SPEECH DISCRIMINATION BY DYNAMIC PROGRAMMING

T. K. Vintsyuk

Kibernetika, Vol. 4, No. 1, pp. 81–88, 1968

UDC 51:681.14:155

In some papers [1, 2] on automatic speech discrimination the proposed methods include a time normalization of the words. The reliability of classification in these methods depends to a considerable extent on the time normalization rule. The essence of these rules is that immediately before attempting recognition the duration of the unknown word is equated to the duration of the standard words.

We shall describe the speech signal of a word by a sequence of readings of the spectral intensity vector x at the output of a spectrum analyzer. Then if we use a decision rule based on the minimum Euclidean distance from the normalized time description of the unknown word to the standards of the classes (words) it is desirable to take a time normalization such that during the comparison of the word with the standard of its class the spectral intensity readings in the vectors x should correspond to each other, i.e., during the comparison the corresponding readings should determine the same sounds. For example, in recognizing the pronunciation of the word "odin" (one) it is required that in comparing this word with the standard, the readings of x corresponding to the sound [i] together with its transitional segments are to be matched with the sound segment [i] and its transitional segments in the standard. For this reason one can speak about a greater or lesser correspondence between the readings of the spectral intensity vectors x under time normalization. Investigations show that under time normalization a correspondence between the readings of x and the standard of a class is achieved only on the average, and often considerable shifts of the readings with respect to each other can be observed.

In our proposed algorithm for the recognition of words the greatest possible match between the readings of x for the unknown signal and for the standard of its class is achieved, since a decision about the class is based on a directed search among the permissible standards, which is equivalent to a complete search over all possible standards (corresponding to all possible time normalization rules). The recognition of words is carried out through discrimination of the components of the word and is accomplished by the method of dynamic programming.

The algorithm makes partial use of the idea of analysis by synthesis, and it can also be explained on the basis of the principle of permissible transformations [3].

A MODEL FOR THE SPEECH SIGNALS OF WORDS. CONSTRUCTION OF WORDS FROM ELEMENTS

One possible direction in research into the automatic speech discrimination is based on the construction of a model for the signals of the various classes on the basis of the known properties of speech signals. From this model a recognition algorithm is derived, which is in some sense optimal. The quality of the recognition system realizing the algorithm is determined by the degree of correspondence between the model of the signals and the actual ensemble of signals. The hypotheses on which the model is based must not contradict the known properties of speech signals.

Learning in the recognition algorithm is reduced to the evaluation of the parameters of the model signals with respect to the learning sequence.

Below we shall give a model of the speech signal from which we derive the recognition algorithm. In this algorithm the problem of finding the standard of a class which is the nearest to the signal to be recognized according to some criterion, is solved by the method of dynamic programming.

We divide the speech signal of a word or its description into elements (parts). By an element we shall mean a spectral segment, i.e., a reading of the spectral intensity vector x at the output of the spectrum analyzer at some instant of time. Let the readings of the vector x be taken uniformly in time. Then the speech signal of a word is a sequence of elements X_l $(x_1, x_2, \ldots, x_l)$ where the number l identifies the duration of the word. We shall assume that the signal is quantized in such a way that the elements x_i reflect completely the possible variations in the spectral picture of the word, or that these variations can be reflected sufficiently accurately by a piecewise-linear approximation to the elements x_i.

Let us construct a model of all possible sequences X_l of the class k as follows.

To each word, whose probability of occurrence is P(k), we assign a family Ω^k of sequences $\mathfrak{Z}^k$, composed of the elements $\mathfrak{z}^k$. The sequences of elements $\mathfrak{Z}_l^k(\mathfrak{z}_1^k, \mathfrak{z}_2^k, \ldots, \mathfrak{z}_l^k)$ used in the process of pronouncing the word, which are ideal, and serve as references, we shall call the standard sequences of elements. The observed sequences $X_l(x_1, x_2, \ldots, x_l)$ will be different from the ideal ones as a consequence of individual properties of the speaker, the variations in the conditions of pronunciation, and so on.

We shall assume that the observed sequences X_l of length l can be derived only from a standard sequence $\mathfrak{Z}_l^k$ of the same length l. The sequences $\mathfrak{Z}_l^k$ of length l constitute a set of standards Ω_l^k.

The standard sequences $\mathfrak{Z}^k$, constituting Ω^k must be determined by the phonemic structure of the word and they must reflect the possible variations in the durations of the individual phonemes in the word.

We shall assume that all sequences $\mathfrak{Z}^k$ of a class can be obtained from one standard sequence $\mathfrak{Z}_q^k(\mathfrak{z}_1^k,$

$\vartheta_2^k, \ldots, \vartheta_q^k$) of length q by means of simple repetition or elimination of some elements ϑ_l^k, still preserving the ordering of the elements in the sequence.

This assumption must be regarded as a hypothesis which appears to be sufficiently justified by a number of well-known physical properties of the speech signal. In particular, this hypothesis is supported by work on artificial speech synthesis.

For example, if voiced sounds, which are often generated purely by the vocal chords, are synthesized by the periodic repetition of the oscillogram of a single period of oscillation, then a high degree of discriminability is achieved in a wide range of variations of the total duration of the sound signal. This corresponds to the case when the sequence of synthetic sound elements are obtained by a repetition of the same element, if the readings of the x elements are taken with a step of T_0, where T_0 is the length of the period.

For the very simple word "ma" the standard sequences ϑ_l^k of different length l differ only in the number of repetitions of the elements ϑ_i^k, corresponding to the stationary parts of the sounds [m] and [a], while the elements of the transient part of the sound [m] and of the sound [a] are unchanged, since it is impossible to pronounce the word "ma" so that the transition is lengthened or (shortened). In the word "pyat'" (five) it is impossible to produce the sounds [t'] and [p'] (apart from pauses) more slowly or more quickly; but it is possible to lengthen or shorten the palatalization and [a], so that effectively it is only possible to repeat (or exclude) certain elements in a word.

We shall start from the assumption that the derivative standards ϑ_l^k are formed from an original standard ϑ_q^k by repetition (or elimination) of certain elements. This makes it possible to construct the permissible standards from the initial standard sequences ϑ_q^k, by a simple and unified method for all classes. The regions of the standards Ω^k for the individual classes will be then somewhat enlarged. Such enlargements have a negligible influence on the results of the recognition process, since the actual standards of the classes from which the observed sequences X_l are derived, will be necessarily synthesized by means of repetitions of these elements.

In the process os synthesizing the standards ϑ_l^k as a result of eliminating elements of the standard ϑ_q^k one can arrive at a standard of minimum possible duration, in which none of the elements can be eliminated, since it is not possible to articulate the word more quickly than some minimal time, and from which the derivative standards are formed only by repetition of elements.

In the minimal-duration standard of a word there are no repeating neighboring elements. We shall take as the initial standard of a class the minimum-duration standard, and we shall form the derivative standards of the class by repetitions of elements. For example, if ϑ_q^k consists of five elements ϑ_5^k (ϑ_1^k, ϑ_2^k, ϑ_3^k, ϑ_4^k, ϑ_5^k), or in a simpler notation ϑ_5^k (1, 2, 3, 4, 5), then the set of standards Ω_5^k, Ω_6^k and Ω_7^k form a sequence:

Ω_5^k

1. 1 2 3 4 5

Ω_6^k

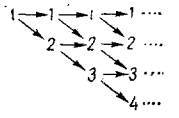
1. 1 1 2 3 4 5
2. 1 2 2 3 4 5
3. 1 2 3 3 4 5
4. 1 2 3 4 4 5
5. 1 2 3 4 5 5

Ω_7^k

1. 1 1 1 2 3 4 5
2. 1 1 2 2 3 4 5
3. 1 1 2 3 3 4 5
4. 1 1 2 3 4 4 5
5. 1 1 2 3 4 5 5
6. 1 2 2 2 3 4 5
7. 1 2 2 3 3 4 5
8. 1 2 2 3 4 4 5
9. 1 2 2 3 4 5 5
10. 1 2 3 3 3 4 5
11. 1 2 3 3 4 4 5
12. 1 2 3 3 4 5 5
13. 1 2 3 4 4 4 5
14. 1 2 3 4 4 5 5
15. 1 2 3 4 5 5 5.

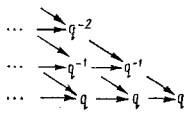
If the standard ϑ_5^k(1, 2, 3, 4, 5) corresponds to the word "odin" (one) then the first standard from Ω_7^k is a standard of the class of length $l = 7$, in which only the initial sound [a] is lengthened, in the second standard the first two sounds are lengthened, represented by the elements 1 and 2 and so on, and in the standard 15 only the last sound [n] is lengthened.

It is easy to notice that the set of standards Ω_l^k are formed from the initial standard of minimal duration, ϑ_q^k (1, 2, . . . , q) by means of a simple branching process. After the first element 1 of the standard we can have either an element 1 or an element 2. After the second element of the standard we can have either an element 1 or an element 2, if the second element is a 1; and an element 2 or an element 3, if the second element is 2. After the third element there will be an element 1 or an element 2 if the third element is a 1; an element 2 or an element 3 if the third element is a 2; an element 3 or an element 4 if the third element is a 3, and so on. This can be shown schematically as the following branching process:

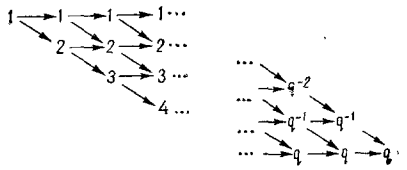


Here the digits in the i-th vertical column determine the numbers of the elements in the initial standard ϑ_q^k, which can occur at the i-th place of the derivative standard of length l.

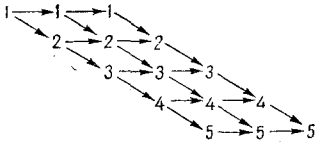
Similarly, starting from the last element of a standard ϑ_l^k of length l one can construct the terminal part of the branching process. The final element q standing at the l-th place can be preceded either by an element q − 1 or q. At the $(l - 2)$-th place there can be either an element q − 2 or q − 1 if at the $(l - 1)$-th place there is an element q − 1; and an element q − 1 or an element q if at the $(l - 1)$-th place there is an element q, and so on. This can also be represented as a branching process:



Since the sequences ϑ_l^k of the set Ω_l^k consist of l elements, therefore by joining the initial and terminal parts of the branching process we obtain exactly l vertical columns of digits. In this way we specify a set of standards Ω_l^k by means of a branching scheme.



This scheme resembles a parallelogram formed by the sides containing $q-1$ oblique arrows and $l-q$ horizontal arrows. Moving along the arrows it is possible to synthesize any permissible standard of a class of length l from the initial standard of a class of length q. We shall say that the set of standards Ω_l^k can be obtained from the initial standard ϑ_q^k of length q by means of a branching scheme of dimensions (l, q), having the structure described above. For example, the set Ω_7^k in our earlier example is represented by a branching scheme of dimensions (7,5)



Now, when all possible standards of the class have been described, for a complete construction of a model of the word signal we have to indicate a method of forming the observed sequences X_l from the standards of the class.

We shall describe the variety of sequences X_l of length l of the class k as the result of the action of the noise r on the elements of the sequences $\vartheta_l^k \in \Omega_l^k$. Let the noise r have the following properties:

1) it is additive, so that every element x_i is the result of summing the standard element ϑ_i^k with the noise r, i.e.,

$$x_i = \vartheta_i^k + r, \tag{1}$$

in expression (1) the quantities x_i, ϑ_i^k, and r should be regarded as n-dimensional vectors;

2) the probability density of the noise is a monotonic decreasing function f of the sum of squares of its components

$$p(r) = f(|r|^2),$$

where $|r|^2 = \sum_{j=0}^{n-1} r_j^2$ and r_j are the components of the noise;

3) the properties of the noise are independent of the parameters of the standard elements ϑ_i^k and of the noise of the other elements in the standard sequence.

Using this model of the speech signal we can formulate some criteria for the recognition of words in natural speech. In using a Bayesian strategy we have to have a preference for one of the hypotheses about the possible value of the class on the basis of observations on the sequence X_l. If the classes have identical a priori probabilities P(k) = const, the measure of the probability that the observed sequence belongs to the k-th class is the a priori probability of the sequence X_l, given the k-th class:

$$p(X_l/k) = \sum_{\vartheta_l \in \Omega_l^k} P(\vartheta_l/k) \cdot p(X_l/\vartheta_l), \tag{2}$$

where $P(\vartheta_l/k)$ is the probability of the standard sequence ϑ_l, given the k-th class, $p(X_l/\vartheta_l)$ is the probability that we shall observe the sequence of elements X_l if the speech signal was generated by the standard ϑ_l.

We shall assume that the standards ϑ_l^k are equiprobable over the region Ω_l^k, i.e., $P(\vartheta_l/k) = \gamma_l^k = 1/N_l^k$, where N_l^k is the number of standards of length l for the k-th class, constituting the set Ω_l^k. It follows from the properties of the noise r that the probability of the sequence of elements X_l, given the standard ϑ_l, is determined by the expression

$$p(X_l/\vartheta_l^k) = \prod_{i=1}^{l} f(|x_i - \vartheta_i^k|^2). \tag{3}$$

Then on the basis of (2) and (3) the recognition criterion is represented by

$$\max_k \gamma_l^k \sum_{\vartheta_l^k \in \Omega_l^k} \prod_{i=1}^{l} f(|x_i - \vartheta_i^k|^2). \tag{4}$$

Criterion (4) can be simplified if the sum of the probabilities is replaced by the largest value of the probabilities

$$\max_k \max_{\vartheta_l^k \in \Omega_l^k} \gamma_l^k \prod_{i=1}^{l} f(|x_i - \vartheta_i^k|^2). \tag{5}$$

In the important special case of Gaussian noise r and equal-size regions of the standards Ω_l^k (N_l^k is independent of k) criterion (5) is maximally simplified and takes the form

$$\min_k \min_{\vartheta_l^k \in \Omega_l^k} \sum_{i=1}^{l} |x_i - \vartheta_i^k|^2. \tag{6}$$

According to (6), in order to make a decision about the membership of an unknown sequence X_l in one of the K classes, it is necessary to synthesize from the standard sequence ϑ_q^k of each class, by repetition (or elimination) of certain elements ϑ_i^k, all possible sequences ϑ_l^k, and to find the one which is nearest, in the sense of the criterion of minimum Euclidean distance, to the unknown sequence X_l. Then the sequence has to be assigned to the class whose standard is in the greatest agreement with this sequence.

It appears to be probable that the recognition criterion (6) will be sufficiently efficient also when the noise r is not Gaussian. The decisive factor in criterion (6) will be sufficiently efficient also when the

the hypothesis that all possible standard sequences of the class can be obtained from one $\mathfrak{Z}_q^k$ by simple repetition (elimination) of certain of its elements.

The model of the elements (1) and the recognition criterion (6) relate to the case of intensity-normalized elements in the sequences $\mathfrak{Z}_l^k$ and X_l. In the model (1) the intensities of the elements x_i obtained from the element $\mathfrak{z}_i^k$ can vary only because of the influence of the noise r. It is possible to imagine a scheme in which recognition is carried out according to length-normalized vector elements x_i in the sequences X_l; then the model of the elements (1) and the recognition criterion (6) can be taken as a basis.

In fact, it is better to use the following model of the elements:

$$x_i = \alpha_i \mathfrak{z}_i^k + \beta_i I + r_i, \tag{7}$$

where α_i and β_i are random variables with identical distributions, the variable α_i characterizing the intensity of the element in the word; I is a vector with unit components. The component $\beta_i I$ corresponds to the action of white noise at the input of the analyzer. Together with r_i reflects the possible distortion of the standard elements due to the noise.

If we assume α_i ($i = 1, 2, \ldots, l$) and β_i ($i = 1, 2, \ldots, l$) to be independent in a first approximation, then in the case of the model (7) the recognition criterion can be taken to be, as is shown in [4],

$$\max_k \max_{\mathfrak{z}_l^k \in \Omega_l^k} \sum_{i=1}^{l} (Q_i^k)^2, \tag{8}$$

where $Q_i^k = \left(x_i, \frac{\mathfrak{z}_{0i}^k}{|\mathfrak{z}_{0i}^k|}\right)$ is the scalar product of the corresponding vectors. The elements $\mathfrak{z}_{0i}^k$ are constructed from $\mathfrak{z}_i^k$ according to a simple rule: the components of the vector $\mathfrak{z}_{0i}^k$ are obtained from the corresponding component of the vector $\mathfrak{z}_i^k$ by subtracting the mean value of the component $\mathfrak{z}_i^k$.

It is also natural to assume for speech signals that the dispersion of the noise r_i in the model of the elements (7) is proportional to the intensity of the element, i.e., $\sigma_i = \varepsilon\alpha_i$, where ε is a proportionality factor, which is the same for all elements. It can be shown that in this case the recognition criterion takes the form

$$\max_k \max_{\mathfrak{z}_l^k \in \Omega_l^k} \sum_{i=1}^{l} (R_i^k)^2, \tag{9}$$

where $R_i^k = \left(\frac{x_{0i}}{|x_{0i}|}, \frac{\mathfrak{z}_{0i}^k}{|\mathfrak{z}_{0i}^k|}\right)$, and the vector x_{0i} is obtained from x_i in the same way as $\mathfrak{z}_{0i}^k$ is obtained from $\mathfrak{z}_i^k$.

In criterion (8) elements x_i having different intensities influence the decision-making in different ways: elements of greater intensity have a larger influence. When changing to the normalized correlations R_i^k in the criterion (9) the contributions of the elements of varying intensities are equalized.

If the recognition criteria (8) and (9) are formulated in the space of values of a weighted correlation analyzer [2], i.e., the elements x_i ($x_{i_1}, x_{i_2}, \ldots, x_{in}$) are expressed in the space of values of a spectral energy analyzer through the readings of a weighted autocorrelation function of the signal b_i ($B_i(0)g_n(0)$, $B_i(1)g_n(1), \ldots, B_i(n-1)g_n(n-1)$), then it will follow from the model (7) that the first reading of the autocorrelation $B_i(0)g_n(0)$ need not be taken into account, and the expressions for Q_i^k and R_i^k take the form

$$Q_i^k = \left(B_i, \frac{B_i^k}{|B_i^k|}\right), \qquad R_i^k = \left(\frac{B_i}{|B_i|}, \frac{B_i^k}{|B_i^k|}\right),$$

where $B_i(B_i(1)g_n(1), B_i(2)g_n(2), \ldots, B_i(n-1)g_n(n-1))$ are the $(n-1)$-dimensional vector readings of the continuous weighted autocorrelation function of the signal and B_i^k are the standard vectors [2].

For normalized vector elements x_i and B_i the recognition criterion (9) is negligibly different from the recognition criterion (6): in contrast to (6) in the criterion (9) instead of the correlations of the elements of the unknown sequence with the standard elements, the squares of these correlations are inserted.

Below we shall give an algorithm for element-by-element recognition of words, which realizes the recognition criteria (6), (8), and (9) for the case when all possible standard sequences of a class are obtained from one standard by a repetition of its elements, still retaining the order of occurrence of the elements in the sequence.

AN ALGORITHM FOR ELEMENT-BY-ELEMENT RECOGNITION OF WORDS

When the standard of the k-th class is a sequence $\mathfrak{Z}_q^k$ of minimum duration q, all possible standard sequences $\mathfrak{Z}_l^k \in \Omega_l^k$ of the class k are formed from $\mathfrak{Z}_q^k$ by means of a simple branching process. As has been shown earlier, the set of standards Ω_l^k can be specified by means of a branching scheme of dimensions (l, q).

The scheme contains exactly l vertical columns of digits. We shall number these columns $i = 1, 2, \ldots, l$ from left to right. Then the digits in the i-th vertical column indicate the numbers j of the elements $\mathfrak{z}_j^k$ of the standard sequence $\mathfrak{Z}_q^k$, which can be compared with the i-th element x_i of the standard sequence X_l in the process of synthesizing the permissible standards $\mathfrak{Z}_l^k$ of length l.

We shall move along the arrows of the branching scheme, starting from the first column of digits and ending in the last one. The pathways determined by these restrictions will be called permissible. If during the movement along a permissible path we note the digits occurring along this pathway, then these digits determine a permissible standard of the class.

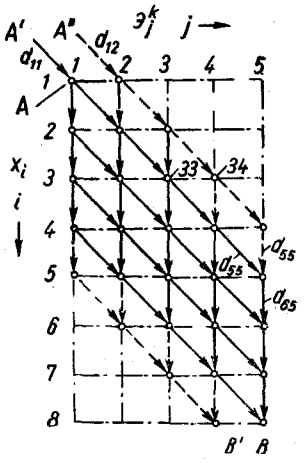
Any permissible path contains exactly $l - 1$ arrows, of which $q - 1$ are necessarily oblique.

The number of permissible pathways (according to the number of permissible standards of length l given an original standard of length q) is determined as the

number of possible combinations of $l-1$ arrows of which $q-1$ are oblique

$$N(l, q) = C_{l-1}^{q-1}. \qquad (10)$$

For example, for the often-required set of values $l = 20$ and $q = 12$ we obtain $N(20, 12) = 75\,582$. For the realization of recognition criteria (6), (8), and (9) it is hardly practicable to scan such a large number of standards in order to determine the probability that X_l will belong to one class.



We are interested in various methods of directed search among the standards of the classes, which quickly lead to a synthesis of the permissible standard of the class nearest to the unknown sequence X_l in the sense of the criteria (6), (8), and (9).

We shall show that the problem of finding the nearest permissible standard of a class is reduced to one of the problems in mathematical programming, i.e., to the problem of finding the shortest path in a graph.

We replace the digits $i = 1, 2, \ldots, l$ of the branching scheme of dimensions (l, q) by points. We assign to each point a pair of subscripts ij. The first subscript i is determined by the number of the vertical column whose digit is replaced by the point. The second subscript j is equal to the replaced digit of the branching scheme.

We assign to each arrow of the branching scheme which goes into the point ij the distance $d_{ij}^k = |x_i - \vartheta_j^k|^2$, if we want to realize the recognition criterion (6); or the value

$$(Q_{ij}^k)^2 = \left(x_i, \frac{\vartheta_{0j}^k}{|\vartheta_{0j}^k|}\right)^2 \text{ and } (R_{ij}^k)^2 = \left(\frac{x_{0i}}{|x_{0i}|}, \frac{\vartheta_{0j}^k}{|\vartheta_{0j}^k|}\right)^2$$

when we want to recognize according to the criteria (8) and (9).

Thus we obtain a graph given in Fig. 1 for the case $q = 5$ and $l = 8$.

If we move from point A of the graph to point B, we pass over exactly l points of the graph (including A and B), the second subscripts of which determine the permissible standard of length l of the class k.

The sum of the values d_{ij}^k, $(Q_{ij}^k)^2$ or $(R_{ij}^k)^2$ of the arrows lying along some pathway from A' and into B, determines the degree of correspondence between the permissible standard corresponding to this pathway and the unknown sequence X_l. It is now clear that the problem of finding the nearest permissible standard of a class has been reduced to the problem of finding the shortest [criterion (6)], or the longest [criteria (8) and (9)] pathways from A' to B in the graph determined by the model in Fig. 1. The length of the shortest or longest pathway from A' to B is a measure of the probability that the unknown sequence X_l belongs to the k-th class.

There are well-known algorithms for finding the shortest or longest pathways in a graph [5]. They are based on a directed search among the possible pathways.

The graph considered here has a special structure so that we can construct a simple procedure for finding the shortest pathway which is quite different from the methods described in the literature. The reason for this is that it is necessary only to find the length of the shortest pathway in the graph.

Let us consider successively the groups of points in the graph with identical subscripts $i = 1, 2, \ldots, l$.

We begin with the points of the graph having the subscripts $i = 1$. There is a unique point of this kind. The point $ij = 11$ can be reached along one pathway. We assign to this point the path-length $D_{11} = d_{11}$. Then we turn to the points with subscripts $i = 2$, and so on. By successively making the transition from all points with subscripts $(i-1)$ to all points with subscript i, we can assign lengths to the points of the graph with subscripts i according to the following recursive formula:

$$D_{ij} = \min(D_{(i-1)(j-1)}, D_{(i-1)j}) + d_{ij}. \qquad (11)$$

Every time the recursive formula (11) is used, we eliminate the pathways from A' to B which go through the point ij, and which are always longer than the remaining ones. For example, assigning according to (11) the value D_{ij} to the point $ij = 42$, we eliminate some of the pathways from A' to B which go through the point 42, and then we have to consider only those pathways from A' to B which are continuations of the shortest initial pathway from the point A' to the point $ij = 42$.

As a result of this procedure, on reaching the last point of the graph $ij = lq$ it will be assigned the value D_{lq}, which determines the length of the shortest pathway in the graph.

For the realization of the recognition criteria (6), (8), and (9) we construct exactly K graphs (according to the number of classes). The size of each of these graphs is determined by the number l of elements x_i in the unknown sequence X_l, and by the number q_k of elements in the standard ϑ_0^k of the class k.

It is easy to see that for the realization of the element-by-element recognition algorithm a relatively small amount of calculation will suffice: we have to determine the quantities d_{ij}^k or $(Q_{ij}^k)^2$, or $(R_{ij}^k)^2$ for each graph, and then we have to find the shortest or longest pathway and assign the unknown sequence to the class whose graph has the shortest or longest path.

When for some reason the unknown sequence X_l can lose some of its elements (the first or the last one),

then the graph (figure) must be completed on the top (see dotted lines) if the loss of elements occurs in the initial part of the sequence X_l; and at the bottom if the loss of the elements occurs at the end of the word; or in both places simultaneously if a loss of both initial and final elements occurs. The completion shown in the figure corresponds to the loss of one initial and one terminal element. After the completion we seek the shortest or longest ones of the shortest or longest pathways between the points A' and B, A' and B', A" and B, A" and B', whose length is a measure of the probability that X_l belongs to the class determined by the graph.

EXPERIMENTAL RESULTS

Below we give the results of some experiments on the recognition of words pronounced by a single speaker.

The speech signal was fed into a digital computer through an MD-55 microphone. The signal-to-noise power ratio was 10:1. The high-frequency components of the speech signal were eliminated by means of an RC-circuit with a time constant of $3 \cdot 10^{-5}$ seconds. Connection to the digital computer was achieved through a five-digit analog-to-digital converter. The sampling frequency was 10 kHz.

The words were uttered by the speaker individually. The speaker manually indicated the beginning of the input signal, and then after about 0.5 seconds he pronounced the word. The total duration of the input was 2.5 seconds.

The program determined the beginning and the end of the word. The separated useful signal of a word was subjected to a running weighted autocorrelation analysis.

The parameters of the analyzer were as follows: analysis interval $\Delta T = 20$ msec (the number $M + 1$ of discrete $\varphi(i)$ speech signals $\varphi(t)$ over the analysis interval is equal to $M + 1 = 200$); the neighboring analysis intervals did not overlap and were separated from each other by 20 msec; and the weighting function of the analyzer [2] was $g_n(s) = (s\pi/2n)\operatorname{ctg}(s\pi/2n)$.

For each analysis interval $n - 1$ values of the autocorrelation function were determined

$$B(s) = \sum_{i=0}^{M-s} \varphi(i)\varphi(i+s) \qquad (s = 1, 2, \ldots, n-1).$$

The weighted readings of the autocorrelation function $B(s)g_n(s)$ determined the elements of the word $B(B(1)g_n(1), B(2)g_n(2), \ldots, B(n-1)g_n(n-1))$.

The number n was chosen to be 21.

The running autocorrelation analysis of the word determined the sequence of elements B_i $(i = 1,2,\ldots,l)$.

The standard sequences $\mathfrak{B}_q^k$ of the classes were taken to be the sequences of elements from correctly and quickly pronounced words, so that the length l of various pronunciations of a word (class) was longer than the length of the standard sequence.

The vector elements B_i were normalized in length. After calculating the scalar products (correlations)

$$R_{ij}^k = \left(\frac{\mathbf{B}_i}{|\mathbf{B}_i|}, \frac{\mathbf{B}_j^k}{|\mathbf{B}_j^k|}\right),$$

these were assigned to the arrows going into the point ij on the graph containing the longest pathway.

If in the process of synthesizing the standards of the k-th class of length l it appeared that $l - q_k + 1 < 0$, then this meant that X_l cannot be found from the standard of the k-th class and that class did not participate in making the decision.

The recognition of the pronounced names of the ten digits 1, 2, 3, 4, 5, 6, 7, 8, 9, 0 and of the words "plus" and "minus" were attempted. There were therefore 12 classes.

No errors were found in a test sequence of 600 pronunciations of words (with 50 in each class). The duration of the pronunciation of each word varied by ±60% from the average duration of the word.

CONCLUSIONS

1. Words in spoken language can be recognized through recognition of the component elements, in spite of noise hindering the correct recognition of individual elements.

2. The recognition of words consists in scanning a huge number of standard sequences of elements. The synthesis of the standard sequence of a class which is nearest to an unknown sequence can be carried out by a directed search through the standard sequences of elements. The problem of recognizing words can be successfully solved by the methods of dynamic programming—which reduces to the problem of finding the shortest pathway in a graph.

3. The algorithm for the element-by-element recognition of words ensures the best agreement between the elements of the standard and of an unknown sequence, and eliminates the theoretical disadvantages of algorithms which involve a time normalization of the words before recognition.

4. No errors were made in an experimental test of the element-by-element recognition algorithm with a test sequence of 600 pronunciations of words from 12 classes for a single speaker.

REFERENCES

1. V.G. Zaitsev and B.B. Timofeev, "Speech discrimination," Avtomatika, no. 2, Kiev, 1965.

2. T.K. Vintsyuk, "Comparative characteristics of analyzers in speech discrimination systems," Proceedings of the Second Symposium on Cybernetics [in Russian], Tbilisi, 1965.

3. V.A. Kovalevskii, "The correlation method of discrimination," collection: Reading Automata [in Russian], izd-vo Naukova dumka, Kiev, 1965.

4. M.I. Shlezinger, "The correlation method of recognizing pattern sequences," collection: Reading Automata [in Russian], izd-vo Naukova dumka, Kiev, 1965.

5. C. Berge, Theory of Graphs and its Applications [Russian translation], IL, Moscow, 1962.

5 April 1967