INTERVAL OF SPECTRAL INFORMATION ACCUMULATION IN PERCEPTION OF NON -STATIONARY VOWELS

INNA CHISTOVICH TAISA MALINNIKOVA ELENA OGORODNIKOVA

Lab.of Speech Physiology, Pavlov Institute of Physiology
Leningrad, USSR, 199164

ABSTRACT

The results of identification experiments indicate that the interval of the auditory spectrum accumulation exceeds 20 ms. The data is compatible with the supposition that the accumulation interval is comparable to the duration of the vowel.

INTRODUCTION

This work is a development of the study of the spectrum shape processing started by L.Chistovich. She suggested a new approach to this problem which allowed to demonstrate that the information about spectrum shape was accumulated over the vowel length, but the data concerning the accumulation mechanism was rather contradictory (see /1/ for a review).

The fact of accumulation can be explained by either one of the following hypotheses:

1. The running auditory spectrum is considerably smoothed in time before extraction of the phonetically relevant parameters.
2. The parameters characterizing spectrum shape are extracted from practically unsmoothed auditory spectrum and then are accumulated. It is evident that in this case the extracted parameters depend strongly on the sampling instants.

The choice between these hypotheses will influence the direction of future studies. If hypothesis 2 is correct, it probably means that the sampling is synchronized to the fundamental tone, and it is necessary to investigate the synchronization mechanism. If hypothesis 1 is correct, the sampling with constant interval or the sampling at the ends of the segments (synchronized to segmentation marks) is to be considered.

We discuss here the previously obtained data /2,3/ and present new experiments designed to test these hypotheses (some of the experiments were suggested by L.Chistovich).

In all the experiments discussed here the same type of the signals, specially designed to have no dynamic cues (formant transitions), was used /4/. To simplify their description we shall introduce some designations.

The signal is a train of n formant pulses. One-formant pulse s_i is a short tonal pulse with triangular time envelope, F_i is the tone frequency, L_i is its intensity. $v_{ij}=s_i+s_j$ is a two-formant pulse, $w_{ijk}=s_i+s_j+s_k$ is a three-formant pulse. The stationary signals, consisting of identical pulses, are denoted S_i, V_{ij} and W_{ijk} respectively. Signals (S_iS_j) contain s_i and s_j; (S_iV_{jk}) contain s_i and v_{jk}; n_i is the number of s_i pulses in such signals. T_0 is the interval between the onsets of two identical pulses, T is the interval between the onsets of any two successive pulses, t is the interval between the onsets of s_i and s_j (or v_{jk}).

The results compatible with hypothesis 1 were obtained in several experiments /1, 3,4/. The most striking is the fact that increasing n_i in (S_iV_{ij}) causes the same changes in identification as increasing L_i in V_{ij} /3/. The main result against hypothesis 1 was obtained in the experiment on identification of S_i, V_{ij} and (S_iS_j) for $T \approx 10$ ms /2/. Signals (S_iS_j) were not identified with the same phonemes as V_{ij}. What is more, (S_iS_j) were mostly identified with either the same phonemes as S_i and S_j, or with [ɨ]. Obviously such result is possible only if hypothesis 2 is correct and the auditory spectrum is so little smoothed that a formant pulse does not affect the next pulse after 10 ms delay. The great number of [ɨ] responses was explained by the fact that Russian subjects often use [ɨ] as a label for indefinite vowels.

As this is the only experiment directly contradicting hypothesis 1, we tried to check its results in Experiments 1 and 2.

EXPERIMENT 1

In this experiment we obtained the identification data on signals S_i and (S_iS_j) for a wide range of F_iF_j. First, we wanted to check if (S_iS_j) would be identified as S_i, S_j or [ɨ] for other values of $F_i,F_j (F_i << F_j)$ than those used in /2/. Then, there were some indications in /2/ that the "local center of gravity effect" (LCGE) could be observed on (S_iS_j). LCGE manifests itself in the fact that a signal with formant frequencies F_1,F_2, $F_2-F_1 < 3\div4$ Bark, is phonetically similar to a one-formant signal with formant frequency F, $F_1 < F < F_2$ /1/. If a) hypothesis 2 is correct, and b) LCGE is a result of smoothing of the auditory spectrum in the frequency domain, LCGE should disappear when the formants are sufficiently separated in time.

Signals of Tests 1,2,3: n=12, T=20 ms or 14 ms, F_i=0.3,0.65,1.15,1.9,3.0 kHz. In (S_iS_j) j=i+2, n_j=4,6,8.

Signals of Test 4: n=8, T=20 ms, F_i=0.3, 0.45,0.65,0.85,1.15,1.5 kHz. In (S_iS_j) j=i+2, n_j=2,4,6.

The results of Tests 1,2,3 were combined, as no significant differences were found between the tests. The results of Tests 1,2,3 do not agree with /2/. All the 3 subjects responded to (S_iS_j) quite differently than to S_i and S_j. Subjects A and B practically always identified S_1, S_3 and S_5 as [u], [a], [i], (the corresponding response rates for 90 trials are 1., 1., 0.99 for A; 1., 0.96, 1., for B). Maximal (for 3 values of n_3) rate of (neither [u] nor [a]) responses to (S_1S_3) is 0.43 for A, 0.62 for B. Maximal rate of (neither [a] nor [i]) responses to (S_3S_5) is 0.87 for A, 0.46 for B. Only subject C frequently identified (S_iS_j) with [ɨ] ; A and B practically never used this phoneme.

In respect of LCGE the results were qualitatively the same as for stationary signals. LCGE was observed in Test 4, where $F_j-F_i \approx 3\div3.5$ Bark: (S_iS_j) were perceived as similar to S_{i+1}. The square distance between the response distributions served as a measure of similarity. In 8 cases out of 12 (3 subjects x 4 F_i, F_j combinations) at least one of three (S_iS_j) with n_j=2,4,6 was nearer to S_{i+1} than to S_i or S_j. In the 4 remaining cases the distances from (S_iS_j) to S_{i+1} and to S_i or S_j were approximately equal (and small). Thus, LCGE does not disappear when the formants are separated in time.

EXPERIMENT 2

In this experiment we tried, using the same F_1,F_2 combination as in /2/, to find the minimal time lag t at which (S_1S_2) begins to be perceived as a mixture of S_1 and S_2 and not as V_{12}.

Signals of Test 1: F_1=0.75 kHz. For S_1,

S_2, V_{12} n=6, T_0=20 ms. For (S_1S_2) $n_1=n_2=6$, $T_0=|t|+20$ ms, $t=\pm5, \pm10, \pm15, \pm20$ ms. We found that for all t values the (S_1S_2) response distribution is not a mixture of responses to S_1 and S_2. All the 5 subjects identified S_2 with [i] (response rate $p_{[i]} \geqslant 0.93$); for all $(S_1 S_2)$ $p_{[i]} \leqslant 0.125$. 3 subjects identify S_1 with [o] ($p_{[o]} \geqslant 0.95$) and never use [o] in responses to (S_1S_2). Only E.Z. gave a lot of [ɨ] responses to (S_1S_2), but she also responded to V_{12} with $p_{[ɨ]}=0.5$. Other subjects had $p_{[ɨ]} \leqslant 0.125$ for all signals. The responses of two subjects were almost independent of t: $p_{[ɛ]}$ fluctuated from 0.58 to 0.87 for $|t| = 0 \div 20$ ms. Others exhibited a strong dependence of identification on t. Increase of $|t|$ increased $p_{[e]}$ and decreased $p_{[ɛ]}$ for S.Zh; increased $p_{[ɨ]}$ and decreased $p_{[ɛ]}$ for E.Z; T.M. changed responses from [ɑ] to [ɛ] and then to [æ]. Thus, the results of one subject (E.Z.) only are similar to those obtained in /2/. The dependence of identification on t is, we suppose, really the dependence on duration or/and pitch, which were not constant. The results of Test 2 support this supposition. Signals of Test 2: F_1=0.75 kHz, F_2=2.5 kHz, T_0=16 ms, $n_1=n_2=12$, t=0, ±4, +8 ms. Four of the subjects of Test 1 took part in Test 2. The table shows the variation of $p_{[ɛ]}$ when t was varied from −5 ms to +5 ms in test 1 and from −4 ms to +8 ms in test 2.

	T.M.	S.Zh.	E.Z.	I.Ch.
Test 1	0.37	0.38	0.2	0.23
Test 2	0.2	0.17	0.07	0.1

As can be seen, though the t range for Test 2 is larger, variation of $p_{[ɛ]}$ is always smaller when duration and T_0 of signals are kept constant.

EXPERIMENT 3

The goal of this experiment was to find out if (S_2V_{13}) could be identified with the same phonemes as W_{123}, and if varying n_2 in (S_2V_{13}) would lead to the same changes in identification as varying L_2 in W_{123}. It is only possible if hypothesis 1 is correct and the auditory spectrum is integrated over several formant pulses. Such an effect was observed for V_{ij} and (S_iV_{ij}) /3/. As (S_2V_{13}) contain no three-formant pulses, the equivalence of varying n_2 and L_2 would be even a stronger argument for hypothesis 1 than /3/.
Signals: F_1=0.3 kHz, F_2=1.1 kHz, F_3=3 kHz, n=12, T=14 ms. For W_{123} $L_1=L_3$, $\Delta L=L_2-L_1=\pm20, \pm10, 0$ dB. For (S_2V_{13}) $L_1=L_2=L_3$, n_2=3,6,9.
The responses to W_{123} strongly depended on ΔL. When ΔL decreased from $+\infty$ (S_2) to $-\infty$ (V_{13}) the obtained sequences of most probable responses were [ɑɛi] for T.M., [ɑɛɨ] for E.Z., [ɑɛɨi] for E.K. and I.Ch., [ɑɛɨu] for S.Zh. All the subjects identified (S_2V_{13}) with the same phonemes as W_{123}, and increasing n_2 in (S_2V_{13}) had the same effect on the identification as increasing ΔL in W_{123}.
To evaluate this effect quantitatively we approximated the (S_2V_{13}) response distribution P_n by the weighted sum of two (closest to P_n) W_{123} response distributions: $P_n=k_1P_1+k_2P_2$. The obtained k_1, k_2 and residual error d^2 are shown in the table. Indices of k indicate ΔL of corresponding W_{123}.
It can be seen from the table that d^2 are quite small. Increasing n_2 is equivalent to increasing ΔL, but ΔL range corresponding to variation of n_2 from 3 to 9 is different for different subjects (from $0 \div 10$ dB for I.Ch. to $-10 \div 20$ dB for T.M.). Thus, all the 3 experiments are compatible with hypothesis 1 and contradict hypothesis 2. The duration of

the time window used for smoothing of the running auditory spectrum should, according to Experiment 2, exceed 20 ms.

	n_2	k_{+20}	k_{+10}	k_0	k_{-10}	d^2
E.K.	9	0.09	0.98			0.004
	6		0.44	0.69		0.042
	3			0.76	0.33	0.042
E.Z.	9	0.07	0.88			0.003
	6		0.17	0.90		0.016
	3			0.61	0.37	0.001
T.M.	9	0.96	0.02			0.001
	6		0.29	0.88		0.086
	3			0.12	0.92	0.024
I.Ch.	9		1.			0.091
	6		0.52	0.60		0.094
	3		0.04	0.97		0.016
S.Zh.	9	0.09	0.85			0.008
	6			1.		0.082
	3			0.67	0.47	0.054

The results of Experiment 3 corroborate the data of /3/ and suggest the duration of time window comparable to the duration of the signal. If this is the case, some sort of amplitude compression or normalization must precede the smoothing, as the identification of (S_iV_{ij}) very weakly depends on the amplitude of V_{ij} pulses /3/. All our results concern only the spectrum shape processing. The formant transitions are probably processed by the system with quite different temporal properties.

REFERENCES

/1/ Chistovich L.A. Central auditory processing of peripheral vowel spectra. – J.Acoust.Soc.Amer., 1985, v.77, pp.789–805.

/2/ Chistovich L.A., Ogorodnikova E.A. Temporal processing of spectral data in vowel perception. – Speech Communication, 1982, v.1, pp.45–54.

/3/ Chistovich L.A., Malinnikova T.G. Processing and accumulation of spectrum shape information over the vowel duration. – Speech Communication, 1984, v.3, pp.361–370.

/4/ Чистович Л.А., Чихман В.Н., Огородникова Е.А. Новый подход к определению фонетической близости стимулов и его проверка в автоматизированном эксперименте. – Физиол.журн. СССР, 1981, т.67, с.704–710.