Mimicking of synthetic vowels

Chistovich, L. and Fant, G. and de Serpa-Leitao, A. and Tjernlund, P.

journal: STL-QPSR
volume: 7
number: 2
year: 1966
pages: 001-018

http://www.speech.kth.se/qpsr
I. SPEECH PERCEPTION

A. MIMICKING OF SYNTHETIC VOVELS
L. Chistovich, G. Fant, A. de Serpa-Leitão, and P. Tjernlund

It was recently suggested \(^1\(^2\)\) that by studying the speech mimicking behavior one could gain some insight in the performance of the initial stages of speech perception without the complication of semantic and syntactic analyses of the message at higher levels of speech processing. In order to elicit a response vowel corresponding to a stimulus vowel the nervous system has to transform the auditory pattern of the stimulus vowel into a set of instructions to the centers, responsible for programming and control of the articulatory organs. Recently there appears to be a certain disagreement concerning the appropriate terminology for dealing with such instructions (distinctive features, motor parameters, and so on) and how closely they are associated with immediate muscle activity and proprioceptive feedback \(^3\(^4\(^5\)\). To avoid discussion we can call this set of instructions the "Intermediate representations" of the vowel, which can be stored and used as input signals both by the mechanism of speech production and by higher levels of speech message processing. These signals correspond to output signals of the subphonemic auditory pattern block in Fant's model \(^4\), to the \(V'\) representation in Stevens and House \(^6\), and to the \(v\) representation in Galonov and Chistovich \(^1\). It seems to us, at the present stage of research, that the physiological nature of these signals and their location in the brain are of secondary importance. More pertinent is the question concerning their logical structure and the rules of transformation from auditory patterns to this representation of signals.

Two different hypotheses concerning such rules of transformation of auditory patterns into instructions for vowel production seem to exist. One can be called the hypothesis of continuous motor representation, the other - the discrete or finite set of representation. Both hypotheses have been advanced by Liberman et al \(^3\) and Fry et al \(^7\) but unfortunately in a rather loose way. In order to facilitate the discussion of experimental methods for checking these hypotheses we have to reformulate them in a more explicit way.

For the sake of simplicity we shall restrict our consideration to isolated stationary vowels and we shall assume that each vowel could be specified by only a single time sample of vocal tract configuration. For a given person with specific anatomical constraints the vocal tract configuration of a vowel can be replaced by the corresponding F-pattern, which in turn can be represented as a point in a three-dimensional \(F_1-F_2-F_3\) space.
It is clear that the range of possible vowel sounds that can be physically produced by a given vocal tract is much more limited than the range of all possible sounds that can be produced by a formant synthesizer with independent control of n formants. That means that a subject mimicking vowels spoken by different speakers or mimicking synthetic vowels in any case reduced the information by transforming a larger inventory of signals into a smaller one. The difference between the two above-mentioned hypotheses concerns the question about how far this reduction of information goes.

Hypothesis of continuous representation

In spite of the physical constraints of vowel articulation a continuous representation provides an infinite number of vocal tract representations and an infinite corresponding number of F-patterns which can be thought of as points along trajectories in this space. No empty intervals occur along these trajectories if they are located within the $F_1$, $F_2$, $F_n$ range of the subjects' speech. The hypothesis of continuous representation is based on two assumptions: (1) There is a complete control of the vocal tract configuration. That means that the nervous system is able to work out the motor commands corresponding to any realizable vocal tract configuration providing the desired position of a vowel sound along a trajectory in the possible range of the $F_1$, $F_2$, $F_n$ space. (2) The numerical values of the motor instruction necessary to produce the given vocal tract configuration derive from the processing of the auditory pattern of the perceived vowel. How this transformation is effected is not clear, of course, but something like an analysis-by-synthesis function could be conceived. It is worth mentioning that the hypothesis of continuous representation is in fact accepted by a majority of phoneticians, who believe that by listening to a vowel they can determine an accurate position of the sound within a vowel triangle.

Hypothesis of discrete (finite set) representation

According to this hypothesis the nervous system "knows" only a finite set of motor instructions corresponding to a finite set of different vocal tract configurations and a corresponding set of discrete points along the trajectories in the $F_1$, $F_2$, $F_n$ space.

The logical consequence of this assumption is that the process of classification takes place at an early stage of transformation of auditory patterns into instructions for production. The problem of evaluating the
size and inner structure of this possible set of instructions immediately arises if this hypothesis is accepted. Experimental data suggest that the vocabulary of instructions has to exceed the vocabulary of phonemes \( (8) \). On the other hand, it is difficult to suppose that the vocabulary of instructions corresponds to the vocabulary of all the allophones observed in the ordinary fluent speech. Some allophones of one and the same phoneme need not originate from separate instructions. A difference in production could also originate from different realization of the same instruction depending on the context (coarticulation effects in a broad sense).

Let us now consider what results we have to expect from mimicking experiments in the cases of the first or second hypothesis being correct and what will be the best procedure for designing experiments that will allow an evaluation.

The useful feature of the mimicking experiment is the possibility of direct comparison of the response vowels with the signal stimulus vowels. Such a comparison is of special interest provided the signal vowels belong to the inventory of vowels which the subject is physiologically capable of producing. This implies that in mimicking experiments it is motivated to use a restricted number of subjects and to synthesize a set of stimuli signals especially adapted to each subject. The simplest procedure is to make detailed experiments on one subject only, and after this to check the conclusions on a group of subjects, using less intricate methods.

If the synthetic stimulus vowels conform with the properties of the subject’s vocal tract it is natural to expect that, in the case that the first hypothesis is correct, the response vowels have to coincide with the stimulus vowels. Or, more precisely stated, only random deviations of response vowels from the stimulus vowels are to be expected.

The results, which the second hypothesis predict, are more complicated and will depend on the relations between the variables involved in the experiment, including the behavior of the subject. In order to discuss these variables it is convenient to refer to the simple model presented in Fig. I-A-1.

Stimulus vowels \( S_1, S_2, \ldots, S_6 \) are shown as points on the x-scale which can be identified with the trajectory in \( F_1, F_2, \ldots, F_n \) space (see Fig. I-A-1. (a)). The auditory patterns of the vowel \( S_1 \) are regarded as a random variable \( A_1 \) with mean value \( \overline{A}_1 \) and standard deviation \( \sigma_{A_1} \) (see Fig. I-A-1. (b)). The auditory pattern is compared with
Fig. I-A-1. Models of categorization. Explanation in the text.
a number of boundaries $T_1$, $T_2$, $T_3$ (Fig. I-A-1. (c)). If $T_1 < A_i < T_2$ it is said that the signal vowel is identified with the category $C_1$ and the instruction $I_1$ is selected. In the following transform of the instruction into a vocal tract configuration motor errors are expected to occur and the response vowel $R_i$ elicited by the instruction $I_1$ is a random variable with mean value $\bar{R}_i$ and standard deviation $\sigma_{R_i}$ (see Fig. I-A-1).

It is clear from Fig. I-A-1 that no conclusions can be reached in case the number of stimulus vowels is limited to one per response category. Statistically significant deviations of response vowel from stimulus vowel are to be expected but this would not prove the second hypothesis. The most favorable experimental and behavioral conditions for a decisive proof is a large number of signals per response category and a small standard deviation of auditory patterns. In that case the response vowel trajectories in the $F_1$, $F_2$, $F_3$-space will appear as step approximations of corresponding stimulus vowel trajectories. The average value of formant frequencies of the response vowel elicited by different stimulus vowels will be constant as long as the stimulus vowels are inside a category region and will shift discretely when the stimuli attain values within an adjacent category. This effect cannot be fully developed unless the standard deviation $\sigma_{A_i}$ of auditory pattern processing is small enough to be neglected. In general it is thus necessary to assign to each input signal a probability of its belonging to different response categories.

As an example the stimulus $S_2$ in Fig. I-A-1 will be described by the probability distribution $P_{S_2}(I_0)$, $P_{S_2}(I_1)$, $P_{S_2}(I_2)$; $P_{S_2}(I_0) + P_{S_2}(I_1) + P_{S_2}(I_2) = 1.0$. This means that observing the responses elicited by vowel $S_2$ we shall deal with a number of samples taken from the separate response distributions $R_0$, $R_1$, $R_2$ (see Fig. I-A-1). It is clear, that stimulus $S_4$ is less likely than $S_2$ to evoke the instruction $I_0$, i.e., $P_{S_4}(I_0) < P_{S_2}(I_0)$ and that $S_4$ is more likely to evoke the instruction $I_3$. The average set of response data will naturally be different for $S_2$ and $S_4$. Accordingly, the expected step functions of categorization will be more or less obscured depending on the value of $\sigma_{A_i}$. According to the model of Fig. I-A-1 additional effects can be observed in case the categorization hypothesis is true. If $\sigma_{A_i}$ is small enough in comparison with the size of the perceptual categories the standard deviation of a response vowel will have a minimum for a signal in the middle of the category and will increase for signals located near the boundaries of the category. It is also clear that the distribution of the observed responses to a wide range of input signals should display definite peaks corresponding
to $\bar{R}_1$, $\bar{R}_2$, ..., $\bar{R}_n$. A requirement for the appearance of these two additional effects is a sufficiently low motor spread $\sigma_R$.

The experiment was planned so as to enable detection of possible categorization. Only one subject took part in the experiment. A set of OVE II synthetic vowels were selected to match the subject’s vowels as closely as possible. An additional aim of the experiment was to compare three kinds of response behavior: rapid shadowing of vowels, delayed mimicking, and letter writing. Letter writing (when the letters of the subject’s native language are used) has a more direct correspondence to phoneme identification than the oral response.

**Stimulus vowels**

12 synthetic vowels were selected along a pathway (trajectory) in the $F_1$ $F_2$ $F_3$-plane constructed to conform with a continuous articulation of the sequence a-e-i produced by the subject L. C. (female Russian). The synthesis was made with the OVE II system of the Speech Transmission Laboratory.

An attempt was made to match the subject’s voice source and constant vocal transfer characteristics as accurately as possible. By inspecting time-frequency-intensity spectrograms of the subject’s natural a, e, and a vowels the stable parts of these vowels were selected and analyzed by means of spectrum sections from the 51-channel analyzer. The spectrum envelope of each vowel was traced directly on the display screen of the analyzer. Alternatively, the synthesizer OVE II was connected to the analyzer and the parameters of OVE II were adjusted for a best match with the natural vowel. The values of formant bandwidths, higher pole correction, and characteristics of voice source gained by this method were then used as constant parameters in the synthesis of the 12 stimulus vowels.

The frequencies of $F_1$, $F_2$, $F_3$ of the stimulus vowels are shown in Table I-A-1. The criterion followed for the sampling was to retain a spacing of 55 mels in $F_1$ of successive members of the set.
The pitch contour typical for the subject's phonation of an isolated vowel was computed from a number of spectrograms and was included in the synthesis program. The duration of all stimuli was fixed at 300 msec. A tape-recording was made with the stimuli arranged in a quasi random order and to allow all possible sequential combinations of two successive stimuli. Out of the 12 stimuli all possible pairs were used including stimulus repetitions.

**Experimental procedure**

The subject was placed in an anechoic chamber and listened to the stimuli through earphones and performed (in turn) one of three different tasks. One was to mimic the vowel as accurately as possible after listening to the whole vowel. The second task was to shadow the vowel with minimal possible delay and the third task was to write an appropriate Russian letter corresponding to the vowel. The subject wrote the letters on an aluminium plate connected to the writing pen via a signal source and an ink-recorder (Mingograph) thus enabling a recording of the time of starting the writing in relation to the onset of

<table>
<thead>
<tr>
<th>F₁ mel</th>
<th>N</th>
<th>F₁ c/s</th>
<th>F₂ c/s</th>
<th>F₃ c/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>365</td>
<td>1</td>
<td>270</td>
<td>2500</td>
<td>3400</td>
</tr>
<tr>
<td>420</td>
<td>2</td>
<td>320</td>
<td>2450</td>
<td>3150</td>
</tr>
<tr>
<td>475</td>
<td>3</td>
<td>375</td>
<td>2420</td>
<td>3000</td>
</tr>
<tr>
<td>530</td>
<td>4</td>
<td>430</td>
<td>2350</td>
<td>2820</td>
</tr>
<tr>
<td>585</td>
<td>5</td>
<td>480</td>
<td>2250</td>
<td>2750</td>
</tr>
<tr>
<td>640</td>
<td>6</td>
<td>540</td>
<td>2140</td>
<td>2650</td>
</tr>
<tr>
<td>695</td>
<td>7</td>
<td>605</td>
<td>1950</td>
<td>2600</td>
</tr>
<tr>
<td>750</td>
<td>8</td>
<td>670</td>
<td>1750</td>
<td>2610</td>
</tr>
<tr>
<td>805</td>
<td>9</td>
<td>740</td>
<td>1570</td>
<td>2650</td>
</tr>
<tr>
<td>860</td>
<td>10</td>
<td>810</td>
<td>1470</td>
<td>2700</td>
</tr>
<tr>
<td>915</td>
<td>11</td>
<td>880</td>
<td>1420</td>
<td>2800</td>
</tr>
<tr>
<td>970</td>
<td>12</td>
<td>950</td>
<td>1400</td>
<td>2900</td>
</tr>
</tbody>
</table>
the stimulus vowel. These recordings accordingly showed the latencies of letter writing. The response vowels produced by the subject in the mimicking and shadowing tasks were recorded on tape and simultaneously on one of the channels of an ink-recorder. The recording of the stimuli in the second channel of the ink-recorder allowed the measuring of latencies of mimicking and shadowing.

The vowels produced by mimicking were analyzed with the 51-channel analyzer. Spectrum sections were taken close to the beginning of the last third of the vowel. These sections were produced with filters of 250 c/s width and 100 c/s spacing. Vowels produced in the shadowing experiment were measured from spectrograms. In instances when a pattern shift was observed during shadowing we selected the final part of the utterance.

Comparison of letter writing, shadowing, and mimicking behavior

Fig. I-A-2 shows the latencies of shadowing, mimicking, and letter writing. Each point corresponds to the average of 36 responses for mimicking and letter writing and to the average of 12 responses for shadowing.

It can be seen that latencies of shadowing and mimicking are independent of the stimulus values. The latencies of letter writing are maximal for stimuli $S_3$, $S_4$, $S_5$, $S_9$, and $S_{10}$. Fig. I-A-3, (a) shows that the stimuli $S_3$, $S_4$ and also $S_9$, $S_{10}$ are non unique with respect to letter selection. They can be regarded as stimuli located in the vicinity of phoneme boundaries.

Increase of the latency of button-pressing reaction (mediated by phoneme selection) at phoneme boundaries was described by Studdert-Kennedy, Liberman and Stevens (9). Increase of reaction time is usually observed in a situation of forced choice experiment when the subject has to select one of alternative reactions and does not receive sufficiently typical signals. It is safe to say that the increase of latencies at the boundaries of categories is a common characteristic of learned classification behavior, for discussion see (5)(10). Absence of similar latency effects in shadowing and mimicking implies that either no classification is included in the particular signal processing or that some more direct classification takes place.
Fig. I-A-2. Latencies of letter writing (a), mimicking (b), rapid shadowing (c).
The very short absolute values of shadowing latencies also point to a primitive reaction. The latencies of 190 msec, obtained in this experiment, are in agreement with data described before (11)(8). The most recent results (12) suggest that the movements, corresponding to vowel shadowing start in fact about 40-50 msec ahead of the beginning of voicing. The effective articulatory reaction time is thus of the order of 140-150 msec, which is about two times less than the established choice reaction time to sounds in the simplest two-alternative situation. In any way it is reasonable to suppose that the processing, responsible for shadowing and mimicking, is not the same as that for letter selection and that the inventory of categories (if categorization takes place in shadowing) can be different in both cases.

According to the letter writing experiment the stimuli vowels form three quite distinct groups*. Group \( /\text{u}/ \) comprises stimuli \( S_1\)-\( S_3 \), Group \( /\text{3}/ \) stimuli \( S_4\)-\( S_9 \), and Group \( /\text{a}/ \) stimuli \( S_{10}\)-\( S_{12} \) (see Fig. I-A-3. (a)). It was of interest to see if the same grouping would be derived from the mimicking experiment.

We asked the same subject to listen to the vowels she had produced in the mimicking and shadowing experiments and to describe each response vowel by symbols. No restrictions for symbols having been made, the subject could use symbols from different languages and could also use a combination of two symbols if the vowel seemed to be intermediate between two phonemes. The results of transcribing mimicked vowels are presented in Fig. I-A-3. (b). It is clear that the grouping of signals in Fig. I-A-3. (b) is different from that in Fig. I-A-3. (a). In transcribing the response vowels, corresponding to \( S_3 \) the subject in some instances made use of the symbol \( /\text{e}/ \) but not when transcribing responses to other stimuli. That means that at least some part of the responses to \( S_3 \) was mimicked in a different way than the responses to \( S_1\), \( S_2 \) or \( S_4\)-\( S_{12} \).

According to the outcome of the mimicking experiment stimuli \( S_4\)-\( S_9 \) form two separate groups (\( S_4\)-\( S_6 \) and \( S_7\)-\( S_9 \)) and not only one, as it was obtained in the letter writing experiment.

Mimicking of \( S_{10} \) is also different from mimicking of \( S_{11} \) and \( S_{12} \). The grouping of stimuli observed from analyzing the results of the shadowing experiment (see Fig. I-A-c. (c)) is rather similar to that of

* The Russian letter \( \text{u} \) corresponds to \( i\)-vowels, the \( \text{3} \) to \( e\)-vowels, and the \( \text{a} \) to \( a\)-vowels.
the mimicking experiment. Thus we can conclude that at least 6 groups of signals have to be taken into account in the analysis of the mimicking data and that the subject used at least 6 different motor instructions for mimicking the 12 synthetic vowels. It is possible too that the number of different motor instructions is much higher (the hypothesis of continuous representation is true) and the grouping of the stimuli in Fig. I-A-3, (b) and Fig. I-A-3, (c) is a result of categorization not at the level of the mimicking or shadowing by themselves but at the level of the symbolic transcription of the responses.

Spectral data of delayed mimicking responses

More direct data for studying possible categorization effects are available from the spectral data of the subject's responses. According to the categorization model discussed in the introduction three effects can be expected. (1) The average values of $F_1$, $F_2$, $F_3$ of the response vowels should follow the stimulus values of $F_1$, $F_2$, $F_3$ by some kind of step-like quantizing function. (2) The standard deviation of the formant frequencies of the response vowels should be minimal at the center of the category defined by the step and should increase towards its beginning and end. (3) The distribution of formant frequencies of the composite data has to show pronounced peaks.

In search of such effects we first summarized the data obtained in the three tests on mimicking. But it was found that the standard deviation was too high and the reason for that appeared to be the systematic shift in formant frequencies of the vowels produced in the first test (session) in relation to those produced in the second and the third tests (sessions). Accordingly the average $F_1$, $F_2$, $F_3$ data were computed separately for the first and second plus third tests. In computing the standard deviations and formant frequency distributions only the data of the second + third sessions have been used.

The results presented in Fig. I-A-4, Fig. I-A-5, and Fig. I-A-6 show the three effects we were looking for. There appears to be four main quantal steps in Fig. I-A-4 and the distribution of $F_2$ accordingly shows four main peaks. Thus the data seem to be in favor of the categorization hypothesis.

More complicated is the problem concerning the exact number of perception categories and the corresponding motor instructions used by the subject in this test. Because the average values of $F_1$, $F_2$, $F_3$
Fig. I-A-4. F₁, F₂, F₃ data of stimuli vowels (broken lines). Average values of F₁, F₂, F₃ of response vowels obtained in the first test (crosses) and in the second and third tests (points) of mimicking.
Fig. I-A-5. Standard deviation of $F_1$, $F_2$ values of response vowels obtained in the second and third tests of mimicking.
Fig. 1-A-6. Distribution of $F_1$, $F_2$, $F_3$ of response vowels obtained in the second and third tests of mimicking.
of response vowels corresponding to $S_4$, $S_5$, $S_6$ are almost the same for all these three stimuli we are able to infer that all these signals were perceived as belonging to the same category and that the same motor instruction was selected in response to all of them. The same can be said about stimulus groups $(S_1, S_2)$, $(S_7, S_8, S_9)$, and $(S_{11}, S_{12})$. This reduces the maximum number of possible categories to 6. Because the average response values for stimulus groups: $(S_1, S_2)$, $(S_4, S_5, S_6)$, $(S_7, S_8, S_9)$, and $(S_{11}, S_{12})$ are clearly different we can infer that the number of categories and corresponding different motor instructions used in mimicking of 12 synthetic vowels is not less than 4.

The analysis of the mimicking response to stimuli $S_3$ and $S_{10}$ is not clear enough. An explanation close at hand would be that $S_3$ is sometimes perceived and mimicked in the same way as $(S_1, S_2)$ and some times in the same way as $(S_4, S_5, S_6)$. But the results of transcribing the $S_3$ responses (see Fig. I-A-3.(b)) contradict to this explanation. Also the spectral data indicate that a number of points belonging to the $S_3$ response distribution lies clearly outside the distribution area of $(S_1, S_2)$ and $(S_4, S_5, S_6)$ responses.

Thus some other explanation is needed. It is possible that some competing categories located in the vicinity $S_3$ and $S_{10}$ are involved in processing of these signals. This explanation seems probable, at least, in the case of $S_3$ responses. Spectral analysis of the subject's /v/-vowel shows that it does not differ very much from some of the $S_3$ responses.

A more subtle explanation is that $S_3$ and $S_{10}$ are processed in a quite different way than other signals. The subject does not associate them with a single category but at once with a number of categories. Then he tries to produce a sound which is intermediate between the associated vowels. This explanation requires more complex and higher order structure of relations between the auditory and the articulatory representation spaces than it was accounted for by the above-mentioned model.

If this explanation is accepted then the responses to $S_{10}$ may be described without assuming an additional category. Anyway, the mimicking data suggest from five to six perception categories involved in the processing of the 12 stimuli.
Comparison of shadowing responses with delayed mimicking responses

Fig. I-A-7 contains the average data of $F_1$, $F_2$, $F_3$ for mimicking and shadowing responses. The uncorrected data for shadowing first to be discussed are presented by dots. A clear difference between shadowing and mimicking can be seen in the range of stimulus vowels from $S_4$ to $S_9$. The shadowing data of $F_2$ and $F_3$ in this range can be approximated by a continuous function and not necessarily by a step function. The range of variation of $F_2$ and $F_3$ from $S_4$ to $S_9$ is less for shadowing than for delayed mimicking. The question thus arises whether this difference indicates a difference in the mode of human transformation or whether the same model is applicable in both cases, though with different parameters of the model. Since the later hypothesis seems more probable we tested it first. The model described in the introduction was accepted with the specific assumptions that in the range from $S_4$ to $S_9$ only two categories ($I_a$ and $I_b$) exist and that the subject uses two different motor instructions only.

According to the model of Fig. I-A-1 we can propose two different explanations for reduction of the response vowel range corresponding to the stimulus ranges $S_4$ to $S_9$. One is the increase in the standard deviation of the auditory pattern ($\sigma_A$). If $\sigma_A$ is large both $S_4$ and $S_9$ will elicit $I_a$ and $I_b$ instructions, the difference between probabilities $P_{S_4}(I_a)$ and $P_{S_9}(I_a)$ decreasing with the increase of $\sigma_A$.

Another interpretation concerns the mode of transformation from the instruction domain, $I$, to the motor response, $R$. It is evident that many variables are involved in speech motor behavior and that one and the same instruction can result in different responses depending on the inner state of the motor centers, the time interval which is available to organize the movements, and so on. There is direct experimental evidence that motor coordination of vowel production in shadowing is different from that in delayed mimicking (12).

If the only source of difference in the average $F_2$, $F_3$ data for mimicking and shadowing is the difference in $S_A$, then the distribution of response vowels in the $F_2$, $F_3$-plane will display the same average trend for mimicking and shadowing. Fig. I-A-8 shows that this is not the case. The location of shadowing responses in the $F_2$, $F_3$-plane (Fig. I-A-8. (b)) is clearly different from that of mimicking responses (Fig. I-A-8. (a)). Fig. I-A-7 indicates the difference in $F_1$ location too. We can thus conclude that vowel production is not the same in both cases. The data of Fig. I-A-7 indicate that in shadowing a more
Fig. I-A-7. Average values of $F_1$, $F_2$, $F_3$ of response vowels.
Solid lines - data from the second and third tests of mimicking.
Points - uncorrected data of shadowing.
Crosses - corrected data of shadowing.
Fig. 1-A-8. Distribution of response vowels corresponding to stimuli 4-6 (points) and stimuli 7-9 (crosses) in the F2-F3 plane. Data from the second test of mimicking (a) and shadowing (b).
centralized articulation is used as judged from a lower $F_1$ and $F_2$.

Fig. I-A-8. (a) (mimicking responses) shows no doubt two distributions of response vowels, one corresponding to $S_4$, $S_5$, $S_6$ (points) and the other to $S_7$, $S_8$, $S_9$ (crosses). In the case of shadowing (Fig. I-A-8. (b)) crosses and points overlap and the existence of two response distributions is not so evident. To facilitate the visualization of the data we introduced a new x-coordinate and plotted all the vowels on this normalized and in equal steps quantified axis.

The results are presented in Fig. I-A-9. They show that also in the case of shadowing two response distributions exist although the separation between them is less than in the case of mimicking. It can be seen that the overlapping of the two response distributions is not large enough to explain the overlapping of crosses and points in Fig. I-A-8. (b). It is also evident that the overlapping of two response distributions cannot change the step function into a smooth one (Fig. I-A-7). The only possible reason for such a transformation, consistent with the model in consideration, is the increase of $\sigma_A$, (resulting in errors in categorization).

If this explanation is true, two effects have to be observed: (1) Points located to the left of the T-axis of Fig. I-A-8, (b) (in the area of the $R_b$ distribution) have to belong mostly to $S_6$ which is closest to the boundary of the category. Crosses, located to the right of the T-axis in Fig. I-A-8, (b) (in the area of the $R_a$ distribution) have to belong mostly to $S_7$ for the same reason. The numbers close to the points and crosses in Fig. I-A-8, (b) indicate the signals provoking this response. It can be seen that 5 out of 8 points correspond to $S_6$ and 3 out of 5 crosses correspond to $S_7$. (2) After excluding from consideration the above-mentioned $R_b$ responses to $S_4$, $S_5$, $S_6$ stimuli and the $R_a$ responses to $S_7$, $S_8$, $S_9$ stimuli and computing the new corrected values of average response vowels to the $S_4$-$S_9$ stimuli we have to obtain the same clear step function dependence as in the mimicking experiment.

The crosses in Fig. I-A-7 represent these new corrected values. The existence of steps is quite evident now. Thus we have to conclude that the "noise of perception" ($\sigma_A$) is higher in shadowing than in mimicking. This seems quite natural because in mimicking the subject can collect the information distributed along all the signal, whereas in shadowing a segment of the length of 70-100 msec only is
Fig. I-A-9. Distribution of response vowels corresponding to stimuli numbers 4, 5, 6, 7, 8, 9 on the x axis in the $F_2$-$F_3$ plane. Converted data of Fig. I-A-8. (a) mimicking, (b) shadowing.
used for response selection, for discussion see (4).

It is known from psychoacoustical data that the difference limens of the frequency decreases with an increase of signal duration up to 150-250 msec which can be treated as a sharpening of the auditory pattern response. A similar explanation can be applied to our case. Another explanation is that in perceiving the speech sound the subject produces a sequence of decisions about this sound and this is his way of collecting the information. In delayed mimicking she may be able to use the whole sequence of decisions to program one response vowel. In shadowing she is only able to realize the individual decisions in sequence. In five spectrograms there appeared quite clear and abrupt pattern changes from one shadowed vowel to another in the course of the vowel observed which proves that the subject is able to make successive decisions in the course of perception of one vowel. A similar effect has been observed in shadowing of consonants (15).

Phonetical interpretation

It is conceivable that the perception categories underlying the subject's response in mimicking and shadowing are identifiable with or include the subject's allophones and vowel qualities learned from other languages. It is accordingly of interest to study the subject's responses in a F1 F2-diagram together with F1 F2-data of such reference sounds. This is brought out in Fig. I-A-10 which for comparison includes the vowel diagram for Swedish female subjects in the early study of Fant (16).

It was found that the data of Swedish female subjects match the response data of the subject's [i] and [e] and [a] closely. The series of stimulus vowels 1, 2, 3, ..., 12 traverses the F1-F2 plane from [i] to [a] along the outer contour of the vowel figure from [i] to [e] and then takes a more centralized path inside the vowel figure down to [a]. This particular path was initially chosen from the sampling of an articulatory gesture [i - e - a] which in the [e-a] region comes closer to a neutral mid vowel than to the subject's [a].

The response group I is close to stimuli 1 and 2 and represents the vowel [i]. The response group II has a mean value close to the corresponding stimulus 3 but the actual distribution of group II is bimodal with a majority of points in the [i] region and a smaller but significantly different part in the [e] region (the velarized i). Response group III, associated with the stimuli 4, 5, and 6, actually lies below
Fig. I-A-10. Small phonetic signs: Swedish female data, Fant 1959. Larger phonetic signs: Vowels of subject L. C. 1, 2, 3, ..., 12: Signal vowels. I, II, ..., VI: Mean responses of six categories from mimicking experiment.
this region at a point of higher $F_2$ and lower $F_2$ which is close to the subject's Russian allophone $[e]$. The average response of category IV is close to the average of the corresponding stimuli 7, 8, and 9 and is not easily given a linguistic interpretation. It represents a kind of centralized or rounded $[\approx]$. The group V, i.e. responses to stimulus 10, appears to be associated with a centralized $[a]$, by the subject described as between $[a]$ and a fronted $[o]$. Group VI, finally, represents the subject's $[a]$.

The response data from the shadowing experiment are close to those of the mimicking experiment with a tendency for a more centralized articulation in groups III and IV as evidence by a lower $F_1$ in the two groups and a lower $F_2$ in group III, (see Fig. 1-A-7), which can be interpreted as a variation towards a neutral vowel.

Identification test

In the previous experiments the subject could use only a very short storage time. The longest latency in mimicking was about 1 sec. The question now arises how the subject performs when she has to store a vowel for a considerable time and more specifically what kind of representation is used for storage.

From the following experiment we hoped to find evidence concerning the particular form in which a stimulus vowel can be stored for a long time.

A kind of identification test was used for this purpose. One of the twelve stimulus vowels was presented twice to the subject in the beginning of the test. The subject had to remember this stimulus, which we shall call standard, and then, listening to a sequence of 144 vowels, decide about each of them whether it was the same as the standard or whether it was different. In the first case the subject had to push a button twice, in the second case to push the button once. The stimulus vowels and the push button response were recorded on the Mingograph ink-recorder.

These identification tests were repeated twelve times with twelve different stimulus vowels serving as the reference. Only one subject, the same as in the first experiment, was involved.

We shall first discuss what kind of results could be expected in the case different forms of representation were used for storage.
Different alternatives concerning the representation used for storage can be suggested. (1) An auditory pattern of the stimuli more or less conforming with the acoustic structure. (2) A symbol or a set of numbers indicating the perception category to which the auditory pattern was found to belong. The motor instruction associated with a specific category may serve this purpose. Phonetically such categories and motor instructions could be identical with the subject's allophones. (3) A symbol or a set of numbers indicating one of the phonemes of the native language of the subject.

If the first hypothesis is true the identification functions and latency functions are to be different for all the standards. The degree of similarity between functions, corresponding to different standards, are to be simply related to the distance between the standards in the acoustical space. If the second or third hypothesis is true the identification and latency functions, corresponding to references belonging to the same category, are not to differ significantly. This means that the 12 identification functions and the 12 latency functions will form a number of groups which would be equal to six or to five in the case of the second hypothesis and is equal to three in order to match the third hypothesis in accordance with the outcome of the letter writing experiment.

The results are presented in Fig. I-A-11,(a) and Fig. I-A-11,(b). It is quite clear that the identification functions and latency functions form six groups, which conform closely with those previously observed in the formant data of mimicking responses to the stimulus vowel and the phonetic transcript of those responses. The identification test thus suggest the ordering: $g_2 \leq S_3, g_3 \leq S_4, S_5, S_6, g_4 \leq S_7, S_8, S_9, g_5 \leq S_{10}, g_6 \leq S_{11}, S_{12}$. Thus we can conclude that the structuring of the subject's memory storage comes closest to the second of the three above-mentioned alternatives.

**Conclusion**

The aim of the present study was to collect experimental data enabling an evaluation of two different hypotheses concerning the nature of the transformation responsible for the mimicking of a vowel. According to one of them the transformation is continuous, according to the other discontinuous - a categorization is involved in the transformation.
Fig. I-A-11. Identification and latency in a test where a subject gives same/different response to each of the 12 vowel signals with one of these selected as a reference.
IDENTIFICATION LATENCY

LATENCY

Continuation Fig. 1-A-11.
Detailed analysis of the $F_1$, $F_2$, $F_3$ data of mimicking responses of one subject to twelve synthetic vowels, located on a trajectory in the $F_1$-$F_2$-$F_3$-space favored the second hypothesis. Additional data from other experiments and with other subjects (to be published later) have supported this conclusion.

An important fact is that the same categorization is involved also in rapid shadowing which is indubitably an automatic or highly automatized reaction with extremely short latencies.

Another important fact was that the categories used by the subject in shadowing and mimicking response were different from the categories of the written response using letters of the native language to represent the perceived vowels. Another difference between writing behavior and shadowing (and mimicking) behavior was that in the first case an increase of latencies was observed near the boundaries of the categories, but there was no such increase in the second case. The absolute values of letter writing latencies were much longer than the latencies of shadowing but shorter than for mimicking. One can speculate that the phoneme selection, included in letter writing, is a result of processing of information at a higher level of perception, where the conscious weighting of different alternatives with a resultant hesitation occurs.

No precise information about the number of perception categories used in shadowing (mimicking) was obtained, but it was clear from the data that this number exceeds the number of phonemes of the native language of the subject. The question as to what extent these categories correspond to allophones is still open and has to be investigated further but there is in evidence that some categories are identical with the subject’s allophones. It is clear that an inventory of allophones rather than phonemes is to be preferred in automatic recognition work. A string of allophones can in theory be used as an input signal of the block, performing the word selection and syntactic analysis of the message. A necessary requirement is of course that the recognition scheme includes means of long time allophone storage.

A special experiment was performed to investigate the form in which the subject is able to store perceived vowels. The results indicate that she stores the same categories, as observed in the shadowing and mimicking experiments, and not the phonemes or the immediate auditory patterns.
An interesting question arises: What code is used to store these categories? Because to each category there corresponds a mimicking response, the response parameters can serve as a dimension of the stored event in conformity with a motor theory of speech perception. On the other hand, these categories may be stored in the form of quite abstract symbols.

More interesting is that categorization takes place and the next object of investigation should be to determine the decision rules underlying categorization.

References:
(2) Chistovich, L. A. et al.: "Psychological methods in the speech perception research", Paper to be presented at the XVIIIth International Congress of Psychology, Moscow, August 4-11, 1966, Symposium: Models on Speech Perception; to be publ. in the Proc. from this congress.
(11) Saslow, M. G.: "Reaction time to consonant-vowel syllables in ensembles of various sizes", MIT, RLE, QPR No. 50, July 15, 1958, pp. 143-144.


