ARCHIVES OF ACOUSTICS
6, 4, 359-370(1981)

THE EFFECT OF CHOSEN PARAMETERS OF A TELEPHONE CHANNEL ON VOICE IDENTIFICATION

CZESŁAW BASZTURA, WOJCIECH MAJEWSKI

Institute of Telecommunication and Acoustics, Wrocław Technical University
(50-317 Wrocław, ul. B. Prusa 53/55)

The method and results of investigations aimed at the evaluation of the effect of chosen parameters of a telephone channel on the masking of the individual voice features are presented. In the experiments on voice samples of 10 speakers (men) a speech signal was represented by the amplitude correlation matrix and the distribution of time intervals between the zero-crossings of the speech signal. The effect of the frequency band of a telephone channel and of distortions determined by different signal to noise ratios on the probability of correct voice identification was investigated. The results obtained show the possibility of voice identification under the conditions of telephone transmission provided some definite values of the parameters of a telephone link are maintained.

1. Introduction

The continuous development of new, specialized computer generations has brought a situation where the peripheral equipment has bottle-necked the application of computer systems of increasingly greater calculation capacity. One of possible solutions to overcome this limitation is equipping computer systems with acoustic terminals. In this connection, in many countries wide research has for a number of years been done on the development of an acoustic output from the computer, i.e. on the solution of the problem of speech synthesis. Efforts to develop an acoustic computer input are equally intensive, which in the field of the transmission of linguistic information requires the solution of the problem of automatic speech recognition, and in the range of the transmission of individual information, that of automatic speaker recognition. Both aspects of the automatic computer input are essential in practice, and they sometimes occur jointly. An example of this may be the case when the access to some information stored in the computer memory is reserved for authorized persons

only, i.e. is available only when the identity of a person applying for this information has been checked on the basis of analysis of his voice sample.

The cheapest and simplest way of achieving a multi-access man-computer communication based on speech can be provided by the existing telephone network. This requires, on the one hand, a number of investigations on the effect of distortions and interferences caused by a telephone channel on the possibility of both speech and speaker recognition, and, on the other hand, needs the development of reliable telephone links of good parameters of speech signal transmission.

Little attention has been paid in the literature on speaker identification to the masking of the individual voice features as a result of a broadly understood effect of a telecommunication channel. Automatic speaker identification, which concerns the procedure of assigning an unknown utterance to a speaker from a given set of speakers, is the most difficult case of the general problem of automatic speaker recognition. It is now in the stage of laboratory investigations and it will be long before it can be used in practice under the conditions of telecommunication. ATAL's publication [1], which reviews the state of art in the investigations of automatic speaker identification, gives no results of research done under the conditions which occur in practice in a telephone conversation. Some information on the effect of a telecommunication channel on the results of identification when simulated under laboratory conditions can be found, however, in other communications in this field [4, 11]. It should be added here that simulation of the effect of a telecommunication channel under laboratory conditions differs from the conditions which occur in practical implementation of a system of automatic speaker identification in that in the first case the learning and the recognized sequences are registered under the same conditions, while in the second case the two sequence are as a rule registered under different conditions, thus additionally deteriorating the identification results.

Automatic speaker verification, which consists in deciding whether the unknown utterance belongs to a particular speaker or not, i.e. a procedure of making a binary decision, is a simpler case of the general problem of automatic speaker recognition, which is closer to practical implementation than speaker identification. Therefore, the ROSENBERG review [9] gives a larger number of works accounting for the effect of the conditions of speech signal transmission from speaker to processor on the verification results. It is of interest to note another ROSENBERG's paper [10] which describes the possibility of implementing a system of automatic speaker recognition with an error less than 5 procent for speech signal transmission via a telephone link.

On the basis of the facts given above and appreciating the necessity of investigating voice recognition on the phonetic and linguistic material of the Polish language, and, additionally, bearing in mind the possibility of using the procedures of automatic voice recognition not only in the generally accessible

telephone information systems but also in crime detection, the present authors set themselves the task of investigating the effect of chosen parameters of a telephone link on the masking of the individual voice features. The object of particular interest was the effect of the frequency bandwidth of a telephone channel on the possibility of voice identification and the effect of typical distortions caused by a telephone channel, represented by different levels of the signal to noise ratio.

2. Identification system

2.1. *Introduction.*

Automatic speaker recognition can be implemented in a system of coupled analogue and digital units, called a recognition system. The starting point for system selection is above all the establishing of the aim of recognition. The aim of the present investigation was to examine the effect of chosen parameters of a telephone channel on the possibility of speaker identification. A simple identification system, with a teacher, and the statistical criteria of decision-making [12], was selected for the implementation of this aim.

2.2. *Block diagram of the identification system.*

A simplified block diagram of the identification systems used is shown in Fig. 1. This identification system can be divided into the three basic units:

(a) the signal source unit;

(b) the unit for measurement, i.e. extraction, of parameter sets,

(c) the classification unit.

The signal source is provided by recorded utterances of M speakers saying the same test (the key phrase).

The measurement unit is an analogue-to-digital system for extraction from a speech signal of the parameters significant from the point of view of the individual features.

The classification unit consists of the information systems: the predetermined information and the "teacher", and the classifier proper including the subunit for making patterns and the identification algorithm with criteria and similarity measures. All the three units are interdependent on each other, i.e. the signal source influences the measurement unit, and the selection of the parameter set affects, in turn, both the way of making patterns and the kind of probability measure and the decision-making criterion of the identification algorithm.

Following the advantages of the model, a short-term recognition model [3] was used in this system.

In order to improve the reliability of the experimental results, two kinds of parameter sets were used. The first was the distribution of the time intervals between the zero-crossings of a speech signal [2, 3], the second was the amplitude

correlation matrix (ACM) [7], based on the short-term spectral analysis. Application of different parameter sets implies another way of pattern making and other similarity measures. Therefore, despite the retaining of one identification algorithm NM (nearest mean), the procedures differ in details and, accordingly, require separate discussion.

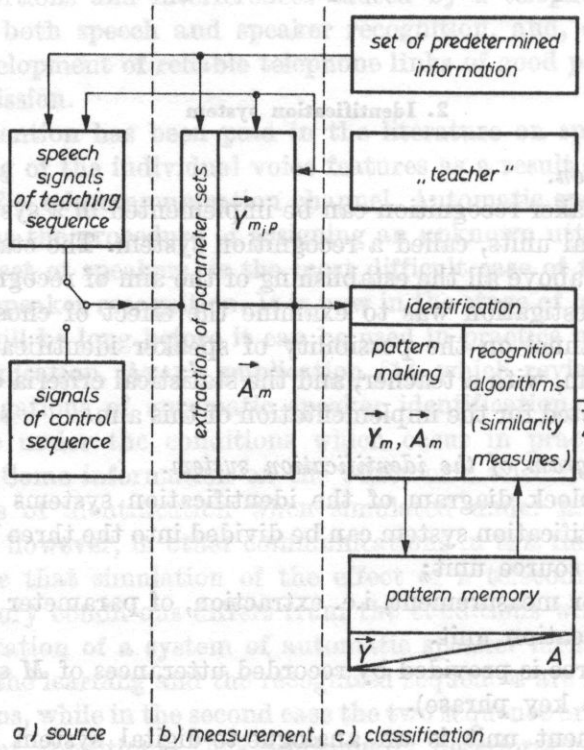

Fig. 1. A block diagram of the identification system

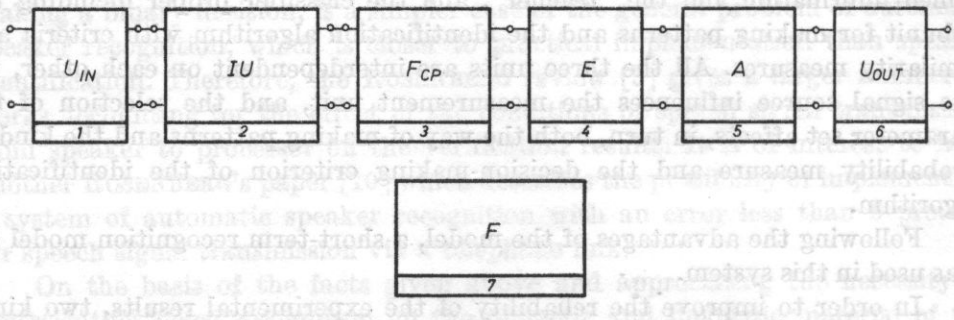

Fig. 2. A block diagram of the telephone channel model

1, 6 — input and output systems, *2* — interference unit, *3* — central-pass filter, *4* — frequency response equalizer, *5* — attenuator, *7* — feeder

2.3. *The distribution of the time intervals between the zero-crossings.*

Analysis of the previous investigations [2, 3] shows a fairly large effectiveness of representing the individual voice features by the parameter set with $\boldsymbol{Y}_{m,p}$, being the distribution of the time intervals between the zero-crossings of a speech signal

$$\boldsymbol{Y}_{m,p} = \{Y_{m,p,1}\, Y_{m,p,2} \ldots Y_{m,p,k} \ldots Y_{m,p,K}\}, \tag{1}$$

where $m = 1, 2, \ldots, M$ (M – the number of speakers), $p = 1, 2, \ldots, P_m$ (P_m – the number of utterance repetitions in the learning or the control sequence for the mth speaker), $k = 1, 2, \ldots, K$ (K – the number of time channels).

The stage following the parameter set making is the making of patterns of voice images. The basis for pattern making in this system is the learning sequence

$$\{Cu\} = (Y_{1,1}, \lambda_1), \ldots, (Y_{1,P_1}, \lambda_1), \ldots, (Y_{m,p_m}, \lambda_m), \ldots, (Y_{M,P_M}, \lambda_M), \tag{2}$$

where λ_m is the speaker class membership predetermined by the "teacher". In the heuristic NM algorithm the voice class patterns are the mean vectors from the repetitions $\boldsymbol{Y}_m$, and the Mahalanobis squared distance [3] was used as similarity functions.

2.4. *Amplitude correlation matrix*

The amplitude correlation matrix was used for speaker recognition in the long-term recognition model, by LI and HUGHES [7], for example. The essence of voice image description in this method consists in using the speaker-dependent correlations beetween the amplitudes for the particular frequency bands. Let a_t represent the vector (of the size K), being a set of the amplitudes for the individual $K \times T$ frequencies (Fig. 3).

Calculating the correlation matrix of amplitudes in the relation

$$A_t(i, j) = \left[(a_t(i) - s(i))\,(a_t(j) - a(j))\right] / \sigma_i \sigma_j, \tag{3}$$

where

$$a(l) = \frac{1}{T} \sum_{t=1}^{T} a_t(l), \qquad \sigma_l = \left[\frac{1}{T} \sum_{t=1}^{T} (a_t(l) - a(l))^2\right]^{1/2},$$

$$i, j, l = 1, 2, 3, \ldots, K, \qquad t = 1, 2, 3, \ldots, T,$$

it can be regarded as a representation of the utterance for a stationary speech signal in the long-term model [2, 4, 7], or for a definite utterance in the short-term model [3].

When using the identification algorithm NM analogously to the distributions of the time intervals, the voice image patterns of the learning sequence are based on the following relation

$$A^m(i, j) = \frac{1}{P_m} \sum_{p=1}^{P_m} A^{m,p}(i, j), \tag{4}$$

where

$$A^{m,p}(i,j) = \frac{1}{T}\sum_{t=1}^{T} A_t^{m,p}(i,j). \tag{5}$$

$A^{m,p}(i,j)$ is the correlation matrix of amplitudes of the size $K \times K$ representing the pth repetition of the mth speaker. Since the matrix $A(i,j)$ is a real, symmetrical matrix, it is sufficient for every second element of the matrix to be stored in the pattern memory, which permits a considerable saving of memory and calculation time.

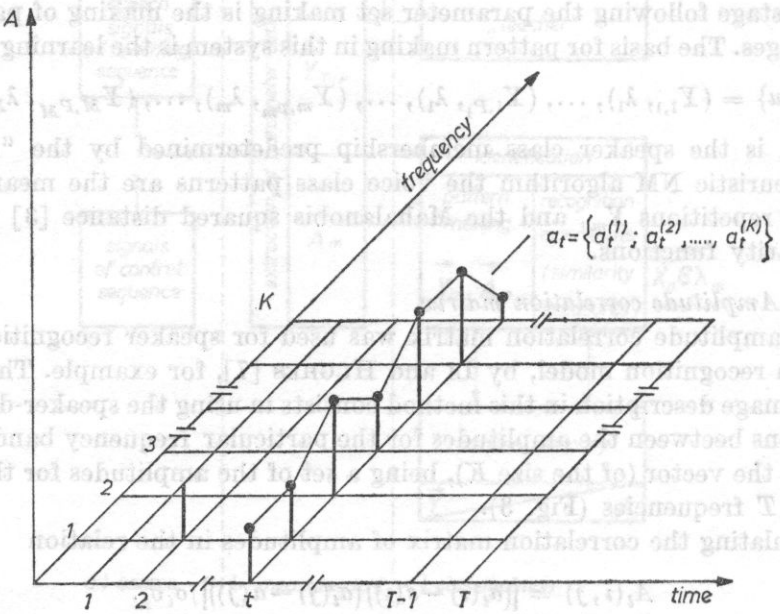

Fig. 3. Discrete representation of the short-term spectrum for the t-sample

The definition of the common decision-making rule of the NM algorithm was given in refs. [2, 3]. In view of the specificity of the parameter set $A(i,j)$, after the authors of ref. [7], two similarity functions were used experimentally. The first, defined as

$$d_{n,m}^{(1)} = \Big[\sum_{j=1}^{K}\sum_{i=j}^{K} B^n(i,j) - A^m(i,j)|^{1/2}\Big] \Big/ \Big[\sum_{j=1}^{K}\sum_{i=j}^{K} (B^n(i,j))^2\Big]^{1/2}, \tag{6}$$

is a normalized Euclidean distance between the two matrices, where $B^n(i,j)$ is, analogously to $A^m(i,j)$, the image of the nth recognized utterance. The second similarity function, defined as the average absolute difference between the matrices $A(i,j)$ and $B(i,j)$, can be calculated from the relation

$$d_{n,m}^{(2)} = \frac{2}{K(K+1)} \sum_{j=1}^{K}\sum_{i=j}^{K} |B^n(i,j) - A^m(i,j)|. \tag{7}$$

3. Identification experiment

3.1. *Introduction*

The considerations in sections 1 and 2 were the basis for the identification experiment aimed at answering the question whether and to what degree the basic parameters of a telephone channel, for which the frequency band and the signal to noise ratio were taken, affect the masking of the individual voice features. This requires sound material recorded under suitable technical conditions. In view of the desired measurement stability, particularly essential with the few identification statistics (the number of speakers, the lengths of the learning and the control sequences), the present investigations did not include measurements under the real conditions of telephone transmission, and these conditions were simulated using a model of telephone channel developed and built in the Institute of Telecommunication and Acoustics, Wrocław Technical University.

3.2. *Model of telephone channel*

The model of telephone channel (Fig. 2) is an analogue system implementing the predetermined physical parameters and the basic characteristics of a typical population of telephone channels. The variability range (the possibility of adjustment) covers the real range of variations in the parameters of the frequency response and of the distortions of typical telephone lines in natural telephony [5, 6]. The channel model consists of input and output systems and the five units of:

(a) additive distortions,
(b) the frequency response equalizer,
(c) the central-pass band filter,
(d) the attenuator,
(e) feeders.

The nominal signal level is 0 dB. The maximum input and output voltage is 3V. The input and output resistance is 600 Ω.

3.3. *Selection of the key phrase*

The present investigation used the text "jutro będzie ładny dzień" as the key phrase. This selection was justified by the easy pronunciation and the frequent use of the words in the phrase and the relatively good approximation of the mean statistics of the Polish language in terms of the frequency spectrum and the occurrence frequency of the phonemes [8].

3.4. *Population of speakers*

In view of the rather time-consuming calculations the population of speakers was limited to 10 persons-men from 20 to 35 years old. The key phrase was recorded in two sessions (I and II) at an interval of one month. The length of the learning sequence was 30, while that of the control sequence was 20, which required three repetitions of the key phrase by each of the speakers in session I

and two repetitions in session II. Irrespective of the kind of experiment the learning sequence was taken from session I, and the control sequence from session II.

3.5. *Preparation of the sound material*

The speakers' utterances were recorded in a listening studio of the attenuation of external distortions of — 30 dB. The speech signal was recorded on Superton C-60 Low-Noise cassettes using a M 601 SD Unitra cassette tape recorder manufactured by ZRK.

In order to introduce distortions and interferences the sound material recorded was fed to the channel model and when it passed through it was recorded again on the same recording equipment.

3.6. *Programme of the experiments*

The programme assumed for the experiments in speaker identification is shown in Fig. 4. Experiment 0 made for a speech signal of the form obtained directly from preliminary recordings, aimed at achieving some reference measure.

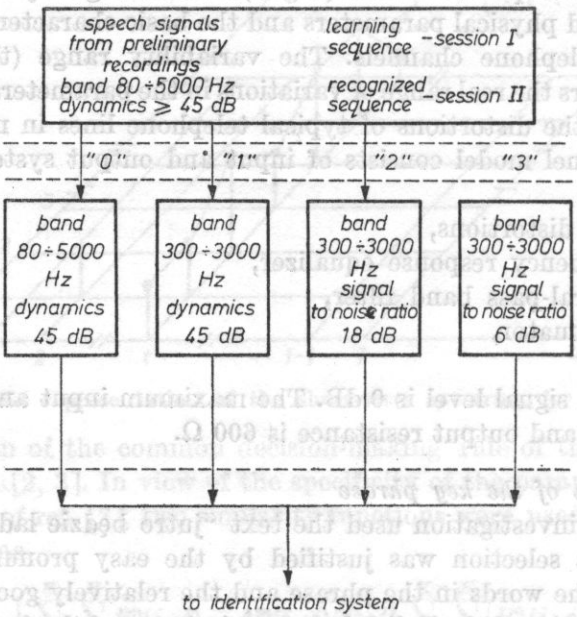

Fig. 4. Programme of the experiments

Experiments 1-3 were made for the parameters of the model representing typical parameters of telephone links and giving essential differences between the mean speech intelligibility measured subjectively (using 10 listeners and 400 *PB* nonsense words corresponding to each setting of the channel model).

The mean intelligibility was 96.3 per cent for experiment 0, while for experiments 1-3 it was 84.5, 73.0 and 66.1 per cent, respectively.

The principle and technical data of the programme implementing the extraction of the parameters Y_m, i.e. the distribution of the time intervals between the zero-crossings, were presented in greater detail in refs. [2, 3]. On the basis of these papers, 16 time channels $K = 16$ of the nominal boundary values shown in Table 1 were taken. In the identification experiments made according to the algorithm given in refs. [2, 3] 16-dimensional vectors of the parameters Y_m were thus used.

Table 1. The boundaries of time channels

k	t_{k-1} [ms]	t_k [ms]	k	t_{k-1} [ms]	t_k [ms]
1	2	3	1	2	3
1	0.15	0.189	9	0.969	1.216
2	0.189	0.238	10	1.216	1.535
3	0.238	0.301	11	1.535	1.937
4	0.301	0.380	12	1.937	2.445
5	0.380	0.478	13	2.445	3.085
6	0.478	0.605	14	3.085	3.893
7	0.605	0.764	15	3.893	4.913
8	0.764	0.969	16	4.913	6.200

The programme of the extraction of the correlation matrices of the amplitudes $A(i, j)$ consisted of two parts. The first stage was a FFT (fast Fourier transform) calculating the amplitude spectrum with a Hamming window of 6.4 m width and the gap between the windows $\Delta t = 20$ ms, which permitted spectral lines at an interval of 156.26 Hz. In view of the differences in duration of the registered utterances of the elements of the key phrase, which did not exceed 10-15 per cent of the mean duration, no time normalization was made and the number T of windows taken, i.e.

$$T = \frac{T_{\text{av}}}{\Delta t}, \tag{8}$$

where $T_{\text{av}} = 0.710$ s is the mean duration for the investigated set of $M \times P$ utterances. $K = 20$ first lines of the spectrum were used to make the pattern and recognized images.

3.7. *Experimental results*

The results obtained from the identification experiments for the two representations of voices Y and A are shown in Table 2.

Table 2. The identification results (percentage of correct decisions) for the parameter sets Y and A

Learning sequence -experiment	Control sequence -experiment	Reprezentation Y [%]	Representation A [%]
	0	90	80
	1	60	55
0	2	65	50
	3	35	25
1	1	85	75
2	2	85	80
3	3	40	40

NB. When the correlation matrix of amplitudes was used for voice description the same results of correct identification were obtained for both similarity functions $d^{(1)}_{n,m}$ and $d^{(2)}_{n,m}$.

4. Conclusion

The experiments permit the following observations and conclusions to be given:

1. Both parameter sets show an almost similarly effective representation of the individual voice features and an almost similar liability to the masking of these features by distortions and interferences caused by the telephone channel, except that slightly better identification results were obtained for the voice description using the distribution of the zero-crossings.

2. The experiments confirmed the thesis, proposed in chapter 1, of the essential effect of the agreement in the conditions of transmission between the signals forming the pattern sequences and those making up the recognized sequences, on the probability of correct voice identification. With this agreement, the identification results were distinctly better.

3. For a recognized sequence composed of a signal contained in the telephone band 300-3000 Hz a change in the value of the signal to noise ratio from 45 dB to 18 dB only slightly affected the identification results. When the pattern sequence was also composed of the signal defined above, the identification results were comparable to the results in the reference system, i.e. for the band 80-5000 Hz and the dynamics 45 dB. It follows therefore that the telephone band and the signal to noise ratio of the order of 18 dB or more do not consitute any essential obstacle to the achievement of safisfactory recognition results provided the agreement in the transmission conditions between the pattern and recognized signals is maintained.

4. Comparison of the identification results shown in Table 2 with the results of speech intelligibility measurements given in section 3.6 shows that distortions nad interferences caused by the telephone channel have a different effect on the masking of the individual voice features, compared to that on the intelligibility of speech. A rapid worsening of the identification results, i.e. weakening of the ability of the telephone channel to transmit individual information, occurred only for the channel parameters defined by experiment 3, while a weakening of the ability of the channel to transmit linguistic information in terms of speech intelligibility occurred in a smooth manner from the condition defined by experiment "0" to those defined by experiment "3". Accordingly, it can be expected that attempts at evaluation of the ability of the telecommunication channel to transmit individual information, based on the quality of the transmission of linguistic information and conversely, will not be succesful.

The future investigations of voice identification under the conditions of telephone transmission should concentrate mainly on the selection of such parameters of a speech signal, which being good carriers of the individual voice features, would not be liable to distortions and interferences caused by the telephone channel, or on the development of such methods of speech signal analysis as permit the compensation for the effect of unknown transmission conditions. Another important task is the definition of the numerical relations between the kind and value of distortions and interferences caused by the telephone channel and the probability of voice identification in the systems assumed.

The solution of these problems and others, difficult and time-consuming though it is, will permit a practical application of systems of automatic speaker recognition under the conditions of telephone communication. The authors hope that the present paper is a contribution to the implementation of this prospect.

References

[1] B. S. Atal, *Automatic recognition of speakers from their voices*, Proc. IEEE, **64**, 4, 460-475 (1976).

[2] C. Basztura, W. Majewski, *The application of long-term analysis of the zero-crossing of a speech signal in automatic speaker identification*, Archives of Acoustics, **3**, 1, 3-15 (1978).

[3] C. Basztura, J. Jurkiewicz, *The zero-crossing analysis of a speech signal in the short-term method of automatic speaker identification*, Archives of Acoustics, **3**, 3, 185-195 (1978).

[4] H. Hollien, W. Majewski, *Speaker identification by long-term spectra under normal and distorted speech conditions*, JASA, **62**, 4, 975-980 (1977).

[5] *Green book of* CCITT, vol. III-1, *Line transmission*, WKiŁ, Warsaw 1976.

[6] *Green book of* CCITT, vol. V, *Quality of telephone transmission, local networks and telephones*, WKiŁ, Warsaw 1976.

[7] K. P. LI, G. W. HUGHES, *Talker differences as they appear in correlation matrices of continuous speech spectra*, JASA, **55**, 4, 833-837 (1974).

[8] W. MAJEWSKI, *Phonetic test for subjective measurements of reference equivalent*, Przegląd Telekomunikacyjny, 8, 237-240 (1979).

[9] A. E. ROSENBERG, *Automatic speaker verification: a review*, Proc. IEEE, **64**, 4, 475-487 (1976).

[10] A. E. ROSENBERG, *Evaluation of an automatic speaker verification system over telephone lines*, Bell System Technical Journal, **55**, 6, 723-744 (1976).

[11] M. R. SAMBUR, *Speaker recognition using orthogonal linear prediction*, IEEE Trans. on Acoustics, Speech and Signal Proc. ASSP – **24**, 4, 283-289 (1976).

[12] V. A. SKRIPKIN, A. L. GORELIK, *Metody rozpoznavanja*, Moscow, Vizhshaya Shkola, 1977.

Received on September 26, 1980; revised version on April 7, 1981.