ARCHIVES OF ACOUSTICS
3, 3, 185-196 (1978)

THE ZERO-CROSSING ANALYSIS OF A SPEECH SIGNAL IN THE SHORT-TERM METHOD OF AUTOMATIC SPEAKER IDENTIFICATION

CZESŁAW BASZTURA, JERZY JURKIEWICZ

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

The aim of this paper is to investigate the possibility of using the zero-crossing analysis of a speech signal in a short-term method of speaker identification. Four sets of parameters obtained with the aid of the zero-crossing analysis are presented which can find application in automatic speaker identification. An experiment of speaker identification for 20 male speakers has been performed. The obtained results have confirmed the applicability of the method of zero-crossing analysis for tracking individual features of voices.

1. Introduction

Thecom plex ityfo the speech process both as regards its mental and articulatory aspect is manifested by the occurrence in a speech signal of a multitude of extralinguistic information, including also that about individual features of speaker's voice. Practice has shown that individual features contained in a speech signal enable the recognition of a speaker on the basis of his statements.

The main stimulus for investigations on the automatic speaker identification is the development of computer techniques and the availability of computers for the processing of the speech signal. Automatic speaker identification can be accomplished by using two identification models which differ primarily by the kind and duration of statement text.

The method based on the long-term analysis features some degree of independence on the text and a relatively long duration of statement. The method of the short-term analysis is based on the individual parameters of the voice obtained from fixed text in a time ranging from a fraction of a second for single phonems, up to several seconds for sentences. The short-term analysis method necessitates the use of time normalization of the signals to be analyzed or of vectors representing these signals in the adopted parameter space.

The speaker identification method should give an adequate and invariant description of the voices of speakers. The sets of parameters for the automatic speaker identification should be [1]:

1° effective in representation of individual features of speakers,
2° easy to measure,
3° stable over the required period of time,
4° little sensitive to ambient conditions changes,
5° hard to imitate.

So far no such set of parameters, which would satisfactorily meet all these requirements, has been found. In view of a complex structure of a speech signal and the lack of explicit premises the choice of the parameters discriminating the voices is dictated mostly by heuristic reasons based on the previous experiments, the acquired knowledge, and even intuition [1, 3, 6].

After such a choice has been made, it is necessary to substantiate it theoretically and experimentally in order to verify the assumed hypothesis of the practicability of the set of parameters used for the automatic speaker identification.

The results of investigations obtained by many authors have convincingly confirmed unquestionable advantages of the zero-crossing analysis of a speech signal, e.g. for the speaker identification and speech analysis [2, 4, 7].

The aim of this paper is to investigate the possibility of using the zero-crossing analysis of a speech signal for automatic speaker identification by the short-term method.

2. Methods of investigations

2.1. *Choice of the statement text.* For the methods based on the short-term analysis one chooses texts which are easy to pronounce, widely used and contain phonetic elements which provide as much as possible information on the individual features of speaker's voice.

The investigations carried out by many authors indicate that vowels as well as lateral, liquid and nasal consonants contribute most to the differentiation of individual features. This results from the fact that the phonation of vowels depends on the shape and size of the vocal tract of a speaker and the properties of the source of his laryngeal tone. The spectra of nasal consonants, however, are closely related to the nasal cavity and to its interaction with the mouth cavity. The position of the tongue and teeth greatly affects the articulation of the lateral consonants. In keeping with this consideration the word ALO has been chosen as a text for a statement.

An additional motivation for the choice of this word is its wide use in colloquial speech (as voiced part of the word HALO), especially when starting a phone conversation.

2.2. *The sets of parameters used for the description of individual features of speakers.* The creation of a suitable parameter space is regarded to be one of the most difficult stages in the process of the automatic speaker identification. It should be stressed that the determination of the values of parameters such as the fundamental frequency F_0 or the frequency of formants using the zero-crossing analysis (ZCA) is not accurate enough to justify the use of these parameters for speaker identification. For this reason it seems advisable to define, with the aid of the ZCA method, the other sets of parameters which satisfy, to some extent, the requirements listed in the introduction.

The description of individual features of voice presented in this paper has been based on the distribution of time intervals and parameters based on the time dependence of the density of zero-crossings of a speech signal.

(a) The function of zero-crossings of a signal.

For the time-dependent signal $U = U(t)$ the function $P(U, t)$ of a zero-crossing has been defined,

$$P(U,t) = \begin{cases} 1 & \text{if there exists } U(t) \text{ satisfying conditions (i)-(iii),} \\ 0 & \text{if there is no } U(t) \text{ satisfying conditions (i)-(iii),} \end{cases} \tag{1}$$

where

$$U(t)\,U(t-\tau) < 0, \tag{i}$$

$$|U(t)| \geqslant a \quad \text{and} \quad |U(t-\tau)| \geqslant a, \tag{ii}$$

$$|U(x)| < a \quad \text{for} \quad t-\tau < x < t, \tag{iii}$$

and a is a threshold level ($a \neq 0$) which prevents counting additional zero-crossings caused by disturbances [4].

The values t_i for which $P(U, t_i) = 1$ are the moments of the zero-crossing of the signal $U(t)$.

(b) Distributions of time intervals.

Using the function $P(U,t)$ makes it possible to present the time dependence of the signal $U(t)$ in a simpler form by means of segments of lengths t_j which are equal to the intervals between successive zero-crossings of a speech signal. For a given signal representing a selected text of a statement one can choose the limiting values t_d and t_g as extreme values of the time intervals between zero-crossings. If in the interval (t_d, t_g) we distribute $K-1$ threshold values, then we obtain K time channels. The values $y(t_{k-1}, t_k)$ represent the number of intervals t_j contained in the interval (t_{k-1}, t_k).

If the time dependent signal $U_{m,i}(t)$ of a duration $T_{m,i}$ constitutes a pattern of the m-th speaker and of the i-th repetition of the speaker's voice, then this pattern can be represented in the form of a K-dimensional vector

$$\boldsymbol{y}_{m,i} = \{y_{m,i,1}, y_{m,i,2}, \ldots, y_{m,i,k}, \ldots, y_{m,i,K}\}. \tag{2}$$

An example of the distribution of time intervals is shown in Fig. 1.

(c) Time dependence of the density of zero-crossings.

The results presented in paper [7] suggest that it is possible to use the time dependence of the density function of zero-crossings for voice discrimination. The measurement of the density $\varrho(kt_N) = \varrho[U(kt_N)]$ from a speech signal in discrete form U_n is made according to the relation

$$\varrho[U(kt_N)] = \frac{1}{t_N} \sum_{n=1}^{N} P\left(U_n, (k-1)t_N + \frac{n}{f_{\mathrm{pr}}}\right), \tag{3}$$

where k is the index of the signal segment of duration t_N, N – the number of signal samples in a given segment, f_{pr} – the sampling frequency, and $t_N = N/f_{\mathrm{pr}}$.

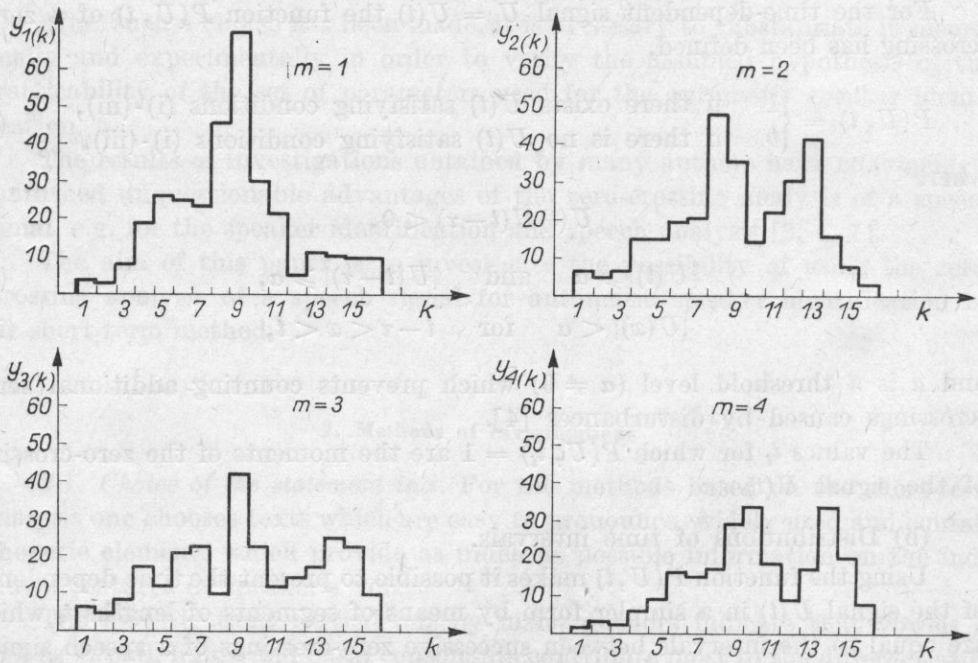

Fig. 1. The averaged distributions of time intervals for 4 speakers (averaging from 5 repetitions of statements)

The function $\varrho(kt_N)$ for the m-th speaker and the i-th repetition can be presented in the form of the vector

$$\boldsymbol{\rho}_{m,i} = \{\varrho_{m,i}(1), \varrho_{m,i}(2), \ldots, \varrho_{m,i}(K_{m,i})\}. \tag{4}$$

The dimension $K_{m,i}$ of the vector depends on the speaker, as also on the given repetition of the statement (especially as regards its rate) and this necessitates the use of time normalization [5].

(d) Parameters defined on the basis of the total time during which the points remain in given sectors of the phase plane $\{\varrho[U(t)], \varrho'[U(t)]\}$.

The time dependence of the zero-crossing density changes can be represented by the zero-crossing derivative. A speech signal can be described by a set of points with coordinates

$$\left\{\varrho[U(kt_N)], \frac{\Delta\varrho[U(kt_N)]}{t_N}\right\}$$

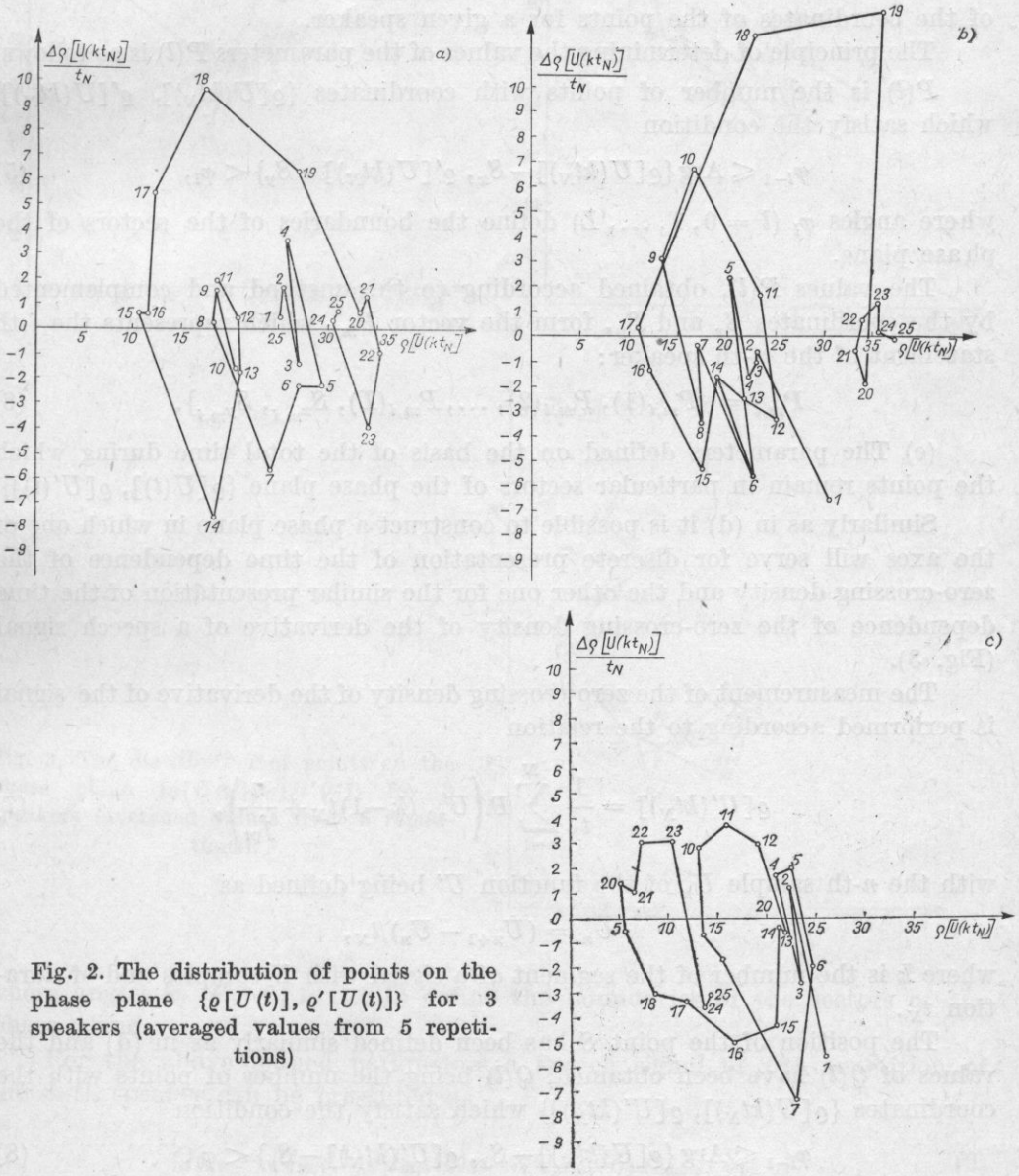

Fig. 2. The distribution of points on the phase plane $\{\varrho[U(t)], \varrho'[U(t)]\}$ for 3 speakers (averaged values from 5 repetitions)

belonging to the phase plane being a discrete representation of the plane $\{\varrho[U(t)], \varrho'[U(t)]\}$.

A given signal can also be characterized by the numbers $P(l)$ corresponding to the total time during which the points with indices $1, 2, \ldots, K_{m,i}$ remain in definite sectors of the phase plane (Fig. 2). By selecting a point S with coordinates S_x and S_y as a centre of the set of these points and assuming it as a central point of partition, it is possible to divide the phase plane into L sectors.

It is convenient to define the coordinates S_x and S_y as arithmetical means of the coordinates of the points for a given speaker.

The principle of determining the values of the parameters $P(l)$ is as follows:

$P(l)$ is the number of points with coordinates $\{\varrho[U(kt_N)], \varrho'[U(kt_N)]\}$ which satisfy the condition

$$\varphi_{l-1} \leqslant \operatorname{Arg}\{\varrho[U(kt_N)] - S_x, \varrho'[U(kt_N)] - S_y\} < \varphi_l, \tag{5}$$

where angles φ_l $(l = 0, 1, \ldots, L)$ define the boundaries of the sectors of the phase plane.

The values $P(l)$, obtained according to this method and complemented by the coordinates S_x and S_y, form the vector $\boldsymbol{P}_{m,i}$ which represents the i-th statement of the m-th speaker:

$$\boldsymbol{P}_{m,i} = \{P_{m,i}(1), P_{m,i}(2), \ldots, P_{m,i}(L), S_{x_{m,i}}, S_{y_{m,i}}\}. \tag{6}$$

(e) The parameters defined on the basis of the total time during which the points remain in particular sectors of the phase plane $\{\varrho[U(t)], \varrho[U'(t)]\}$.

Similarly as in (d) it is possible to construct a phase plane in which one of the axes will serve for discrete presentation of the time dependence of the zero-crossing density and the other one for the similar presentation of the time dependence of the zero-crossing density of the derivative of a speech signal (Fig. 3).

The measurement of the zero-crossing density of the derivative of the signal is performed according to the relation

$$\varrho[U'(kt_N)] = \frac{1}{t_N} \sum_{n=1}^{N} P\left(U'_n, (k-1)t_N + \frac{n}{f_{\mathrm{pr}}}\right), \tag{7}$$

with the n-th sample U'_n of the function U' being defined as

$$U'_n = (U_{n+1} - U_n)/t_N,$$

where k is the number of the segment of a signal with N samples and of duration t_N.

The position of the point S has been defined similarly as in (d) and the values of $Q(l)$ have been obtained, $Q(l)$ being the number of points with the coordinates $\{\varrho[U(kt_N)], \varrho[U'(kt_N)]\}$ which satisfy the condition

$$\varphi_{l-1} \leqslant \operatorname{Arg}\{\varrho[U(kt_N)] - S_x, \varrho[U'(kt_N)] - S_y\} < \varphi_l, \tag{8}$$

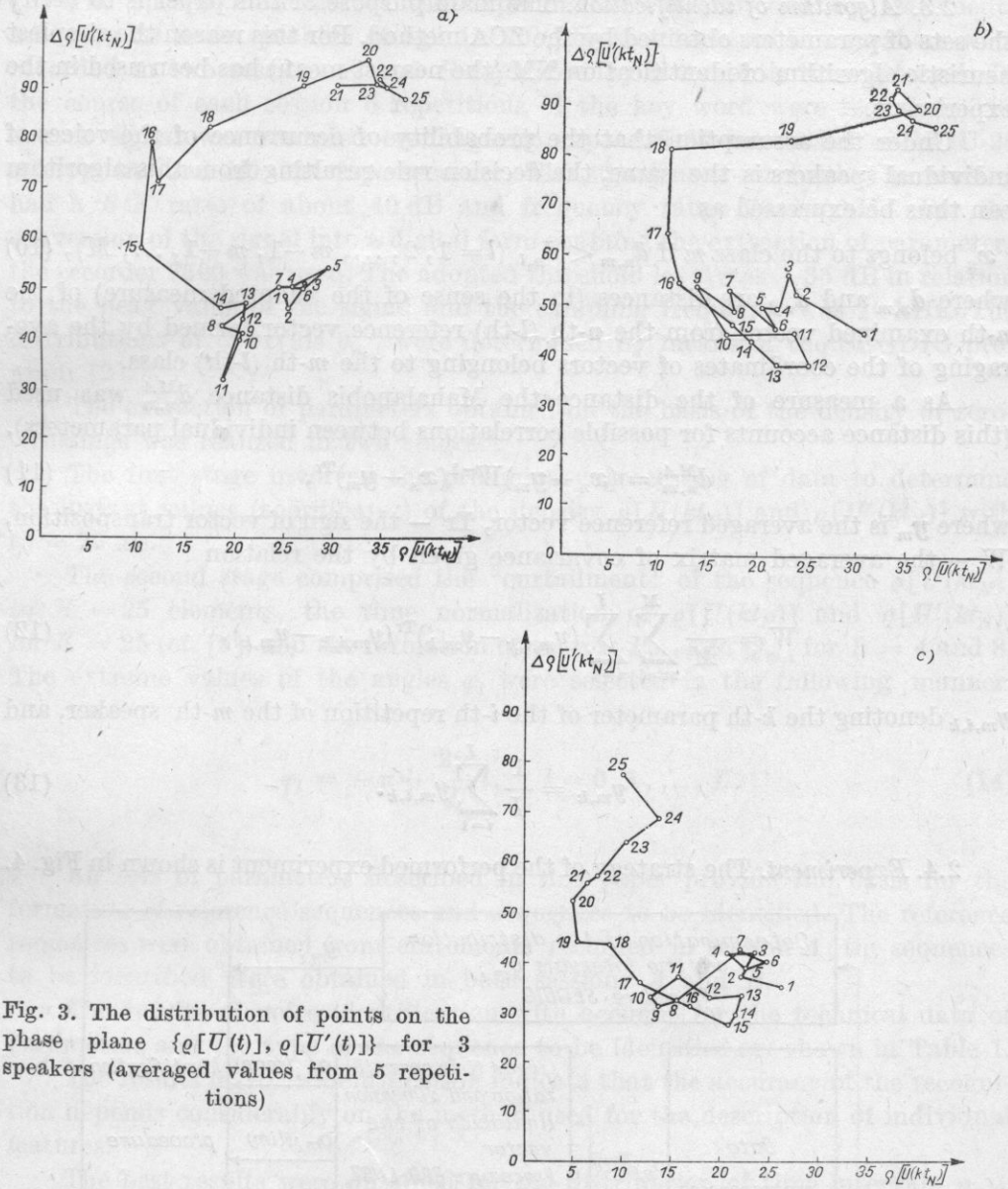

Fig. 3. The distribution of points on the phase plane $\{\varrho[U(t)], \varrho[U'(t)]\}$ for 3 speakers (averaged values from 5 repetitions)

where angles φ_l $(l = 0, 1, \ldots, L)$ define the boundaries of the sectors of the phase plane.

For the above-defined parameters a speech signal of i-th repetition of the m-th speaker can be presented as

$$\boldsymbol{Q}_{m,i} = \{Q_{m,i}(1), Q_{m,i}(2), \ldots, Q_{m,i}(L), S_{x_{m,i}}, S_{y_{m,i}}\}. \tag{9}$$

2.3. *Algorithm of identification.* The main purpose of this paper is to verify the sets of parameters obtained by the ZCA method. For this reason the simplest heuristic algorithm of identification NM (the nearest mean) has been used in the experiment.

Under the assumption that the probability of occurrence of the voices of individual speakers is the same, the decision rule resulting from this algorithm can thus be expressed as

$$\boldsymbol{x}_n \text{ belongs to the class } m \text{ if } d_{n,m} < d_{n,l} \quad (l = 1, 2, \ldots, m-1, m+1, \ldots, M), \tag{10}$$

where $d_{n,m}$ and $d_{n,l}$ are distances (in the sense of the adopted measure) of the n-th examined vector from the n-th (l-th) reference vector formed by the averaging of the coordinates of vectors belonging to the m-th (l-th) class.

As a measure of the distance the Mahalanobis distance $d_{n,m}^{MA}$ was used (this distance accounts for possible correlations between individual parameters),

$$d_{n,m}^{MA} = (\boldsymbol{x}_n - \boldsymbol{y}_m) W^{-1} (\boldsymbol{x}_n - \boldsymbol{y}_m)^{\mathrm{Tr}}, \tag{11}$$

where $\boldsymbol{y}_m$ is the averaged reference vector, Tr — the sign of vector transposition, W — the averaged matrix of covariance given by the relation

$$W = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{I} (y_{m,i,k} - y_{m,k})^{\mathrm{Tr}} (y_{m,i,k} - y_{m,k}), \tag{12}$$

$y_{m,i,k}$ denoting the k-th parameter of the i-th repetition of the m-th speaker, and

$$y_{m,k} = \frac{1}{I} \sum_{i=1}^{I} y_{m,i,k}. \tag{13}$$

2.4. *Experiment.* The strategy of the performed experiment is shown in Fig. 4.

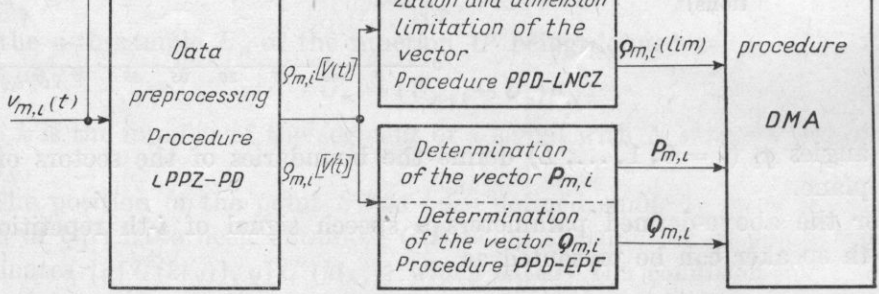

Fig. 4. The strategy of the identification experiment

The speech material for the experiments was provided by statements made by 20 male speakers of age 19-30. The statement of the speakers were recorded in two sessions A and B separated by a 3-month time interval. In the course of each session 5 repetitions of the key word were recorded. The speech signal was recorded on magnetic tape AN-25 by means of MDU-26 microphone and MP-224 tape recorder. The signal used for further processing had a S/N ratio of about 40 dB and frequency range 75-5000 Hz. For the conversion of the signal into a digital form enabling the extraction of parameters the recorder 7502 was used. The adopted threshold level was −35 dB in relation to the peak value of the signal and the sampling frequency was 10 kHz. The distributions of intervals $y_{m,i}$ were determined by means of the SEGDIG program [2].

The extraction of parameters obtained on the basis of the density of zero-crossings was realized in two stages.

The first stage involved the preliminary processing of data to determine the instant values (coordinates) of the density $\varrho[U(kt_N)]$ and $\varrho[U'(kt_N)]$ with $t_N = 20$ ms.

The second stage comprised the "curtailment" of the sequence $\varrho[U(kt_N)]$ to $K = 25$ elements, the time normalization of $\varrho[U(kt_N)]$ and $\varrho[U'(kt_N)]$ for $K = 25$ (cf. [5]) and the formation of vectors $\boldsymbol{P}_{m,i}$ and $\boldsymbol{Q}_{m,i}$ for $L = 4$ and 8. The extreme values of the angles φ_l were selected in the following manner:

$$\varphi_l = -\pi + \frac{2\pi}{L} l, \quad l = 0, 1, \ldots, L. \tag{14}$$

All sets of parameters described in this paper provide the basis for the formation of reference sequences and sequences to be identified. The reference sequences were obtained from statements recorded in session A; the sequences to be identified were obtained in both sessions A and B.

The results of voice identification with account for the technical data of parameters and the type of the sequence to be identified are shown in Table 1.

The results of voice identification indicate that the accuracy of the recognition depends considerably on the method used for the description of individual features.

The best results were obtained for the distribution of time intervals $\boldsymbol{y}_{m,i}$. In view of a not too large dimension K of this vector and no need for the time normalization, the priority should be given to this approach as being convenient and effective for the description of the reference pattern of voice obtained by the ZCA-method. The worse results were obtained for the sets $\boldsymbol{\rho}_{m,i}$ and $\boldsymbol{Q}_{m,i}$. Furthermore, prior to the formation of reference patterns and the patterns to be identified, the dimension of the set $\boldsymbol{\rho}_{m,i}$ must be reduced to the constant value K by normalization of "curtailment". The set $\boldsymbol{Q}_{m,i}$ for the adopted two

Table 1. Summary of the results of correct voice identification (in %)

Vector of parameters	Identified sequence from session *A*	from session *B*
$\boldsymbol{y}_{m,i}$ $K = 16$	92	78
$\boldsymbol{\rho}_{m,i}$ $K = 20$ (normalized)	90	72
$\boldsymbol{\rho}_{m,i}$ $K = 20$ (unnormalized)	89	68
$\boldsymbol{P}_{m,i}$ $L+2 = 6$	60	39
$\boldsymbol{P}_{m,i}$ $L+2 = 10$	63	49
$\boldsymbol{Q}_{m,i}$ $L+2 = 6$	77	48
$\boldsymbol{Q}_{m,i}$ $L+2 = 10$	84	61

dimensions of the vector $L+2$ is less efficient; for $L+2 = 10$ a considerably higher probability of correct identification was obtained and this gives evidence of the influence of the number of sectors of the phase plane on the results of identification for this set of parameters. Similar relations apply to the set $\boldsymbol{P}_{m,i}$ for which the least accurate results of identification were obtained.

For all sets of parameters a considerably smaller probability of correct identification in the case of sequences to be identified taken from session *B* was obtained, thus supporting the hypothesis of a pronounced effect of time lapse upon individual features of voices [1, 3].

On the basis of the results obtained it can be concluded that the ZCA--method can be effectively used for the formation of the sets of parameters discriminating the speakers in a short term analysis, provided a suitable method for the description of individual features of voice is used.

References

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

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

[3] R. Gubrynowicz, *Problem of the recognition of individual features of voice*, Prace IPPT PAN, 28 (1969) (in Polish).

[4] — *Zero-crossing method in the analysis of a speech signal and automatic recognition of a limited set of words*, Prace IPPT PAN, 37 (1974) (in Polish).

[5] A. PAWLAK, C. BASZTURA and W. MAJEWSKI, *Time normalization of a statement in the process of speaker identification*, Proceedings of 24. Opened Seminar on Acoustics, Gdańsk-Władysławowo, September 1977, 84-87 (in Polish).

[6] M. SAMBUR, *Selection of acoustic features speaker identification* IEEE, **ASSP 23**, no. 2, 176-182 (1975).

[7] D. A. WASSON and R. DONALDSON, *Speech amplitude and zero-crossings for automated identification of human speakers* IEEE, **ASSP 23**, 390-392 (1975).

Received on 28th December 1977