmožnosti gortani i ich ispolzivanije v russkoj reči",Problemy teoretičeskoj i experimental´noj lingvistiki.Publ.OSIPL MGU,vyp.8,M.,1977

/31/ See papers of special conference:"Prosodija texta",Sbornik naučnych trudov MGPIIJA im.M.Toreza,1982,vyp.169.

PHRASIERUNG: ZENTRUM UND PERIPHERIE DER PROSODISCHEN KONTUR

Christian Sappok, Seminar für Slavistik

Ruhr Universität Bochum
4630 BOCHUM, B.R.D.

ABSTRACT

The latest development in modelling intonational contours tends towards a complicated internal structure, for example the coexistence of different tones within one and the same contour. The question árises wether the peripheral portions such as boundaries or connections between contours show a comparable amount of structural diversification. After the discussion of alternative hypotheses a series of experiments is presented where manipulations of prosodic parameters in the place between adjacent contours are tested in perceptional experiments. The results show a nonlinear dependency between these changes and the reactions, this being interpreted as evidence for the functioning of categorial features in the observed utterance portions.

I. DIE FRAGESTELLUNG

Ein wesentliches Kennzeichen in der jüngsten Entwicklung der Intonationsforschung ist ein zunehmender Komplexitätsgrad innerhalb der Grundeinheit, der Phrasierungseinheit (des Syntagmas in der Terminologie von L.V. ŠČERBA). Die interne Zusammensetzung dieser Einheit wird komplexer, und zwar nicht bezüglich der relevanten Tonstufen oder Intensitätsgrade; es geht vielmehr um die mögliche und grammatisch geregelte Verknüpfbarkeit von Trägern dieser Merkmale zu funktionellen Verbänden. Im Modell von PALMER 1922,7 konnte die Intonationseinheit nur eine hervorgehobene Silbe haben. Dieses Konzept findet seine Fortsetzung bei CRYSTAL und QUIRK, allerdings mit einer Weiterentwicklung in Form möglicher Subordinationsverhältnisse zwischen Nuklei (1964, 15ff.). NIKOLAEVA 1982, 13 et passim geht von der Opposition von Phrasenbetonung und akzentueller Hervorhebung aus; die beiden Mechanismen sind miteinander verträglich und können innerhalb ein und derselben Intonationskontur wohlunterschieden koexistieren. Die Kontur bleibt bei dieser Entwicklung, von der hier nur wenige markante Punkte genannt wurden, in ihrer Rolle als oberstes Gliederungsprinzip des lautsprachlichen Äußerungsprozesses unangetastet; sie ist aber nicht mehr ein unteilbares Ganzes, sondern wird zur Folge unterscheidbarer, funktionell selbständiger Komponenten. Hieraus ergibt sich für die Linguistik die Aufgabe, Regeln für die Kombinatorik der Bestandteile und ihre funktionellen Entsprechungen zu formulieren.

Ein Teilaspekt dieser Aufgabe besteht in der Neubestimmung des Stellenwertes für die nichtnuklearen, peripheren Bestandteile der Kontur. Solange jeder Gipfel oder Nukleus ein monolithischer Block von Ausdrucksmerkmalen und Inhaltsfunktionen ist, kann die Peripherie im Status einer rein mechanischen Begleiterscheinung, eines prosodischen Übergangsgliedes verbleiben, dem gar keine oder höchstens periphere Funktionen wie der Ausdruck emotionaler Färbung oder organisatorischer Besonderheiten wie Delimitation oder Weiterverweisen zukommen.

Läßt das Beschreibungsmodell aber polynukleare Sequenzen mit heterogenen Funktionen und eigener, sprachspezifischer Regularität zu, dann muß auch die Frage nach dem Status peripherer Komponenten neu gestellt werden. Vor allem wird die Frage von Bedeutung sein, ob und inwieweit dieser Bereich als Abbau oder Umkehrung der nuklearen Merkmalskonstellation zu bewerten ist, oder ob es positive, spezifische Regularitäten gibt, die den peripheren Bereich aus der prosodischen und funktionellen Abhängigkeit vom Nukleus bzw. des komplexen nuklearen Breichs herausheben.

II. DIE ALTERNATIVE

Wir formulieren die bis hierher entwickelte Fragestellung in Form einer Alternative um, die es ermöglichen soll, gezielt Beobachtungsdaten zu sammeln und ihre Auswertung zur Beurteilung von Hypothesen heranzuziehen. Grundlage ist ein Modell vom lautsprachlichen Äußerungsprozeß, dessen integraler Bestandteil ein Konzept der Dynamik darstellt, d.h. einer Veränderung, die von einem Zustand x ausgeht, einen Zustand y erreicht und wieder zum Ausgangszustand x zurückkehrt. Im Modell überlagern sich dabei mehrere Dynamiken; die hierarchisch höchste (die A - Prosodie nach TILLMANN und MANSELL 1980) ist dadurch ausgezeichnet, daß sie der Aufmerksamkeit als solche, d.h. als Dynamik, zugänglich ist. So wie die Zuordnung von Silbenstruktur (als einer untergeordneten Dynamik) und Morphemstruktur nicht willkürlich oder trivial ist, sondern von sprachspezifischen Regeln gesteuert wird, unterliegt auch das Verhältnis von A - Prosodie und Gesamtäußerung sprachspezifischen Regularitäten, den Phrasierungsregeln.
Die Herausbildung eines Nukleus innerhalb dieser Dynamik, die Umformung einer untergeordneten Hervorhebung in eine übergeordnete (etwa eines stress in einen accent in der Terminologie von BOLINGER 1965) ist der zentrale Gegenstand jedes Intonationsmodells; hinsichtlich der Herausbildung eines Ausgangs- oder Übergangsbereiches ist folgende alternative Betrachtungsweise möglich.

A. Der Ausgangs- und Endpunkt, d.h. der Zustand x gemäß der oben gewählten Formulierung, ist der virtuelle Nullpunkt im Prädikatenraum der prosodischen Parameter. Im konkreten Äußerungsfall ist er dadurch gekennzeichnet, daß sich die Parameter des Signals auf diesen Punkt zubewegen, sich ihm annähern. Dabei ist es keineswegs nötig, daß sie diesen absoluten Nullwert in Form einer Schweigepause tatsächlich erreichen. Das Wahrnehmungskorrelat für das Ab- und Ansteigen im Falle einer Gliederungszäsur ist gewährleistet, wenn die Bewegung eine bestimmte Schwelle überschreitet. Von linguistischer Relevanz ist die Untersuchung dieser Schwellenwerte, des Zusammenhangs zwischen den Werten der einzelnen Parameter im Verhältnis zu den entsprechenden Werten im Zentrum, wobei nur den letzteren distinktive Funktionen zukommen.

B. Der Ausgangs- und Übergangsbereich befindet sich in unmittelbarer Nachbarschaft des Zentrums, er ist folglich von dessen Merkmalen abhängig, er ist aber darüber hinaus in spezifischer Form gegliedert und geregelt: Er ist geprägt durch den gleichen Prädikatenraum wie das Zentrum, doch gibt es auch hier kategorial geprägte Grenzen, geregelte Verhältnisse zwischen den Werten der beteiligten Parameter, die folglich den Status von Merkmalen besitzen. Die Peripherie ist gekennzeichnet durch ein Merkmalbündel, d.h. bestimmte Merkmalskombinationen sind in einer bestimmten Sprache als zulässig ausgezeichnet, während andere unzulässig, markiert oder unnatürlich sind.

Es folgen einige Beobachtungen, die die Entscheidbarkeit dieser Alternative betreffen und eine endgültige Konstruktion von Hypothesen zum Status der Grenzbereiche im Intonationsmodell ermöglichen sollen.

III. DIE DATEN

Zur Gewinnung von Daten für die Beurteilung der Ausgangsalternative hinsichtlich der prosodischen Merkmale im Grenzbereich dient uns die folgende empirische Grundlage. Es wurde eine Reihe von Sätzen konstruiert, deren lautsprachliche Realisation zwei Versionen zuläßt, und zwar eine mit einer internen Phrasierungsgrenze, die den Satz in zwei Phrasierungseinheiten zerteilt, und eine ohne diese Grenze. Die prosodischen Parameter dieser Paare wurden gezielten Veränderungen unterworfen und anschließend einer auditiven Beurteilung unterworfen[1]. Die Aufgabe bestand darin, die mehrfach in Zufallsreihenfolge gebotenen Stimuli einer von zwei schriftlichen Versionen zuzuordnen, deren eine durch Interpunktionszeichen die Position der virtuellen Phrasierungsgrenze signalisiert. Um die generellen Möglichkeiten des Vorhandenseins von Übergangsfällen und den kategorialen Status von Opposition zu testen, wurden im Beispielsatz (1) alle prosodischen Parameter einer Manipulation
(1) Uvidel Tonju (,) na vokzale.
Er traf Tonja(,) auf dem Bahnhof.
unterworfen und ein Kontinuum von acht Übergangsstufen generiert. Die Ergebnisse zeigen das für die kategoriale Wahrnehmung charakteristische Interpretationsprofil, vgl. 1.1 als die der Kommaversion am nächsten stehende Interpolationsstufe, 1.8. die dem ungegliederten Original am nächsten stehende Stufe und den Sprung zwischen dem 6. und 7. Stimulus.

[1] Die digitale Bearbeitung und Manipulation der Stimuli erfolgte am Seminar für Phonetik der Universität München, Vorstand H.-G. Tillmann. Die auditiven Tests erfolgten an der Kafedra fonetiki der LGU Leningrad, Vorstand L.V. Bondarko sowie an der Pracownia fonetyki der PAN Poznań, Vorstand W. Jassem. Diesen Institutionen, vor allem den Versuchspersonen, möchte ich an dieser Stelle für ihre Mitarbeit danken.

+ gegliedert	Konf.-Interv.	Mittel	Varianz
1.1. 80.0%	3.528-4.499	4.014	2.118
1.2. 78.1%	3.456-3.344	3.905	1.817
1.3. 74.6%	3.281-4.179	3.730	1.814
1.4. 66.8%	2.842-3.833	3.338	2.209
1.5. 65,4%	2.865-3.676	3.270	1.480
1.6. 54.6%	2.289-3.170	2.730	1.744
1.7. 29,2%	1.041-1.932	1.486	1.785
1.8. 24.3%	0.726-1.706	1.216	2.160

Welche Merkmale auf der Inhaltsebene entsprechen dieser Opposition? Und läßt sich die Beeinflussung dieser Interpretation auch durch auf den Grenzbereich beschränkte Manipulationen erreichen? Im Satzpaar (2) bewirkt der Gliederungseinschnitt

(2) Teraz biegiem (,) na dworzec!
Jetzt los (,) zum Bahnhof!

eine Aufspaltung der pragmatischen Funktion in zwei Portionen: Der Vorgang der Auslösung wird redupliziert. Das Interpretationsprofil zeigt die relativ geringfügige Wirkung einer Absenkung des Grundtons im Bereich der Silbe vor der potentiellen Phrasierungsgrenze, vgl. 2.1., mit einer geringen Verstärkung des Effekts bei größerer Senkung. Ein Sprung im Bewertungsprofil wird durch die Kombination von Absenkung und Dehnung erzielt, vgl. 2.4., während der Effekt der Absenkung weiter abgeschwächt wird, wenn die nachfolgende Silbe (vgl.2.5) miteinbezogen wird und so die Diskontinuität im Grundfrequenzverlauf als Begleiterscheinung der Manipulation ausgeglichen wird.

Art der Modifikation	Bewertung: ohne Zäsur
2.1. Original ohne Zäsur	88.3%
2.2. ...biegi/e/m 160>145 Hz	77.6%
2.3. ...biegi/e/m 160>140 Hz	68.0%
2.4. ...biegi/e/m 160>145 Hz,+ 60 ms	19.3%
2.5. ...biegi/emna/ 160>140 Hz	75.6%

Die bisher dargestellten Manipulationen lassen die Reaktionen der Versuchspersonen als Funktion von Eingriffen erscheinen, die die Peripherie in ihrem Verhältnis zum unmittelbar vorausgehenden Nukleus betreffen. Im folgenden russischen Beispiel (3) mit der pragmatischen Funktion der Aufforderung bleibt die Dehnung und

(3) Nu nu (!) Nu nu nu!

die Frequenzabsenkung im Grenzbereich auch bei guter Erkennbarkeit relativ schwach wirksam, vgl.3.1 bis 3.4, während die Verlegung der End- (vgl.3.5) bzw. der Anfangssilbe (vgl.3.6) in die Umgebung des virtuellen Einschnitts einen starken Gliederungseffekt nach sich zieht.

Art der Modifikation	Bewertung: Ohne Zäsur
3.1. Original ohne Zäsur	91.5%
3.2. nun/u/ ... 235 > 210 Hz	56.9%
3.3. nun/u/ ... + 60 ms	81.2%
3.4. nun/u/ ... + 80 ms	63.3%
3.5. nu/nu/ = Endsilbe	5.6%
3.6. nunu/nu/ = Anfangssilbe	22.5%

Die manipulative Veränderung der prosodischen Parameter im potentiellen Grenzbereich beeinflußt das Interpretationsprofil auf graduell abgestufte Weise oder in Form deutlich markierter Sprünge. Dabei bleibt unklar, wo der Bezugsrahmen zu suchen ist, von dem aus gesehen prosodische Veränderungen auf die Interpretation Einfluß nehmen. Von der eingangs entwickelten Alternative aus gesehen sind zwei Möglichkeiten denkbar: Entweder ist der vorausgehende und der nachfolgende Nukleus ausschlaggebend; also muß von dessen Tonhöhe, Amplitude und Länge ausgesehen der Grenzbereich eine Abweichung darstellen, und je stärker diese Abweichung ausgeprägt ist, desto eher ist die Interpretation in Richtung auf die intern gegliederte Version hin verschoben. Oder es gibt eine spezifische Konstellation von Parameterwerten, die als Merkmal für die Peripherie interpretiert werden; auch in diesem Fall stellen die Parameter der Peripherie Abweichungen vom Zentrum dar, doch ist der spezifische Weg, wie aus den Parameterwerten Merkmale mit Signalfunktion werden, nicht direkt in Abhängigkeit zu den Nuklei der unmittelbaren Umgebung zu modellieren. Um in dieser Frage einen Schritt weiterzukommen, wurde eine möglichst weitgehend parallele Manipulation in einem Aussagesatz und einem entsprechenden Fragesatz vorgenommen; vgl.(4) und (5).

(4) On živet zdes' (,) na Mojke.
Er wohnt hier (,) an der Mojka.

Um die Manipulationen in Anlehnung an die vorangehenden Stimuli durchführen zu können, wurden aus den Originalversionen solche Fälle ausgesucht, die die Betonung auf živet tragen, während zdes' zur Peripherie gehört. Die Manipulationen beziehen sich zunächst nur auf den Grundfrequenzverlauf, vgl. 4.11 bis 4.4. und entsprechend 5.1.

Art der Modifikation	Bewertung: ohne Zäsur
4.1.Original ohne Zäsur	92.5%
4.2.živ/e/tzd/e/s' 180/170>150/140 Hz	76.4%
4.3.živ/e/tzd/e/s' 180/170>150/120 Hz	59.5%
4.4.živ/e/tzd/e/s' 180/170>150/110 Hz	51.3%
4.5.dito 4.3.,+ 50 ms	8.5%
5.1.Original ohne Zäsur	91.2%
5.2.živ/e/t zd/e/s'200/180>180/160 Hz	76.1%
5.3. - " - 200/180>180/140 Hz	77.6%
5.4. - " - 200/180>180/130-110 Hz	79.4%
5.5. dito 5.2., /zdes'/ + 50 ms	76.8%
5.6. dito 5.3., /zdes'/ + 50 ms	23.5%

Es zeigt sich, daß in Satz (4) die kontinuierliche Absenkung einen kontinuierlichen Einfluß auf die Interpretation hat, während in (5) die Interpretation trotz analogen Veränderungen unbeeinflußt bleibt. Eine begleitende Dehnung führt in (4) einen Sprung in Richtung auf die Gliederungsinterpretation herbei, was in (5), wie 5.5. und 5.6. zeigt, nicht bzw. in abgeschwächter Form der Fall ist. Die unmittelbare Abhängigkeit zwischen den Merkmalen des Zentrums und den davon abweichenden Merkmalen der Peripherie ist in diesem Falle gelockert bzw. aufgehoben.

IV. ERGEBNISSE

Das linguistische Modell vom lautsprachlichen Äußerungsprozeß muß eine Verbindung zwischen der Dynamik prosodischer Prozesse und der diskreten Abfolge grammatischer Formative herstellen. Der Phrasierungsdynamik kommt dabei als oberster Ebene in der Hierarchie prosodischer Strukturierung insofern eine besondere Bedeutung zu, als sie der bewußten Aufmerksamkeit als Dynamik zugänglich ist und daher eine Reihe von kommunikativ zentralen Funktionen wie Fokussierung, Satzmodusfestlegung usw. übernehmen kann. Prosodisch distinktive Merkmale konzentrieren sich dabei auf die zentralen Bereiche der Intonationskontur; die Tatsache, daß das Zentrum seinerseits aus mehreren Nuklei zusammengesetzt sein kann, läßt dem Grenzbereich eine wichtige organisatorische Funktion zukommen: nicht jede Übergangsphase zwischen zwei Nuklei ist als Grenze zu bewerten. Handelt es sich dabei um graduierbare Übergangsformen? Oder gibt es spezifische Merkmale für den Grenzbereich, deren Zusammensetzung von der unmittelbaren Umgebung unabhängig einen funktionell differenzierten Bestand an Grenztypen zuzuweisen ist? Vorläufige Ergebnisse der unter diesem Aspekt auf den Grenzbereich konzentrierten Beobachtungen lassen sich folgendermaßen zusammenfassen.

1. Die Tatsache, in welchem Maße ein Segment zwischen zwei Nuklei oder Nukleuskombinationen als Grenzbereich interpretiert wird, ist abhängig von der syntaktischen und semantischen Ebene der zugrundeliegenden Struktur. Es gibt keine absolut wirksame Abhängigkeit zwischen dem Ausmaß der Veränderungen prosodischer Parameter und der Interpretation der Äußerung als gegliedert bzw. ungegliedert. Dabei kann es zu ambivalenten Beurteilungen kommen, ohne daß die Wohlgegliedertheit der Äußerung als ganze gestört ist.

2. Eine der zentralen Funktionen der Peripherie von Intonationskonturen ist die differenzierte Verknüpfung von pragmatischen Funktionsträgern. Die Merkmale signalisieren den Übergang von einem pragmatischen Typ zum anderen bzw. die Reduplikation ein und desselben Typs.

3. Die Ableitung der peripheren Merkmale aus den umgebenden zentralen Bereichen ist nicht obligatorisch an die unmittelbar umgebenden Nuklei gebunden. Diese Beobachtung berechtigt die Annahme, daß die Strukturierung der peripheren Merkmalbündel einen autonomen, von den zentralen, nuklearen Merkmalen unabhängigen Regelkomplex unterworfen ist.

Literatur

BOLINGER, D.L.M. (1965): Forms of English. Accent, Morpheme, Order. Tokyo.

CRYSTAL D., QUIRK R. (1964): Systems of Prosodic and Paralinguistic Features in English. The Hague.

NIKOLAEVA T.M. (1982): Semantika akcentnogo vydelenija. Moskva.

PALMER H.E. (1922): English Intonation with Systematic Exercises. Cambridge.

ŠČERBA L.V. (1955): Fonetika francuzkogo jazyka. Moskva.

TILLMANN H.-G. , MANSELL Ph. (1980): Phonetik. Lautsprachliche Zeichen, sprachliche Signale und lautsprachlicher Kommunikationsprozeß. Stuttgart.

SENTENCE INTONATION IN LANGUAGE AND LINGUISTICS

N.D.SVETOZAROVA

Department of Phonetics
Leningrad University
Leningrad, USSR 199034

ABSTRACT

In spite of the great interest of modern linguistics in the problems of intonation and a great amount of experimental data, intonation has not yet taken its due place in general linguistics.This can be accounted for by the fact that intonation is often viewed as peripheral to language system, belonging rather to speech than to language. Another reason is an extreme diversity in understanding the nature of intonation units - either as language signs or as a means of the plane of expression. It is assumed that intonation - with the multitude of its functions - belongs to the plane of expression of complex signs which should not be reduced to phonemics alone. The phenomenon of intonation demonstrates the inconsistency of the view on sound matter of language as the lowest level of language structure.

INTRODUCTION

One of the noticeable peculiarities of modern linguistics is increasing interest of linguists of various trends to the problems of sentence intonation. It's proved by the fact that intonation that has been the "cinderella" of linguistics for a long time, nowadays turns out to be its "core".

But in spite of the fact that a lot of experimental data on intonation have been accumulated, their introducing to the theory of linguistics does not seem to be a simple task although many linguists realize the necessity of their generalization and interpretation. Even the best of modern works in general linguistics lack serious discussions of intonation problems.

There are several interrelated reasons for this.

THE NATURE OF INTONATION UNITS AND THE PLACE OF INTONATION IN LINGUISTICS

The first reason is the still-remaining treatment of intonation as something peripheral to the language system and attributing it rather to speech than to language.

The second reason is the fact that it is not clear what aspect of linguistics should deal with intonation. Information on intonation in a particular language (if there is any) is given either in "Phonetics" or in "Syntax" or even in both of them. The latter is illustrated for Russian by the new academic edition of Russian Grammar /1/ (parts devoted to intonation are written by E.A. Bryzgunova).

The fact that different aspects of linguistics deal with the same subject seems to be quite natural as different branches of linguistics do not just study various language phenomena, but analyze them from different points of view.

The word is analyzed both in phonetics and lexicology. Nevertheless, one can hardly imagine that word - the main language unit - should be analyzed only in phonetics, that is, from the point of view of its form (plane of expression) or only in lexicology, that is, from the point of view of its meaning (plane of contents).

As to intonation units, the main debateable point is whether to attribute them only to the plane of expression or to regard them as bilateral units, that is, language signs.

Both viewpoints have their adherents.This is reflected in different definitions of intonation and its units.

The adherents of the view point that intonation units are signs do not give any special arguments to support their position. Probably they consider the ability of intonation to convey definite meanings (thus be connected with meaning) to be quite a weighty reason. However, everything in the language is connected with meaning to a certain extent. The material form of the language does not exist just by itself but as a means of conveying information.

To prove the sign nature of intonation its signs should be compared with other language signs to see whether everything that is defined by "intonation" is equally and in the same way connected with meaning.

On the contrary, those who attribute intonation only to the plane of expression, i.e. to phonetics as the aspect dealing with the sound matter (form) of the language, think it is their duty to give special reasons for their position. Thus they present the well-known fact that the same formal means (intonation pattern) combined with different lexical and grammatical structures conveys different meanings, so that the meaning is not conveyed by intonation pattern alone, but by its combination with other linguistic means.

It seems, though, that special reasons for proving the non-sign nature of intonation units are required only if they are treated like phonemes. In case we attribute intonation to the plane of expression of language but do not confine it to phonemic alone admitting the existence of other sound means that have special functions, then we may regard intonation as an element of the plane of expression capable of conveying some specific meanings (elements of the plane of contents) - communicative, modal and emotional.

According to the conception of the Shcherba Phonological School (or Leningrad Phonologocal School) the main function of the phoneme is the constitutive one /2, 3,4/.

Phonemes, which have no meaning of their own and are singled out due to their potential link with meaning, constitute the sound matter of morphemes - the smallest meaningful language units.

The plane of expression of complex language units, such as words, word-groups, syntagms, sentences, paragraphs, texts, cannot be reduced to a mere chain of phonemes. Together with the relationship between the units of the lower levels / 4: 257/ it contains a special constituent - prosody - performing some specific functions.

Intonation takes part in forming the sound matter of a complex language unit primarily as a means of organizing its components.

However, the fact that speech units (syntagms, utterances, paragraphs) can be organized in various ways enables intonation to express some specific meanings, either alone or in combination with other non-intonational means.

In some of its functions intonation is more conventional, that is its units are very close to the conventional language signs. This can be seen in some communicative types of utterances and some emotional reactions where such "meanings" as for example "question", "agreement", "non-agreement", "surprise","doubt" have their "own" corresponding intonation patterns, which in some cases are used as the only means of conveying information. It can be proved by the fact of correct perception of the so called "pure" intonation (intonation without words) in colloquial Russian (hm? - hm. - hm! - hm?! - hm!!).

In its other manifestations (significantly more common) intonation reveals its non-sign nature. I believe that in such functions of intonation as the delimiting or the prominence-lending, intonation patterns can hardly be viewed as bilateral linguistic signs (for the opposite point of view see, for examlpe /5/).

So, obviously, the author of an intonation theory will regard intonation units as more or less sign units depending on the fact which of its functions are in the centre of his investigation.

Thus, the third reason for rather a vague status of intonation among the other linguistic aspects is different interpretation of the essence of intonation, the extent of its functions and the character of its units.

INTONATION: SYNTACTIC OR SEMANTIC?

The great variety of view points on intonation , in my opinion, can only be explained by the complexity and heterogeneity of its phenomena.

The fact that intonation has long been outside the sphere of main linguistic problems can be accounted for not by an insufficient amount of experimental data but by the fundamental difference of intonation from other linguistic phenomena. The term "intonology" coming nowadays into use in the Soviet Union is a symptom of the realization of this difference.

Paradoxical as it may seem any further

deepening of the investigation of intonation aggravates the state of things.

Large quantities of experimental data show that intonation correlates more closely with the semantics of an utterance than with its syntax.

Thus, the so called communicative types of utterance as well as many phenomena of sentence accentuation are, in fact,semantic by their nature. For example, some features of the intonation structure of an utterance may be caused by the specific meaning of words and word-groups constituting this utterance.

The results of the experiments proved that the metaphorical use of the word, its specific semantic capacity, the presence of a number of meanings (semantic components) in its semantic structure (as for example, the meanings of evaluation, contrast, result, negation /6,7/) serve as factors causing a greater degree of its prominence. On the contrary, the semantic "emptiness" of some of the words is the cause of their weaker accentual prominence.

Thus, intonation which has long since been called "syntactic phonetics" has a real chance of being called "semantic phonetics".

INTONATION AND THE LEVEL STRUCTURE OF LANGUAGE

The problem of the place of intonation in the language system as far as the levels of the language structure are concerned is even more complicated than the problem of the place of intonation in linguistics.

It is significant that in the majority of conceptions of the language levels system the place of intonation is not put forward for special discussion.

On the other hand attempts "to insert" intonation with its all-embracing means and functions into the proposed hierarchy of the levels show the inadequacy of the construction itself.

One of the main difficulties of the traditional approach to language levels,where the phonetic (phonological) level is considered to be the lowest, is the impossibility of explaining how the non-sign elements of language level construction - phonemes - form the meaningful linguistic units - morphemes - at the next level. The inclusion of intonation into this lowest level aggravates the difficulty still more, for intonation units form neither morphemes nor other linguistic units, the relation between them being quite different.

Thus, not everything that is included in phonetics can equally be included in the phonetic level, considered to be the lowest level of the language structure.

To avoid this difficulty we might assume that intonation units are language signs (which - in my opinion - is true only for a smaller part of intonation phenomena), but this gives rise to another difficulty.

The attempt to include these "intonation signs" into one of the traditional sign levels reveals the variety and the specific character of such signs.

On the other hand it is impossible to place "intonation signs" between the traditional sign levels, as no sign level can either be composed of or decomposed into intonation units alone.

The only solution is to remove intonation from the hierarchy of levels and to assign intonation phenomena to different sign levels in accordance with the variety of intonation functions. Then intonation will easily find its place in the language structure, but only as an inherent element of the plane of expression of the complex signs. Intonation is related, in some of its functions, to such complex units as the sentence, the paragraph and the whole text. In its other functions , intonation is connected with smaller units, such as the syntagm and even the rtythmic group.

This approach, natural and even traditional as it is (compare the usual distinction between segmental and suprasegmental means, or features), involves a certain contradiction. How does this high level use of phonetic means correlate with the fact that the phonetic level is considered to be the lowest one? I think that intonation facts demonstrate the inadequacy of this conception.

The phonetic means, i.e. sound matter in all of its aspects (both segmental and suprasegmental) naturally correlate with different levels of language structure (beginning with the word and up to the text) due to the fact that they do not form a separate level but an aspect without which no level can exist.

The acknowledgement of the specific role of the sound matter of language predetermined by human nature, which in its turn predetermines the main properties of the language, is extremely important for the conception of the Shcherba Phonological School. It is well-defined in the works by L.R.Zinder /2/ and L.V.Bondarko /3/ and especially in the recent book by V.B. Kasevich "Phonological Problems in General and Oriental Linguistics" /4/, in which the author proposes a new conception of the phonological component of language.

CONCLUSION

The study of such a complex and specific phenomenon as sentence intonation leads us to the conclusion that the conventional division of language into separate levels is too straightforward and should be considered more critically.

A revision of some of the ideas concerning the place of phonetics and intonation in the language system makes it possible to describe intonation in all its manifestations within the framework of phonetics as linguistic science.

ACKNOWLEDGEMENT

I want to avail myself of this opportunity to thank my teacher Professor L.R.Zinder for his constant assistance in my study of Russian intonation and for helpful comments on an earlier version of the paper.

REFERENCES

/1/ Russkaja grammatika /Russian Grammar/ Moskva, vol.I - 1980, vol.II - 1982
/2/ L.R.Zinder Obshchaja fonetika /General Phonetics/, 2nd edition, Moskva, 1979
/3/ L.V.Bondarko Foneticheskoje opisanije jazyka i fonologicheskoje opisanije rechi /Phonetic description of language and phonological description of speech/, Leningrad, 1981
/4/ V.B.Kasevich Fonologicheskije problemy obshchego i vostochnogo jazykoznanija / Phonological problems in general and oriental linguistics /, Leningrad, 1983
/5/ T.M.Nikolaeva Frazovaja intonacija slav'anskikh jazykov / Sentence intonation of slavonic languages /, Moskva, 1977
/6/ T.M.Nikolaeva Semantika akcentnogo vydelenija / Semantics of accentual prominence /, Moskva,1982
/7/ A.V.Pavlova, N.D.Svetozarova Faktory, opredel'ajushchije stepen' akcentnoj vydelennosti slova v vyskazyvanii /Factors determining the degree of accentual prominence of a word in an utterance/.-In: Slukh i rech v norme i patologii /Hearing and speech in norm and pathology/,Leningrad, 1987

INTONATION AND WORLD CONCEPT IN A LITERARY TEXT

IRINA G. TORSUYEVA-LEONTYEVA

Translator's Department
Maurice Thorez Moscow State Institute of Foreign Languages
Moscow, USSR 119034

ABSTRACT

The report presents one of many possible ways to interpret the phonetic structure of a text i.e. the way a "world concept" is expressed by phonetic means.

A world concept, or a comprehensive ideological model intrinsic to a given type of culture, is an invariant represented through different variants, i.e. scientific, artistic, religious or epic, etc. pictures of the world. These variants, in their turn, are actualized in certain types of texts that can be considered concrete realizations of a general world concept.

Certain varieties can be found within the artistic concept of the world such as, for instance, the poetic concept of the world. A concrete text is the reflexion of an author's vision of the world. Analysis of this produces an individual world model. It does not exist by itself, but is conditioned by the epoch's general world model so that the movement takes place here from the invariant to the variant and thence to the concrete realization.

The main constituents of a world concept are Man and his system of values within objective reality, and Space and Time as forms of the existence of matter.

As the author's internal individual world model is implemented in texts, language resources are being used inherent to a language community and thus participating in their turn in producing a reflexion of the objective world. Study a text's world concept expressive means is mainly centered around lexico-semantic means, grammatical patterns and stylistic devices. The part played by intonation in communicating the literary text world concept has practically eluded investigation and up to now only the first steps have been made in this respect.

Intonation is latent in the text and intonational actualization of this, if adequate to the author's conception, corrects the decoding by the recepient.

Different possibilities of perceiving the text are now widely studied in contemporary linguistics and literary criticism. However, such analysis is incomplete without using the data of intonational analysis, both auditive and acoustic.

Text linguistics deals mainly with written texts. From its viewpoint any concrete text is seen as one text. But for an intonation specialist any reading of the text is a new text. Hence we would suggest that the written text be considered the basis for analysis, and each concrete reading (actualizing in sound form), of the text be considered as its intonational realization. It is in this case that the possibility of multiple perceptions of the basic text (plurilecture) stands out most clearly.

It is only natural that in different

realizations common structural features inherent to the intonation of a given language will be seen. The common element in reading (or intoning) the text is also determined by referring the basic text to a certain style and genre. An eventual set of interpretations of the author's conception is not unlimited, although it is an open list of possibilities. Further constraints on the number of possible interpretations are also imposed by the fact that the recepient belongs to a certain epoch and perceives the text within the limits of a world concept intrinsic to a given epoch and type of culture. In this way the variance of text intoning by different speakers of a language is within the bounds of variance set by the concrete language and the epoch's world concept.

A student of the literary text intonation may have the following objectives:

(1) studying structural characteristics of the intonation form of the text;

(2) studying whatever is common among the different interpretations of the basic text;

(3) studying the reflexion of the author's internal individual world model in a concrete realization.

The latter, on the one hand, provides the ground for seeking cut the generic element, and on the other is interesting in itself as an individual interpretation. Every individual actualization of the text is both the reflexion of the general intonational parametres and, at the same time, of the author's conception. Perhaps with a greater number of such studies general regularities could be ascertained concerning the functioning of intonation in a literary text.

Intonation can not be studied in isolation from other means of text formation. To bring out clearly the world model of a text it is especially important to ascertain the relationship between lexico-semantic means and intonation. Studies of this type have been rare in linguistics.

Text is a system, i.e. a structurally and functionally integral entity where in internal relation among the elements is more deterministic and stable than their relation to the environment or elements of other systems. Its main principles as a system are that it is structured, hierarchical, integral, sustaining a mutual relationship between itself and the environment, functional, objective, admitting of multiple descriptions. Actualization of all language recources in a text is determined by its whole. These recources may be oriented towards the same direction or towards a number of directions providing opportunities for different interpretations of the text. The study of the world concept intrinsic to the text shows the relationship between the system and the environment whose constituents are the author and the recepient.

The study of lexical units in order to ascertain the underlying concept of the world is carried out by Yu.N.Kharaulov /1/ and Yu.K.Lekomtsev /2/. Such studies may provide the ground for bringing out the world concept in the text through intonation. Therefore we think it necessary to investigate ways and types in which intonation interacts with the lexico-semantic matter of the text. It is essential to distinguish here between the content and the sense (message) of the text. Its content is the reflexion of a certain fragment of the real (or imagined) world. The sense of the text includes its appraisal, both intellectual and emotional, of this fragment. This sense, which is potentially present in the basic text,finds its expression when the recepient comes in contact with it. In intonation studies the recepient is the speaker reading the text and the listener who hears it as it is being read. There is a complex overlapping of the two types of perception, the comparison of which permits to appreciate the author's internal -individual world model.

The studies we carried out or supervised deal with different aspects of the world concept viz. reflexion of mythical poetic semantics in a folklore text through intonational means; finding out latent semantics of lexical units in the text,prose in fiction being used as source material; bringing out the world concept underlying descriptive texts; ascertaining space-time organization of the text by means of intonation, revealing with more precision the way literary plots are expressed, etc.

The foundation for such studies is provided by semantic analysis of the text. Results of this should be further compared with the data obtained by the auditory and intonographic analyses.

Hence the relationship between lexical semantic units and intonation is brought out in the actualization of the world concept reflected in the text.

Let us cite by way of example the intonation used with adjective of colour of a French fairy-tale. In French, a very important part is played by these adjectives in the nomination process. Colour designation is a vivid means of describing the world concept. The colour spectrum is represented in the tale by achromatic (blanc, gris, noir) and heraldic colours (or, argent) as well as by phrases with the word 'couleur'. A prominent place is held in it by the epithets d'or, doré, blanc, rouge, bleu, noir. Colour is involved in suggesting images of its characters (protagonists, their helpers, givers and harmdoers), various magic objects, elements of space. Fairy-tale possesses a glossary of colour symbolism with origins in the remote past. This symbolism is related to manifestations of mythological thinking. Thus, for instance, colour designation of the 'gold' group is connected with the reflexion of the sun cult.

Intonation actualizes potential semes contained in the structure of colour designation. So, for example, the semes 'superb , beau'are brought out in the lexeme 'doré'. The acoustic correlates of semê actualization are the temporal characteristics, ASP_{max} (amplitude of sound pressure), the fundamental frequency direction. 'White' (blanc, blanche) marks the outstanding, which finds its expression on the intonational level as well. The acoustic correlates of accentual prominence given to this meaning are the temporal characteristics F_0 direction. The colour terms 'bleu' and 'noir' potentially have negative meaning. The acoustic correlates of potential semê actualization is F_0 level and voice quality.

Intonation is an important indicator of the space and time organization of the text. Textual lexical elements directly designate action,time and place. In the novel 'La neige en deuil' by Henry Troyat there is a passage where action takes place in the protagonist's hut in the mountains at night:

"La bise sifflait derrière le vantail. Des poutres craquaient. La nuit de neige s'appuyait contre la fenêtre aux carreaux constellés de givre. Un peu plus tard, la porte de l'écurie étant restée ouverte, les brebis pênétrèrent l'une après l'autre, dans la maison. Elles marchaient â petits pas, humant les meubles, lêchant le salpêtre des murs, s'appelant et se rassurant d'une voix tremblante. La lumière les guidait. Et l'odeur du maître. Elles s'assemblèrent autor du lit. Un grouillement laineux emplit la chambre. L'hindoue semblait flotter sur un nuage de toisons pâles et touffues. Isaîe caressait le dos des bêtes et disait:

- Ne faites pas de bruit ... Vous voyez, elle dort."

The space of artistic creation within this passage hinges around the opposition between the external hostile world (world of evil) and the internal space of home (world of good). The semantic complex reflecting the former is grouped around the word 'nuit' and characterized by the lexemes 'bise, neige, givre, siffler, craquer'. The semantic complex reflecting the latter is built around the word 'lumière' and represented by the lexemes 'maitre, nuage, grouillement, toisons, laineux, pâle, touffu, se rassurer, flotter'.

The simple action of looking up a dictionary definition does nothing to reveal what the semantics of these words has in common. At the intonational level they all stand out. The centre of emphasis in the passage are the lexemes 'la nuit de neige' and 'lumière'. Thus intonation marks the conceptual, and not the factual information in the text, i.e. the author's vision of the world. This is even more vividly so in descriptive texts. So in John Galsworthy's 'The Forsyte Saga' series ('The Man of Property') there is a passage describing approaching spring in its influence upon Man and Nature:

"The driver turned once or twice, with the intention of venturing a remark, but thought better of it. They were a lively couple! The spring had got into his blood, too; he felt the need for letting steem escape, and clucked his tongue,flourishing his whip, wheeling his horses, and even they, poor things, had smelled the spring, and for a brief half hour spurned the pavement with happy hoofs.

The whole town was alive; the boughs curled upward with their decking of young leaves, awaited some gift the breeze could bring. New-lighted lamps were gaining mastery, and the faces of the crowd showed pale under that glare, while on high the great white clouds slid swiftly, softly, over the purple sky.

Men in evening dress had thrown back overcoats, stepping jauntily up the steps of Clubs; working folk loitered, and women - those women who at that time of night are solitary, solitary and moving eastward in a stream - swung slowly along with expectation in their gait, dreaming of good wine and a good supper, or, for an unwonted minute, kisses given for love.

Those countless figures, going their ways under the lamps and the moving sky, had one and all received some restless blessing from the stir of spring. And one and all, like those clubmen with their opened coats, had shed something of caste, and creed, and custom, and by the cock of their hats, the pace of their walk, their laughter, or their silence, revealed their common kinship under the passionate heavens."

Thematic grouping were individualized in the text which are hierarchically organized into three levels of the semantic structure of the fragment:

Level 1 - Town, Time, People;
Animals, Sky, Earth
Level 2 - Spring, Love, Blessing;
Level 3 - Common Kinship.

The integration of the text, or creation of a comprehensive world concept (kinship of all living things on earth in love and renewal) takes place at the third hierarchical level. Intonationally the lexemes 'spring, blessing, love' etc. are outlined most vividly. They could be seen as keywords.

The correlates to this emphasis are the high falling tone, the incomplete high falling tone, short pauses, the broken descending scale, emphatic stress.

Hence preliminary studies make it possible to conclude that intonation is functionally charged in reflecting the world concept in a literary text.

REFERENCES

/1/ Yu.N.Kharaulov, "Linguistic Foundations of the functional approach towards literary criticism", in Problems of Structural Linguistics, 1980, Moscow, 1982.

/2/ Yu.K.Lekomtsev, "Antonomic Text", in Text - Semantics and Structure, Moscow, 1983.

TOWARDS A UNIFIED FRAMEWORK OF RUSSIAN INTONATION

OLGA T. YOKOYAMA

Harvard University

ABSTRACT

A generative framework of Russian intonation is proposed that incorporates both semantically significant pitch contours and so-called sentential stress, two phenomena usually treated separately in existing accounts of Russian intonation.

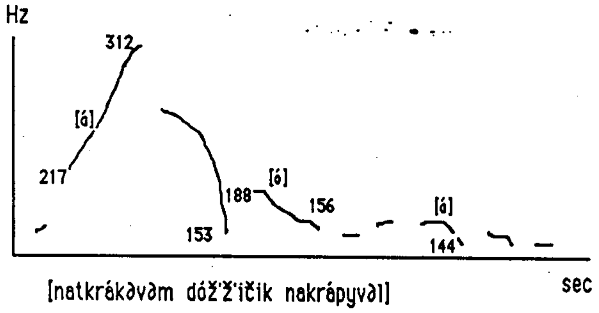
The proposed framework involves at least two levels: a phonemic level consisting of both pitch level and contour tone sequences, and a phonetic level accessible to perceptual and instrumental analysis. The phonetic level is generated as a result of intertonal level mapping processes and general implementation processes like downstep, upstep, or declination. The descriptive data include those observed by previous researchers, as well as the author's own instrumental measurements of fundamental frequency.

§0. In this paper I outline some basic considerations for a framework of Russian intonation that incorporates both semantically significant pitch contours and so-called sentential stress (SS). In existing accounts of Russian intonation, the two intonational phenomena are usually treated separately. Some accounts of intonation do not mention SS (usually called "logical stress" in the Russian scholarly tradition) at all, as is the case with the authoritative description of Russian intonation in the 1982 Academy Grammar [3]; under this approach, intonation is limited to a taxonomy of meaningful pitch contours, i. e., to a lexicon of Russian intonational meanings. Other accounts (e.g. [12]) treat SS together with other types of stress (e.g. word stress, phrase stress); such accounts generally do not examine pitch contours associated with SS, but define SS in terms of intensity and length. Finally, in accounts of speech melody that examine pitch contours of utterance types both with and without SS (e.g. [1] and [7]), no attempt is made to construct a unified intonational system.

In the following presentation, I consider both SS and certain melodic contours within an intonational system that involves (at least) two levels: a phonemic level that consists of both pitch level and contour tone sequences, and a phonetic level accessible to perceptual and instrumental analysis. The phonetic level is produced as a result of intertonal level mapping processes, as well as general implementation processes like downstep, upstep, or declination. Specifically, I concentrate below on the following three points: (§1) characterization of SS in terms of the direction of the pitch in the stressed syllables of post-SS segmental material; (§2) accounting by means of downstep for iterated sequences of rising contours in non-utterance-final syntagms; and (§3) positing separate phonemic pitch level boundary tones (BTs) not associated with lexical stress.

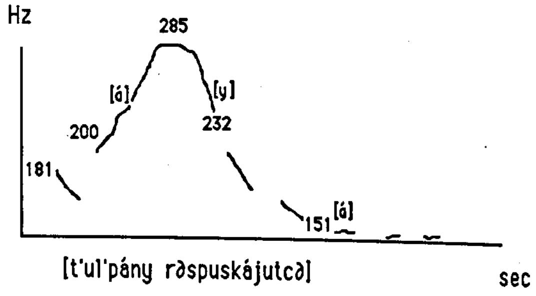
§1. As is well known, the concept of SS is indispensable for a comprehensive description of Slavic word order. In the absence of an explicit phonetic or phonological definition of this concept, however, it has been taken as a primitive by all scholars dealing with word order. In the actual process of investigation, this amounts to relying on an ill-defined introspective criterion, which has led to some misguided analyses of word order data. It is clear that neither absolute amplitude nor absolute pitch signals SS in Russian, since SS can be placed toward the end of the sentence, where both absolute amplitude and absolute pitch are always lower than they are at the beginning. Also, the duration of the syllable carrying SS is often shorter than that of some other intonational centers in the same sentence, especially the stressed vowel of the sentence-initial rising syntagm (*neokončennaja sintagma* 'incomplete syntagm' in [1], or *načinatel'naja melodema* 'initial melodeme' in [4]). The direction of the pitch in the stressed syllable of the SS itself is not distinctive either, since SS can have either a rising or a falling pitch contour (cf. [1] and [7]). I suggest that the SS site is determined without reference to its own prosody, intensity, or duration, but rather relatively and negatively, as the leftmost intonational center after which no syntagms and no rising contours occur. Thus in (1), SS is on the word *doždiček* 'rain', on whose stressed syllable [ó] a falling accent HL is implemented; the first syntagm has a rising intonation LH on its stressed vowel [á], while the other stressed syllable in the same syntagm with *doždiček* 'rain', i. e., the post-SS stressed [á], has falling pitch:[1]

(1) Nad Krakovom doždiček nakrapyval.
'It was drizzling over Krakow.'



When the direction of the pitch in SS is rising or rising-falling (which happens mostly in interrogatory and exclamatory utterances), the pitch contour of the post-SS stressed vowels is still falling; consider (2), where the SS is on *tjul'pany* ' tulips', a word with a rising-falling pitch on its stressed [á] , and the stressed [á] after the SS has falling pitch again:

(2) Tjul'pany raspuskajutsja!
'The tulips are opening!' (with stress on 'tulips')



When SS occurs in a monosyntagmatic utterance, it can be found anywhere in its syntagm. In monosyntagmatic utterances where SS falls on the first stressed syllable, no rising stresses occur at all. In such cases, however, when the post-SS segmental material is too extensive, and the pitch cannot continue falling due to the limits imposed by the base line [9], upstep is implemented [5]. When SS appears in non-initial position within its syntagm, the pitch level before SS is generated by a phonemic pitch level accent, which may or may not generate phonetically rising word stresses. When SS occurs in a multisyntagmatic utterance, on the other hand, it is always preceded by a syntagm with rising intonational center; significantly, segmentaion into syntagms is not possible after SS.

. The definition of SS as formulated above connects certain previously made observations concerning logical stress [10]. with Bryzgunova's IK inventory. It becomes clear, for example, that when the intonational center in Wh-questions uttered with IK-2 is found on the Wh-word, the Wh-word is the SS; or that the intonational center in IK-3 also turns out to be the SS. Moreover, since the function of SS is to mark that piece of information which is not part of the addressee's knowledge and/or current concern [11], this definition of SS sheds light on the functional dichotomy of certain utterance types. Thus, while the SS on the Wh-word of IK-2 marks the rheme, the rest of the information can automatically be judged to be part of the addressee's current concern. Similarly, in questions with IK-3, the SS marks the disjunctive information (i. e. *x* or *y* ?) that is outside the speaker's knowledge, and the rest of the proposition is also part of the addressee's current concern. In both cases, the SS status of the intonational center is not only consistent with the fact that the pitch contour of SS itself can be either rising or falling, but it is also corroborated by the contour of the tail, which lacks rising stresses and forms no syntagms.

These considerations indicate that some of the more abstract meanings of IKs, specifically those associated with the structure of the discourse (such as theme/rheme) as opposed to attitudinal factors (cf. some IK meanings like skepticism, disapproval, enthusiasm) constitute on the one hand a separate group within the intonational lexicon, while on the other, they must be considered as an intergal part of the framework of Russian intonation as a whole.

§2. As suggested by Ščerba and repeated by most subsequent scholars, some utterances do not have SS at all; they are composed of one or more syntagms, each of which has its own syntagmatic (or phrasal) stress. Consider (3):

(3) V našu komnatu / vošla požilaja ženščina /
v bol'šix / mužskix / sapogax.
'A middle-aged woman in big men's boots walked into our room.'

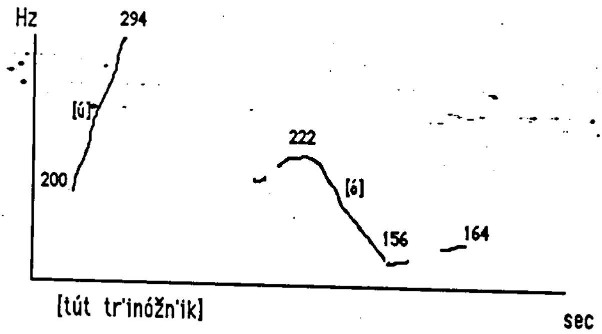
In (3), which can easily be uttered with as many as 5 syntagms, each of which but the last has a rising center, the only falling stress is realized on the final syntagm *sapogax* 'boots', which is nevertheless clearly not the carrier of SS. I will call this utterance intonation "Type I".

Type I intonation brings us to the second problem, that of downstep. The intonational contour described for (3) is essentially that of *Moskovskoe vremja / četyrnadcat' časov / pjatnadcat' minut* 'Moscow time is fourteen hours and fifteen minutes', which is analyzed as IK-6 / IK-4 / IK-1 [2:195]. Note that the number of rising stresses (like those in IK-6 and IK-4) in sentences with Type I intonation can easily be increased; in (3), for example, four rising stresses are quite possible, depending on the rhythm of the speaker, and this number can be increased by lengthening the sentence. Attributing each rise to a separate phonemic IK that differs from the preceding one only in having a slightly lower rise misses the generalization that such iteratively rising "slopes" call for. I suggest that this pattern of Type I intonation can be accounted for in terms of downstep.[2] If downstep is accepted as part of the Russian intonational system, the invariable core of Type I intonation can be described as $[LH]_n$ HL, where n is the number of non-final syntagms, and [ ] indicates the implementation of downstep. The surface pitch level is then generated as a result of a combination of downstep, intertonal pitch mapping between adjacent phonemic tones, and the general declination of the utterance.[3]

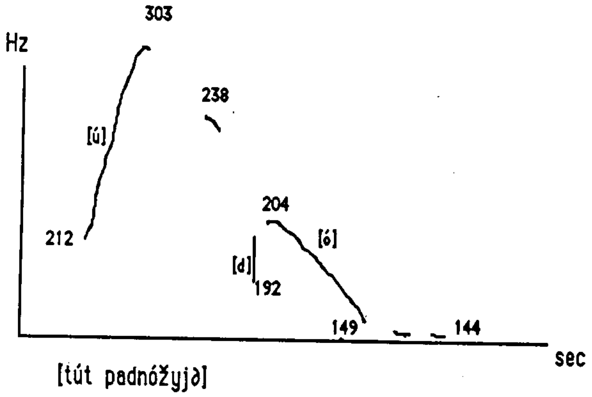
This solution eliminates the *ad hoc* assignment of different phonemic rising IKs that gradually decrease in height to an open set of syntagms, which runs counter to the obvious fact that the number and height of the intermediate pitch levels is nothing more than a function of the length of the sentence and its division into syntagms. This solution accounts, moreover, for numerous redundancies observed in the current system of IKs, such as the otherwise unexplained synonymy of IK-3, IK-4, and IK-6, all three of which are said to signify "incompleteness", among other things. An additional benefit of such a deep structure is the fact that it also accounts for our intuition that sentential Type I intonation and so-called "citation intonation" for single words (cf. also Bryzgunova's observation that IK-1 is used as "title" intonation [3]) are quite similar. Thus one can posit a core phonemic representation for both Type I and citation intonation as $[LH]_n$ HL, where $n = 0$ in citation intonation.[4]

§3. I suggested in §2 that the underlying sequence of contour tones $[LH]_n$ and HL constitutes the core of Type I intonation. This core, however, is not entirely sufficient for generating all utterances that should be considered Type I. Consider the contrast between (4) and (5):

(4) Tut trenožnik. 'There is a tripod here.'



(5) Tut podnožie. 'There is a pedestal here.'



Both (4) and (5) have Type I intonation LH HL. But the post-tonic pitch levels in the final syntagms of (4) and (5) differ significantly. To account for this, as well as other phonemic differences realized on unstressed material, I propose phonemic pitch level boundary tones (BTs).[5] Thus, if we posit utterance-final BTs H# for (4) and L# for (5), we can capture both the Type I intonation that (4) and (5) clearly share, as well as the non-finite nature of (4), as expressed by the H#. The phonetic realization of BTs is of course affected by the adjacent tone level, as well as by the overall declination of the base line.

BTs can also account for similarities and differences between several other contours that must otherwise be distinguished as a whole. For example, the difference between the first and the second type of syntagmatic stress as described in [1] can be reduced to LH for the first type, versus LH H# for the second. Similarly, the difference between "qualificational"

IK-3 and IK-6 (as in *Kakoj sup vkusnyj! 'What a yummy soup!'*) can be reduced to that of the BT, which is L# in IK-3 and H# in IK-6.

The specific points discussed above do not of course exhaust the theoretical questions associated with the description of Russian intonation. Among remaining problems, the problem of the definition of a syntagm is crucial for the understanding of the generative process of utterance intonation. Of the many suggestions for defining a syntagm, at least the intonational definition, as the domain of a single intonational center, appears to be generally valid (with the exception of the "bicentral" IK-5, cf. fn. 4). But this leaves at least two important questions unanswered: (a) how the boundaries of syntagms are determined (cf. e. g. [8]), and (b) what determines the position of the intonational center within the syntagm itself. Although various authors have offered speculations on both questions, no rules have yet been proposed that would *generate* a correct segmentation in a given context. There are also general descriptive issues to be addressed, such as whether or not Russian intonation is best represented by a sequence of pitch level and/or contour tones (as assumed in this paper) or by head-nucleus-tail configurations (this is essentially the approach taken by Soviet scholars), or how many basic tone levels can be posited for Russian, or which tones, if any, can spread. The answers to these questions would enable us to determine the underlying tones for various IKs, and to incorporate *all* of the items of the intonational lexicon, including discourse features marked by SS, into a unified framework of Russian intonation.

NOTES

1. The graphs given in (1), (2), (4), and (5) were produced by a computerized analysis of changes in fundamental frequency over time. The informant was an ethnic Russian female from Leningrad in her thirties; the graphs use a regular (i.e. not a logarithmic) scale. This instrumental research was supported by NSF Grant BNS 8206064.

2. Downstep, which is an important feature of the tonal systems of many African and native American languages (see e.g. [5]), has also been proposed for English [9].

3. For declination, see [9].

4. The only bicentral in the system, namely IK-5, may also be essentially represented by the same phonemic sequence, where it is perhaps the intonational meaning of this IK that obligatorily requires $n = 1$, along with some other peculiarities in the underlying structure of this contour.

5. For boundary tone, see [6].

REFERENCES

[1] Bryzgunova, E. A. 1963. *Praktičeskaja fonetika i intonacija russkogo jazyka.* Moskva: Moskovskij universitet.

[2] ________. 1977. *Zvuki i intonacija russkoj reči.* Moskva: Russkij jazyk.

[3] ________. 1982. 'Intonacija', in N. Ju. Švedova (ed.), *Russkaja grammatika* , 96-122. Moskva: Nauka.

[4] Čeremisina, N. V. 1976. 'Melodika i sintaksis russkoj sintagmy', in G. A. Zolotova (ed.), *Sintaksis i stilistika* , 65-85. Moskva: Nauka.

[5] Clements, George N. 1979. 'The description of terraced-level tone languages', *Language* 55:536-58.

[6] Liberman, Mark. 1978. *The Intonational System of English.* MIT Dissertation.

[7] Matusevič, M. I. 1976. *Sovremennyj russkij jazyk: fonetika.* Moskva: Prosveščenie.

[8] Nikolaeva, T. M. 1973. 'Smyslovoe členenie teksta, ego individual'nye varianty', in M. Mayenowa (ed.), *Semiotyka i struktura tekstu* , 71-80. Wroclaw: Wydawnictwo Polskiej Akademii Nauk.

[9] Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation.* MIT Dissertation.

[10] Raspopov, I. P. 1957. 'Smyslovye i stilističeskie funkcii logičeskogo udarenija', in *Russkij jazyk v skole* 1957, 4:18-22.

[11] Yokoyama, Olga T. 1986. *Discourse and Word Order.* Amsterdam: John Benjamins.

[12] Zinder, L. R. 1960. *Obščaja fonetika.* Leningrad: Leningradskij universitet.