AUDITORY PROCESSING OF SPEECH

Ludmilla A. Chistovich, Pavlov Institute of Physiology of the Academy of Sciences of the USSR, Leningrad, USSR

I shall outline the approach (Chistovich, Ventsov, Granstrem et al., 1976) adopted by our group in studying auditory levels in speech perception. After reading Studdert-Kennedy's report, I realized that some explanation of our reasons should be presented.

When phoneticians describe the acoustic cues they usually refer to some "objects" or "events" seen on the dynamic spectrogram, such as gaps, transitions and so on. Studdert-Kennedy has presented very good evidence that the parameters of these events (for instance, the duration of a gap, the direction of a formant transition) as well as the temporal order of the events displayed over intervals of roughly syllabic length are utilized by the human being in the phonetic interpretation of the message. Unlike Studdert-Kennedy we were not able to suggest any procedure for automatic phonetic interpretation which would conform with known experimental data on speech perception without assuming the preliminary conversion of the speech signal into a flow of events. To make this clear I shall mention only one quite trivial problem - the problem of the measurement of duration. Duration is the interval of time between two events. This parameter does not exist at all if the events delimiting the interval are not specified. That seems sufficient to explain why our group became interested in the auditory bases for the detection of events.

Neurophysiological studies of the central auditory system have revealed a highly ordered tonotopic organization at each anatomical level, with several representations of the frequency scale at the same level. This suggests that the original peripheral excitation pattern is transformed into a number of versions, with the axes of the pattern remaining unchanged. Stimulus-response relations for the central auditory neurons hint at the extraction of irregularities in the pattern along both frequency and time axes. There are indications that the width of the window of processing increases both in time and in space (frequency range) at higher levels of the system.

This led us to believe that the detection of irregularities in the speech-induced excitation pattern might be an essential part

of signal processing, and that the psychoacoustic study of the detection of irregularities might be a good starting point. To see how models, based on psychoacoustical data, will react to real speech, one has to build them as working signal-processing systems. A functional model of the cochlea is a necessary instrument for this approach. Our group is exploiting a linear model of the cochlea built as a 128 channel analyzer (Goloveshkin et al., 1978). Parameters of the model have been adjusted according to tuning curves for auditory nerve fibers. Dynamic spectrograms of speech obtained from this model are somewhat different from the conventional dynamic spectrograms, but almost all the details believed to be important are preserved.

So far we have confined ourselves to the simple kinds of irregularities: irregularities of the envelope and irregularities of the spectrum shape of a steady-state stimulus.

Processing of the envelope

Slow changes in the stimulus envelope are perceived as loudness changes. Rapid solitary irregularities such as jumps, drops, small gaps, hills and valleys give rise to associations with consonants. Although subjects cannot indicate any particular consonant with certainty, they have no difficulty in deciding whether the consonant associations are the same for two stimuli and whether they are present or absent. This allows us to use the classical psychoacoustical approach and to measure the quantitative relations between parameters when their effects are equal. The data concerning relations between the parameters of envelope changes (Chistovich, Ventsov, Granstrem et al., 1976; Stoljarova and I. Chistovich, 1977; I. Chistovich, 1978) suggest processing close to band-pass filtering, with the center frequency of the envelope filter being around 25 Hz. Having assumed band-pass filtering, we had to decide whether it is sufficient to use a single filter in the model, with something like instantaneous loudness being the input signal, or whether it is necessary to use a number of filters, each processing information within a restricted frequency range. It was found that although the association of a small gap in a pure tone with [r] does exist over a wide frequency range, it disappears when the tones preceding and following the gap differ in frequency (Lesogor, 1977). The critical mistuning of the two tones appears to be close to the critical band well known in psychoacoustics. The masking of the jump in the envelope of one tone by a simultaneously presented second tone has also been studied (Lesogor et al., 1978). The tone with the jump was held constant in frequency, intensity and jump amplitude, while the frequency of the masker ($F_m$) was varied and the minimal level ($L_m$), necessary to make the "consonant" disappear, measured. The resulting $L_m$ vs. $F_m$ curve appeared to be very similar to the so-called "psychoacoustical tuning curves". These data indicate that the detection of envelope irregularities requires our model to be multichannel. The simplest solution is to place one envelope filter at the output of each channel of the "cochlea".



In the first version of our frequency selective model for processing stimulus envelopes (I. Chistovich, 1978), half-wave rectification with a memoryless compressive nonlinearity was used to simulate the mechanical-to-neural transformation in the cochlea. Two (positive and negative) thresholds were placed at the output of the envelope filter, their crossing resulting in onset- and offset-markers. Better agreement with psychoacoustical data was achieved when peripheral short-term adaptation was also incorporated in the model.

This multichannel model, including the "cochlea", has been built as an analog system (Kozhevnikov et al., 1978). The model is good at detecting the rapid spectral and intensity changes in speech signals, and could be used for the automatic segmentation of speech. Although it is blind to the formants in a steady-state stimulus, it succeeds in tracking formant transitions. The model is not yet satisfactory from the psychoacoustical point of view. It cannot reproduce the above mentioned frequency selective effects in jump and gap detection, which seem to require that some spatial (interchannel) interaction must be incorporated in the model.

A serious problem concerns the combining of markers over the frequency scale. Data on perception of amplitude irregularities on the widely spaced components of a complex stimulus (Rodionov et al., 1976; Lesogor and Chistovich, 1978) indicate some kind of summation over a wide frequency range. The temporal threshold for jump detection (minimal interval between the stimulus onset and the jump) appeared to be equal to the threshold of nonsimultaneity for the onsets of two tones (Kozhevnikova, 1978). This also points to

summation. The threshold was found to be insensitive to "selective adaptation" (Ogorodnikova, 1978).

Summation of the markers would be useful for locating exactly the moment of change, but it will lose information about the frequency region where the change occurs. So far we have failed to find any evidence that the subject is able to pick up the frequency component which is the carrier of the amplitude jump or the gap. At the same time we have found that the subject "knows" the stimulus spectrum shape at the moment when the irregularity occurs (Zhukov et al., 1974; Zhukov and Lissenko, 1974). When a small gap was moved along a [iu] stimulus (F1 - steady-state, F2 - time-varying), subjects were able to locate the point corresponding to the shift from [iru] to [iɕu]. Gaps with different durations were adjusted by subjects in such a way that the end of the gap always coincided with the same value of $F_2$. Subjects could do this just as easily when the time-varying $F_2$ was presented to one ear while the gap (in the stimulus with steady-state formants) was presented to the other ear.

Segments of the signal between onset and offset markers cannot be regarded as phonetic elements at this stage of processing. Temporal rules are used by the subject in accepting (or rejecting) the vowel-like segment as a vowel - the element of rhythmic pattern. These rules are based on segment duration as well as on the duration of the onset-to-onset interval (between one segment and its successor) and on the offset-to-offset interval (between one segment and its predecessor) (Chistovich, Ventsov, Granstrem et al., 1976).

Spectrum shape processing

Two-formant stimuli with widely spaced formants are convenient for measuring the formant peak detection threshold, since the criterion of a shift in vowel quality can be used. The fact that the threshold depends on both the formant spacing and the steepness of the spectrum slope (Mushnikov and Chistovich, 1971; Chistovich et al., in press) suggests a process such as spatial differentiation of the excitation pattern. Stimuli with spectral peaks just below threshold and just above threshold have been used to adjust the parameters of a lateral inhibition model processing the output pattern of the "cochlea". The weighting function (spatial window) appeared to be quite narrow. The output of the model to a natural steady-state vowel is a spatial pattern with a number of peaks separated by zero-excitation intervals. To convert this pattern into a conventional formant description of the vowel, one has to identify its peaks with formants of the appropriate serial number and pick up the coordinate values (frequency position and amplitude) corresponding to the peaks. This procedure seems rather unrealistic from the point of view of the neurophysiology of hearing. The "center of gravity" effect (Delattre et al., 1952) indicates the spatial integration of a spectral pattern and suggests that the intermediate formant description of a stimulus might not be necessary for it to be identified as a vowel.



To test the "center of gravity" effect single-formant stimuli and two-formant stimuli with 350 Hz formant spacing (AI > A2 and AI < A2) have been used. Clear evidence for the effect was found in both the identification data and the matching data (Bedrov et al., 1978).

The "center of gravity" effect can be described in a qualitative way as $F_1 < F^* < F_2$, where $F^*$ is the frequency of the single-formant stimulus most close in vowel quality to the two-formant stimulus. Maximal spacing of the formants in the two-formant stimuli and the range of the formant amplitude ratio delimiting the area of the existence of the effect have been measured (Chistovich et al., in press). The critical spacing appeared to be equal to 3.0 - 3.5 Bark and the amplitude ratio range could reach 40 dB. Experiments on two-formant to two-formant matching for stimuli with more-than-critical formant spacing indicate that in this case the formant amplitudes are of minor importance, provided both formant peaks are above threshold. Stimuli with quite different A1/A2 values are most similar in vowel quality when their formant frequencies coincide. The data suggest a model with spectral peaks extracted at a lower level of processing and spatial integration at a higher level.

Our current attempts to simulate both the "center of gravity" effect and the unimportance of formant amplitudes with widely spaced formants apply a rather small set of spatial summators with overlapping summation intervals, each summator corresponding to one particular cardinal vowel. Assuming that the stimulus is described in terms of the distribution of the amount of excitation in the subset of excited summators, one is able, by using the model as an

instrument, to evaluate the similarity of two stimuli in vowel quality and to carry out the matching experiment. The set of cardinal vowels seems to be a better approximation to the inventory of vowels "known" by a Russian subject than the set of Russian phonemes. This follows from both the mimicking data (Avakjan, 1976) and the similarity scaling data (Kuznetsov, 1978). I should like to note that identifying summators with cardinal vowels is in fact one version of the template approach of which Studdert-Kennedy seems to disapprove.

There is no doubt that at least some of the parameters of the model must be made context-sensitive. A very strong adaptation-like effect was observed in the experiments on formant peak detection (Chistovich et al., in press). The nature of the effect is not yet analyzed.

Spectral cues in nonstationary vowels

The temporal parameters of spectrum shape processing are not yet known. It seemed useful to find out first what cues in the time-varying spectral shape pattern are important to the subject. Experiments with short two-formant vowel-like stimuli with linear and close to triangular (up-down and down-up) F2 contours revealed three cues used in phonetic interpretation (Lublinskaja and Slepokurova, 1977; 1978). One cue corresponds to $F_2$ or the spectrum shape value at the "target" point: the extreme point of the triangular contour and the end-point of the linear contour. The second cue corresponds to the initial value of $F_2$ or of the spectrum shape. The third cue is the direction of the initial F2-transition. This last cue appeared to be effective only in a restricted frequency range since it only serves to differentiate between [ʉ] and [ɨ] and between [ø] or [œ] and [ɛ] or [ə]. The boundary in the direction of the F2-transition is somewhat displaced from the zero transition, the amount of the displacement being systematically different in different subjects. It would be very interesting to find out whether the transitions utilized in consonant perception are represented by different complex events (for instance, the transitions which occur not later than some critical interval from the onset marker) or whether they are the same as the transitions differentiating vowels.

In conclusion I would like to present one topic for discussion. We (our group) believe that the only way to describe human speech perception is to describe not the perception itself but the artificial speech understanding system which is most compatible with the experimental data obtained in speech perception research. The main point is that artificial systems are based on many sources of scientific information, speech perception data being only one of these sources. If our point of view is accepted (I doubt that speech psychologists will agree with us), then it will be practical to direct experimental research to those problems which arise in automatic speech processing research.



Let us discuss some problems in automatic "phonetic processing"; they are most relevant to this meeting. The main problems concern the input parameters (representation of the signal at the input of the processor), the output representation and the rules and procedures of transformation. To specify the output one has to decide what kind of inventory (phonemes, allophones or something else) to use and how to represent the prosodic information. These problems are especially important from the point of view of simulating the higher levels of processing. Fortunately, research in this field does not really depend on exact knowledge of the lower levels of processing. In the case of the identification rules the situation is basically different. The rules are bound to depend strictly on the form of the signal representation: if you change the parameters extracted from the signal you must change the identification rules. It would be a good strategy to concentrate effort on the problems of auditory processing and on constructing automatic systems to simulate this processing. With these systems in hand it would be possible to approach the problem of rules by using both speech perception methods and the statistical methods applied in automatic speech recognition research.

References

Avakjan, R.V. (1976): "Study of the perception of isolated vowels by Russian language users', Fiziologia cheloveka 2, 81-90.

Bedrov, Ja. A., L.A. Chistovich and R.L. Sheikin (1978): "Frequency location of the "center of gravity" of the formants as the useful parameter in vowel perception", Akust. Zh. 24, 480-486.

Chistovich, I.A. (1978): "A functional model for envelope processing in the frequency channel of the auditory system", Fiziologia cheloveka 4, 208-212.

Chistovich, L.A., A.V. Ventsov, M.P. Granstrem, S.Ja. Zhukov, M.G. Zhukova, E.K. Karnickaja, V.A. Kozhevnikov, D.M. Lissenko, V.V. Lublinskaja, V.N. Mushnikov, N.A. Slepokurova, N.A. Fedorova, R.Haavel, I.A. Chistovich and V.S. Shupljakov (1976): Physiology of speech. Speech perception, Leningrad: Nauka.

Chistovich, L.A., R.L. Sheikin and V.V. Lublinskaja (in press): ""Centers of gravity" and spectral peaks as the determinants of vowel quality", in Frontiers of Speech Communication Research, B. Lindblom and S. Öhman (eds.), London: Academic Press.

Delattre, P., A.M. Liberman, F.S. Cooper and Gerstman (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.

Goloveshkin, V.T., V.S. Shupljakov, L. Bastet and J.M. Dolmazon (1978): "Functional model of auditory spectral analyzer (linear version)", in Avtomaticheskoe raspoznavanie slukhovykh obrazov 10, 23-24, Tbilisi.

Kozhevnikova, E.V. (1978): "Natural categorization of stimuli with different delay of amplitude jump", Fiziol. zh. SSSR 64, 1843-1849.

Kozhevnikov, V.A., V.D. Rodionov, E.I. Stoljarova and I.A. Chistovich (1978): "Study and modelling of the auditory extraction of amplitude irregularities", in Avtomaticheskoe raspoznavanie slukhovykh obrazov 10, 37-38, Tbilisi.

Kuznetsov, V.B. (1978): "Phonetic interpretation of steady-state vowels", in Avtomaticheskoe raspoznavanie slukhovykh obrazov 10, 97-98, Tbilisi.

Lesogor, L.V. (1977): "Measurement of the admittable mistuning in r-gap perception", Fiziologia cheloveka 3, 85-87.

Lesogor, L.V., V.D. Rodionov and L.A. Chistovich (1978): "Masking of the irregularity on the tone envelope by a tone of different frequency", Akust. zh. 24, 563-568.

Lesogor, L.V. and L.A. Chistovich (1978): "Detection of consonant in two-component complex sounds and interpretation of stimulus as a sequence of elements", Fiziologia cheloveka 4, 213-219.

Lublinskaja, V.V. and N.A. Slepokurova (1977): "Perception of vowel-like sounds with time-varying spectra", Fiziologia cheloveka 3, 77-84.

Lublinskaja, V.V. and N.A. Slepokurova (1978): "Perception of synthetic vowels with initial and final second formant transition", in Avtomaticheskoe raspoznavanie slukhovykh obrazov 10, 99-100, Tbilisi.

Mushnikov, V.N. and L.A. Chistovich (1971): "Auditory representation of the vowel. II. Detection of the second formant in the synthetic vowel", in Analiz rechevykh signalov chelovekom 11-19, Leningrad: Nauka.

Ogorodnikova, E.A. (1978): "Investigation of 'selective adaptation' effect in simple nonspeech perception", Fiziol. zh. SSSR 64, 1831-1835.

Rodionov, V.D., P. Carre and V.A. Kozhevnikov (1976): "Combining the information on short changes of the envelopes of the signals in different channels of the auditory system", Fiziologia cheloveka 2, 1021-1027.

Stoljarova, E.I. and I.A. Chistovich (1977): "Frequency response of the model for auditory processing of envelope and the output threshold devices", Fiziologia cheloveka 3, 72-76.



Zhukov, S.Ja. and D.M. Lissenko (1974): "Segmentation markers and their role in the interpretation of the formant contour", in Avtomaticheskoe raspoznavanie slukhovykh obrazov 8, Lvov.

Zhukov, S.Ja., M.G. Zhukova and L.A. Chistovich (1974): "Some new concepts in the auditory analysis of acoustic flow", Akust. zh. 20, 386-392.