УДК 621.391
B.M. Lobanov, G.S. Slutsker, A.P. Tizik

RECOGNITION OF SOUND COMBINATIONS IN THE CURRENT SPEECH SYGNAL

The article solves the task of current recognition of sound combinations in continuing speech as the task of recognition of a number of signals with noise of unknown beginning, duration and nonlinear time scale. It contains a functional block-scheme of the automaton, which realizes the described mathematical apparatus of recognition.

Introductoon

Building of phonemic vocoders and systems controlled by speech is directly connected with the task of converting speech signal normalized parameters [l]

$$\overline{\xi}(t) = \left[\xi_1(t), \xi_2(t), \mathrm{K}\ , \xi_p(t)\right], \tag{1}$$

coming to the input of the recognizer, into discrete sequences numbers of sound combinations (diphones, syllables and so on) from the finite alphabet [2, 3]

$\lambda_1, \lambda_2, \mathrm{K}\ , \lambda_k, \mathrm{K}$, where $(1 \le \lambda_k \le M)$.

Each of these numbers is put in accordance with the signal segment (1):

$$\lambda_k \leftrightarrow S^k = \left\{G^k, \overline{\xi}(t)\right\}, G^k = \left[t_0^k, t_1^k\right] \tag{2}$$

where t_0^k is beginning of λ_k sound combination in the signal (1),

$T^k = t_1^k - t_0^k$ is duration of sound combination realization.

The task (2) can be considered from the view-point of the theory of optimal signal processing as the task of M signals distinguishing at the presence of noise and distortion [4], [5] with the unknown λ, t_o and t_1. Let's consider deviations from the pattern pronunciation after normalization [1] to be Gaussian. Let the patterns of sound combinations are:

$$\varepsilon^\lambda = \left\{G^\lambda, \overline{E}^\lambda(\tau)\right\}, G^\lambda = \left[0, T^\lambda\right], \tag{3}$$

where τ is pattern time. Then the task (2) will be solved with the inequality solution:

$$\min_\lambda h^\lambda R^2\left(S, \varepsilon^\lambda\right) < H\ , \tag{4}$$

where S is looking for segment in signal (1), R is some functional determining the distance between the signals, h^{λ}, H - weight and threshold set on the bases of a' priori data [6] and chosen optimization criterion [4].

1. Method

Calculation of the distance R by the standard formula of mean square deflection of signals [4], [5] is impossible because of looking for segments (2) coincide with the patterns (3) only to within non-linear change of time scale. Using almost everywhere differentiated non-linear reflections

$$C = \{t(c), \tau(c)\}, \frac{d_t}{d_c} \geq 0, \frac{d_\tau}{d_c} \geq 0 \tag{5}$$

of the segments (2) and (3) into each other with keeping the direction of counting (7), we'll determine the distance R by the functional:

$$R^2(S, \varepsilon^{\lambda}) = \inf_C \left\{ \frac{1}{\mu C} \int_C r^2[t(C), \tau(C)] d\mu \right\}, \tag{6}$$

where $r(t,\tau)$ is the distance between samples of the signal (1) and the pattern (3) at the moments t and τ.

$d\mu = \max(dt, d\tau)$ - measure of time at the reflection C.

Because the task of current recognition differs from the task with the known boundaries [7] the value r in the formula (6) has another meaning. Let's consider in the details the metric space of the signals. In the p-dimensional vector-phase space

$$\Phi = \{\overline{X}\}, \overline{X} = (x_1, x_2, \text{K}, x_p) \tag{7}$$

let's determine the phase distance p between the vectors $\overline{X}'$ and $\overline{X}''$ by the formula

$$\rho^2(\overline{X}', \overline{X}'') = \sum_{i=1}^{P} \alpha_i (x_i' - x_i'')^2 \tag{8}$$

where α_i - weight coefficients. Let's call (p+l)-dimensional multiplication Ω of time t and phase space (7) to be the samples space

$$\Omega = t \times \Phi = \{\overline{\omega}\}, \overline{\omega} = (t, \overline{X}) = (t, x_1, x_2, \text{K}, x_p) \tag{9}$$

The distance r between the samples $\overline{\omega}'$ and $\overline{\omega}''$ of space (9) we'll determine by the formula

$$r^2 = \left(\overline{\omega}', \overline{\omega}''\right) = \delta^2\left(t', t''\right) + \rho^2\left(\overline{X}', \overline{X}''\right), \quad (10)$$

where $\delta = \sqrt{\alpha_0\left(t' - t''\right)^2}$ is time distance, p – phase distance (8) (figure 1).

Let's say that the final signal is determined as

$$S = \left\{G, \overline{\xi}(t)\right\}, G = [t_0, t_1],$$

if the time space G of its existence and p-dimensional lasting vector-function of time are set. It's easy to understand that the signal (11) is seen in the (p+l) – dimensional space of counting (9) as a space curve stretching along the axis of time t and mutually simply projected on the piece G (figure 2). A lot of signals (11) form a functional space of the final signals determined at the spaces of time with arbitrary beginning and $T=t_1-t_o$ durability. Let's say that counting $\overline{\omega} = \left(t, \overline{X}\right)$ longs to the signal (11), $\overline{\omega} \in S$, if $t \in G$ and $\overline{X} = \overline{\xi}(t)$. For a degenerative case when one of the signals converts into counting $\overline{\omega}^0$ let's call average quadrate distance (10) from $\overline{\omega}^0$ to all $\overline{\omega} \in S$, the distance between the converted signal and the signal (11) (figure 3):

$$R^2\left(\overline{\omega}^0, S\right) = \frac{1}{T}\int_G r^2\left(\overline{\omega}^0, \overline{\omega}\right)dt \quad (11)$$

For the two signals S' and S'' (figure 4):

$$R^2\left(S', S''\right) = \inf_C\left\{\frac{1}{\mu C}\int_C r^2\left(\overline{\omega}', \overline{\omega}''\right)d\mu\right\} \quad (12)$$

where - r - distance (10) between counting $\overline{\omega}' \in S'$ and $\overline{\omega}'' \in S''$, put by reflection into correspondence (5). We'll use the formula (12) to determine the distance (6) between the sample (3) and the piece (11) of the signal (1) considering

$$\tau = t - t_0, \quad (13)$$

where t_o - beginning of the piece (11).

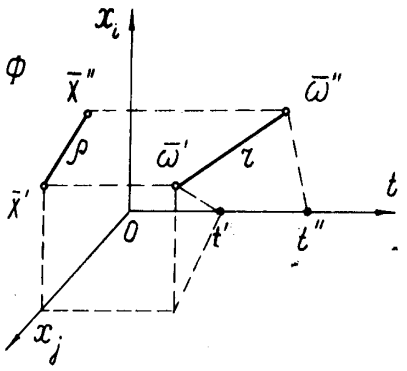

Fig. 1. A phase distance

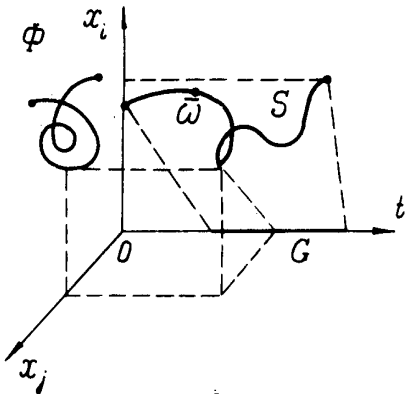

Fig. 2. A space curve stretching

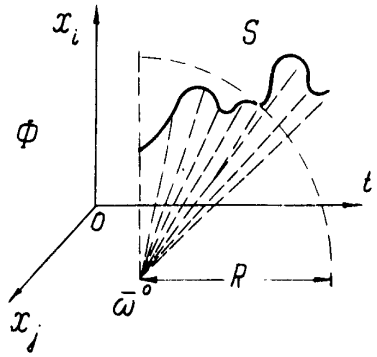

Fig. 3. The signals distance.

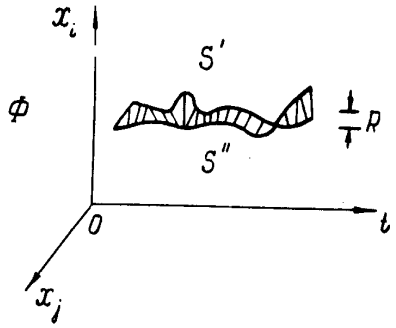

Fig. 4. Illustration of two signals

The task solution (4) consists of 4 steps:

a) looking through all the pieces (11) of the signal (1);
b) calculation of the distance between the chosen piece (11) and all the samples (3), (13);
c) choosing of the nearest sample;
d) comparing of the minimal distance got and the threshold H for decision making about the fact that the piece (11) is really a realization of one of the sound combinations which are to be recognized.

Let's put order in searching for pieces (11) at the coordinate of the end t_1. Let's enter the current time $t=t_1$. Now we'll consider all the pieces of the signal (1) with different beginning t_o and given t:

$$S^{t_0} = \left\{G^{t_0}, \overline{\xi(t)}\right\}, G^{t_0} = \left[t_0, t\right].$$

Let's determine functions of the current distance for each sample (3)

$$L^{\lambda}(t)=\min_{t_0\le t} R^2\left(S^{t_0},\varepsilon^{\lambda}\right)=R^2\left(S^{t_0^{\lambda}},\varepsilon^{\lambda}\right) \tag{14}$$

and the current length

$$T^{\lambda}(t)=t-t_0^{\lambda} \tag{15}$$

of hypothetical realization $S^{t_0^{\lambda}}$. Now we'll enter the function of minimal distance

$$L(t)=\min_{\lambda} h^{\lambda}L^{\lambda}(t) \tag{16}$$

Comparison of the function of minimal distance and the threshold will give the pieces Δ_k which will lead to

$$L(t)<H, t\in\Delta_k \tag{17}$$

The smallest meaning of the function (16) at the piece (17) will produce the coordinate of the end t_1^k and of the beginning (together with (15))

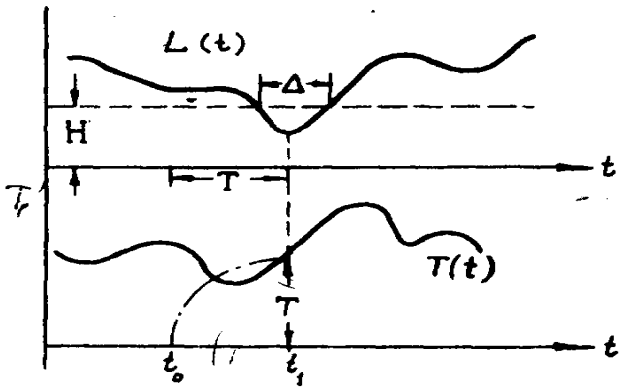

Fig. 5. Coordinates of the end and of the beginning

$$t_0^k=t_1^k-T^{\lambda_m}\left(t_1^k\right)$$

of the K piece, which realizes λ_m-e sound combination (figure 5).

2. Algorithm

Here's a recurrent algorithm for calculation of functions (14), (15). Let's say, Δt - is a step of discretization by time. The sample (3) will be in the shape of matrix

$$\varepsilon^{\lambda} = \left\{ \overline{\omega}^{\tau} \right\} = \left\{ \left(\tau, \overline{E}^{\tau} \right) \right\}, \tau = 0,1,\text{K}\ , n^{\lambda} = \frac{T^{\lambda}}{\Delta t}\ . \qquad (18)$$

The multiplier in front of the integral (6) can be omitted because $\mu C \approx T^{\lambda}$ and then can be taken into consideration in (16). We'll omit the index λ in the algorithm description because of independence of functions calculating (14), (15) for all λ.

Let's consider (n+l) of partial samples with the common beginning in 0:

$$\varepsilon_{\tau} = \left\{ \overline{\omega}^{0}\ ,\text{K}\ , \overline{\omega}^{\tau} \right\}, \tau = 0,1,\text{K}\ , n \qquad (19)$$

We'll calculate functions (14), (15) for all the partial samples (19) at the same time. Let's say, that the values L_{τ}^{t-1} of function (14) and durability T_{τ}^{t-1} of the signal pieces (l) (corresponding to the partial samples) at the moment (t –l) of the current time are known. We'll call ξt a newly come counting of signal (1). New values L_{τ}^{t} and T_{τ}^{t} for all τ are calculated beginning with τ=0, with the help of the old T_{τ}^{t-1} and L_{τ}^{t-1}.

Let's suppose

$$T_0^t = 0, L_0^t = \rho^2\left(\overline{\xi}^t, \overline{E}^0 \right) \qquad (20)$$

We'll consider $L_{\tau-1}^{t}$, $T_{\tau-1}^{t}$ for τ>0 to be known. Let's consider (figure 6) three hypotheses offering their value of T^k for the sought T_{τ}^{t}.

1. Counting $\overline{\xi}^t$ is a continuation of the piece with durability T_{τ}^{t-1} which corresponds to the part of the sample ε_{τ} at the moment (t–l). Then $T^{'} = T_{\tau}^{t-1} + 1$. Let's think $L^{'} = L_{\tau}^{t-1}$.
2. The piece with durability $T_{\tau-1}^{t}$ corresponds to the partial sample ε_{τ}. Then $T^{''} = T_{\tau-1}^{t}$. Let's suppose $L^{''} = L_{\tau-1}^{t}$.
3. Counting $\overline{\xi}^{\tau}$ is a continuation of the piece (with the durability $T_{\tau-1}^{t-1}$), which corresponds to the part of the sample $\varepsilon_{\tau-1}$ at the moment (t –l) . Then $T^{'''} = T_{\tau-1}^{t-1} + 1$. Let's suppose $L^{'''} = L_{\tau-1}^{t-1}$.

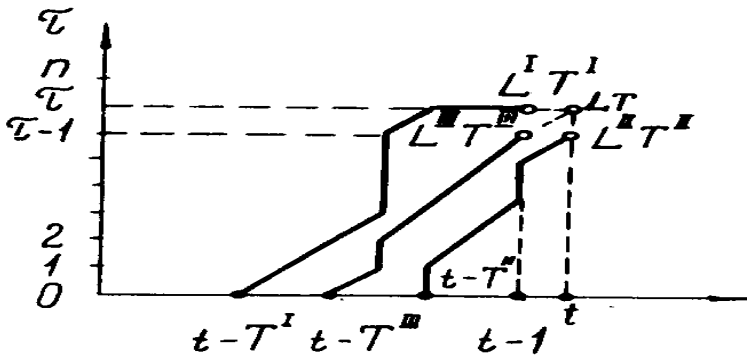

Fig. 6. Illustration of three hypotheses

Function value (14) will be

calculated by the formula

$$L_\tau^t = \rho^2\left(\overline{\xi}^t, \overline{E}^\tau\right) + \min_{K=I,II,III}\left[L^K + \delta^2\left(T^K, \tau\right)\right] \qquad (21)$$

Value T_τ^t is considered equal to T^m, where m is the number of the hypothesis chosen in (21).

3. Hardware implementation

Figure 7 contains a functional scheme of a standard elementary hardware cell that realizes the calculation of L_τ^t and T_τ^t.

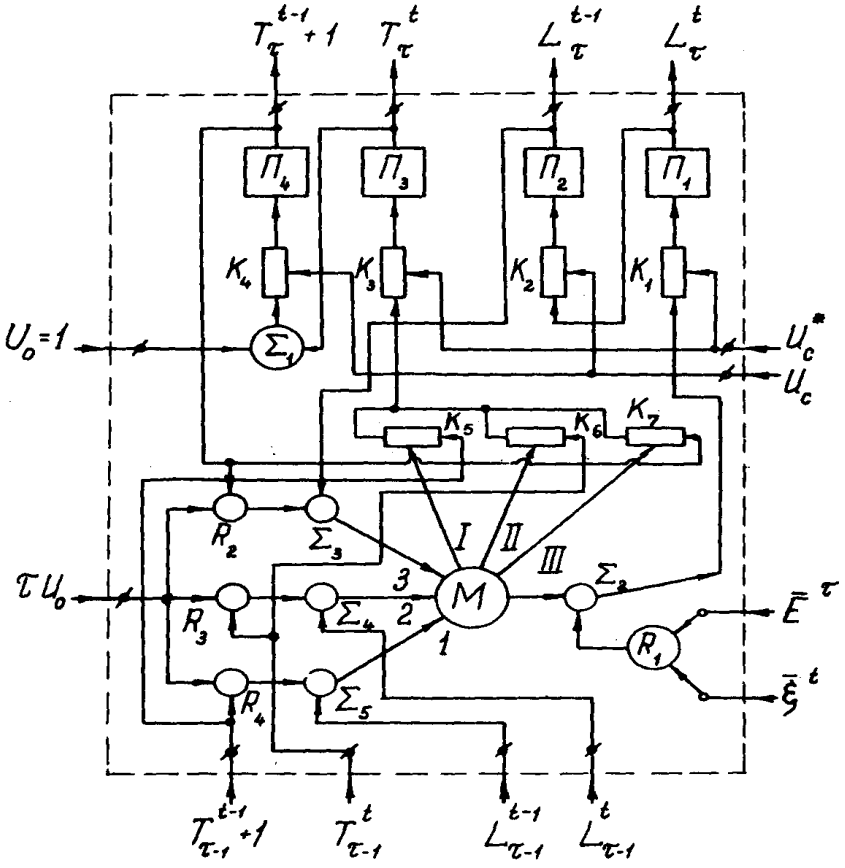

Fig. 7. A functional scheme of a standard elementary automaton cell

The cell has 10 inputs and 4 outputs. It is convenient to sort all the inputs into 3 groups according to the type of the source of input signal:

- with outside sources common for all the cells ($\overline{\xi}^t$, u_c, u_c^*, u_0);

- with outside sources which are individual for the certain τ cell (τu_0, $\overline{E}^{\tau}$);
- with sources which are outputs of the previous (τ-1) cell.

In the first group of inputs the signal $\overline{\xi}^{t}$ is a set of values p of slowly changing normalized (l) parameters of speech signal; u_c is an impulse synchronizing signal with the frequency equal to the discretization frequency $\frac{1}{\Delta t}$; u_c^* is inversion of the signal u_c; u_0 is direct voltage source which means a quantum of time.

In the second group of inputs the signal $\overline{E}^{\tau}$ is a set of p direct voltages , values of which together with the signal τu_0 describe τ counting of the sample (3).

In the third group of inputs the signals $T_{\tau-1}^{t-1}+1$, $T_{\tau-1}^{t}$, $L_{\tau-1}^{t-1}$, $L_{\tau-1}^{t}$, are output signals of (τ-1) cell, they change their values if synchronizing impulses u_c appear.

All the cell outputs can be divided into two groups: main (L_{τ}^{t}, T_{τ}^{t}) and additional (L_{τ}^{t-1}, $T_{\tau}^{t-1}+1$). The signals from the main outputs carry information about durability T_{τ}^{t} of the signal piece (1) with optimum correspondence to τ partial sample (19) and about the value L_{τ}^{t} of function (14). The signals from the additional outputs L_{τ}^{t-1}, T_{τ}^{t-1}, being saved from the previous moment of time (t – l), are transferred (adding l to time T_{τ}^{t-1}, together with the main signals to the input of the next (τ+1) cell.

A cell consists of standard functional elements:

- M is a choice scheme of minimal voltage with indication;
- $R_1 \div R_4$ – calculation schemes of vector distances;
- $\Sigma_1 \div \Sigma_5$ - calculation schemes of voltages adding ;
- $K_1 \div K_6$ - key schemes;
- $\Pi_1 \div \Pi_4$ - memory schemes for analogous voltages.

A cell works in the following way. When the signal u_c is absent the keys K1, K3 are opened and the keys K2, K4 are closed. The values L_{τ}^{t} and T_{τ}^{t} are calculated with their simultaneous remembering in the elements Π_1 , Π_3. Calculation is made in accordance with the expression (21). The three values of the expression in the square brackets are calculated with the help of calculation schemes for the vector distance R2, R3, R4 and summing schemes Σ3, Σ4, Σ5. Minimum of the values is chosen by the choice scheme M and is summed up in Σ2 with the value of vector distance $\rho^2(\overline{\xi}^{t}, \overline{E}^{t})$, calculated in R1. The result of summing up is remembered in the memory element Π_1.

The scheme M simultaneously with choosing minimal of the three voltages at the inputs and transferring it to the output, makes indication of input, which had

minimal voltage. Indication is described as appearance of voltage (at one of the three additional outputs I, II, III), which opens one of the three keys K5, K6, K7. As a result, the value T^m, to which minimum of the expression in the square brackets of the formula (21) corresponded, is remembered in the memory element Π_3.

Now let's consider the processes going on in the cell at the moments of the signal u_c transfer. The signal is transferred in the form of impulses with the discretization frequency of speech signal parameters (l). Impulse durability is chosen to be a little longer than the time of non-stationary processes in the cell. When the signal u_c appears, the keys K2, K4 are opened, and the keys K1, K3 are closed. A new voltage value equal to the voltage at output of memory element Π_1 is remembered in the memory element Π_2, and the sum of voltage equal to output voltage Π_3 and logical "1" voltage from the input u_0 is remembered in the memory element Π_4.

After the signal vanishing u_c the cell converts into a state which is initial for the next moment (t+l) of the current time.

Figure 8 contains a block- scheme of the device for sound combinations recognition.

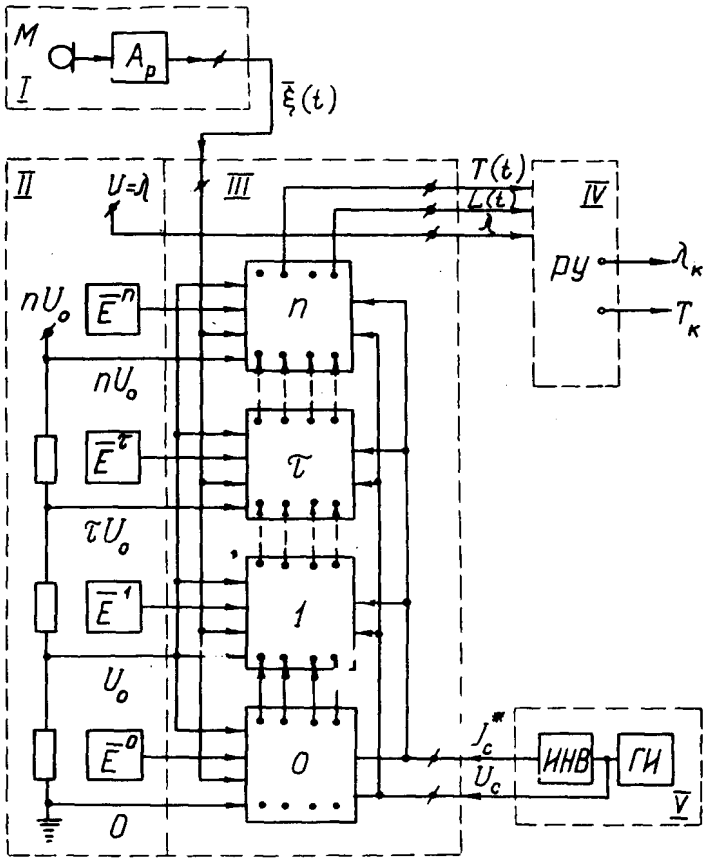

Fig. 8. Block- scheme of the device for sound combinations recognition.

The device consists of five different blocks. Block I is a production block of p-parametric normalized description of speech signal (1), which includes a microphone and an analyzer A_p.
Block II is a block of sample voltages corresponding to (n+l) counting of the sample (3).
Block III is a block of monotype cells, which realize calculation of function L_τ^t and T_τ^t (21), and which are taken according to sample counting.
Block IV is a decision-making block which makes comparison with the threshold and choice of function minimum L(t) in (17).
Block V is a synchronizing block which produces synchronizing impulses with discretization frequency of parameters of speech signal normalized description. The block includes an impulse generator IG and an inventor.
The blocks I, IV, V are common for all the sound combinations. The blocks II are monotype blocks according to their structure and differ only by the values of the sample voltages, which come to the block III inputs. The blocks III are totally monotype blocks and differ only by the number of elementary cells.

Concludion

So, we can see that the task of sound combinations recognition in continuing speech is a task of function calculation of the current distance (14) and the current length (15) which allows to determine the location (18) of their realization in the signal (1). Realization of the described algorithm of decision-making in the technical mechanism will allow to make sound combinations recognition in the real-time scale which is necessary in the systems of synthetic telephony and direct speech control.

References

1. Pirogov A.A.: To The Question Of Phonetic Coding Of Speech, Electrosvyaz, N 5, 1967.
2. Pirogov A.A.: Synthetic Telephony, Svyazizdat, Moscow, 1963.
3. Sapozhkov M.A.: Speech Signal In Cybernetics And Communication: Speech Conversion Concerning Tasks Of Communication Techniques And Cybernetics, Svyazizdat, Moscow, 1963.
4. Kharkevitch A.A.: Noise Avoiding, Fizmatgiz, Moscow, 1963.
5. Tikhonov V.I.: Statistic Radio Engineering, Soviet Radio, Moscow, 1966.
6. Yelkina V.N., Yudina L.S.: Statistics Of Russian Speech Syllables, Calculator Systems, Novosibirsk, 1964.
7. Slutsker G.S.: Non-Linear Method Of Speech Signals Analysis Works of Scientific Research Institute of Radio, N2, 1968.

Institute of Engineering Cybernetics
National Academy of Sciences Belarus,
Radio Research Institute of Moscow