ARCHIVES OF ACOUSTICS
3, 1, 3-15 (1978)

THE APPLICATION OF LONG-TERM ANALYSIS OF THE ZERO-CROSSING OF A SPEECH SIGNAL IN AUTOMATIC SPEAKER IDENTIFICATION

CZESŁAW BASZTURA, WOJCIECH MAJEWSKI

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

This paper investigates the possibility of using long-term analysis of the zero-crossings of a speech signal for speaker identification. The applied method of identification is based on an analysis of the signal in measuring windows of a duration which should ensure the stationarity of the statical distributions of the time intervals between successive zero-crossings, in 16 pre-set time channels.

An objective method of defining the minimum length of the measuring window for a selected set of parameters is presented. It is based on the stationarity test and the ergodic theorem for stochastic processes, as is the transformation of the speech signal mentioned above. An experiment in speaker identification for 10 speakers with 10 repetitions for each speaker has been performed. The results obtained, in well exceeding 90% correct identification for 30- and 40 second samples of the speech signal, have confirmed the practicality of the method of zero-crossing analysis for speaker identification.

1. Introduction

The problem of speaker identification on the basis of the analysis of a speech signal still arouses the interest of scientists. The investigations in progress [3, 4, 5, 9, 10] are aimed at using such parameters for voice recognition as would be effective from the viewpoint of the storage of information on an individual, and at the same time being suitable for digital processing without the costly and complicated transformation of the speech signal.

Previous investigations [8] have shown that zero-crossing analysis is a method which can be used for speaker identification. It thus satisfies to some extent the first of the above-mentioned postulates.

In addition this method fully meets the other postulates. Its main advantages, i.e. ease of extraction and subsequent digital processing of a selected set of parameters, have determined that it, primarily, is used for speech analysis and recognition [2, 6].

The purpose of this paper is the investigation and explanation of certain problems that result from the use of the method of the analysis of the zero-

-crossings of a speech signal for speaker identification with the aid of a long-term analysis of this parameter.

Worthy of note is that the speaker identification methods can be divided generally into the methods based on the analysis of a short-term and long-term analysis of a speech signal.

The methods of short-term analysis are based on the individual parameters of the voice obtained from a established text in a time ranging from a fraction of a second for single phonemes, up to several seconds for sentences.

The methods based on long-term analysis are characterized by the fact that during the recognition process use is made of a set of parameters obtained from the speech signal which is of such a duration that the parameters may be considered to be stationary.

Advantages of the long-term methods of analysis are thus their lack of dependence on the text of a statement (i.e. utterance, pronouncement), the elimination of the associated problem of time normalization and also a high probability of correct speaker identification [4]. A real disadvantage, however, is the comparatively long duration of speaker statement necessitated. The problem of determining the minimum length of time for the long-term analysis of a speech signal is of considerable practical importance and constitutes one of the main aims of this paper.

2. The statistical distribution of the time intervals between the zero-crossings of a speech signal

Let $U(t)$ be the time function of a speech signal. If this signal is subjected to some amplification and infinitive peak clipping, then the resulting function $V(t)$ (Fig. 1) can be written in the form

$$V(t) = \text{sign}\ [U(t)]. \tag{1}$$

It can be seen from Fig. 1 that the positions of the zero-crossings after such a transformation of a speech signal remain unchanged under the assumption that the shaper does not introduce any distortion noise which may change the positions of the zeros.

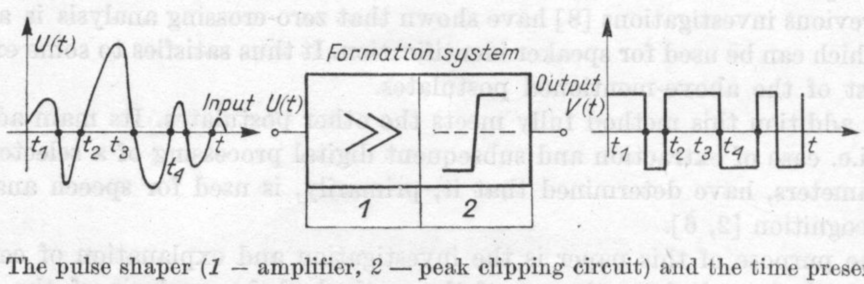

Fig. 1. The pulse shaper (*1* — amplifier, *2* — peak clipping circuit) and the time presentation of signals $U(t)$ and $V(t)$

For the signal $V(t)$ it is possible to define a function $R(t)$ giving the distribution of the time intervals between successive zero-crossings in a given signal segment of duration T_N

$$R(t) = \sum_{j=1}^{J} \delta(t-T_j), \tag{2}$$

where $\delta(x)$ is the Dirac delta function, $j = 1, 2, \ldots, J$ (J is the number of zero-crossings), T_j denotes the interval between a pair j-1 and j of zero-crossings in the segment T_N, with $T_N = \sum_{j=1}^{J} T_j$.

An example of the distribution of the function $R(t)$ is shown in Fig. 2. t_d and t_g are the limiting values of the time intervals for a given signal $V(t)$ in the measuring segment T_N.

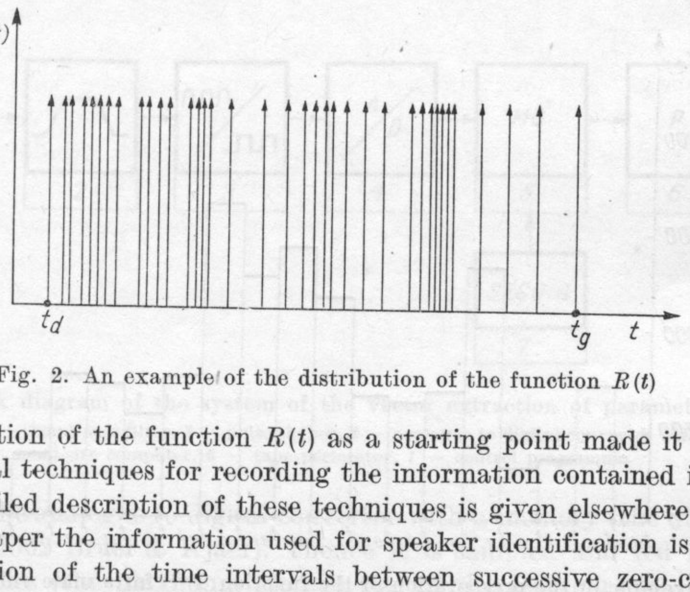

Fig. 2. An example of the distribution of the function $R(t)$

The adoption of the function $R(t)$ as a starting point made it possible to develop several techniques for recording the information contained in a speech signal. A detailed description of these techniques is given elsewhere [6].

In this paper the information used for speaker identification is the statistical distribution of the time intervals between successive zero-crossings in a measuring segment. If in the interval t_d to t_g the $K-1$ of the threshold values are distributed, then one obtains K time intervals called subsequently time channels, or more shortly, parameters. The signal representation mentioned above in the form of the distribution of the zero-crossing in the K time channels will be obtained by summing the number of intervals in suitably chosen ranges.

Let us denote these numbers by

$$y(t_d = t_0, t_1), \quad y(t_1, t_2,), \quad \ldots, \quad y(t_{k-1}, t_k), \quad \ldots, \quad y(t_{K-1}, t_g = t_K). \tag{3}$$

Between these numbers and the function $R(t)$ there the relationship

$$y(t_{k-1}, t_k) = \int_{t_{K-1}}^{t_K} R(t)\,dt. \tag{4}$$

The inclusion of the interval of length T_j in the k-th time channel agrees with the dependence

$$y(t_{k-1}, t_k) = \begin{cases} y(t_{k-1}, t_k)+1 & \text{for } T_j \in (t_{k-1}, t_k), \\ y(t_{k-1}, t_k) & \text{for } T_j \notin (t_{k-1}, t_k). \end{cases} \tag{5}$$

The combined function of the distribution of time intervals in the K time channels can be written in the form

$$Y(t) \stackrel{\text{def}}{=} \sum_{k=1}^{K} y(t_{k-1}, t_k)[1(t-t_{k-1})-1\ (t-t_k)], \tag{6}$$

where $1(t) = 0$ for $t < 0$ and $1(t) = 1$ for $t \geqslant 0$.

Fig. 3 shows an example of the distribution of time intervals as expressed by the function $Y(t)$.

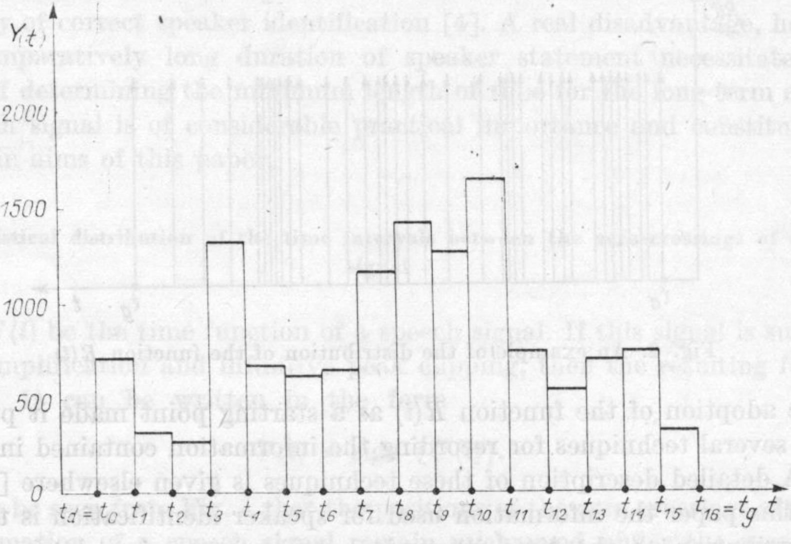

Fig. 3. An example of the presentation of the function $Y(t)$ for a male voice. Time analysis $T^s = 30$ s (the time axis is not graduated)

If the time representation $V(t)$ of duration T_N constitutes pattern (or a time segment of the pattern) of the m-th speaker and of the i-th repetition of the speaker's voice then after having obtained the distribution as a function $Y(t)$ this pattern can be represented in the form of a K-dimensional vector

$$\overrightarrow{y_{m,i}} = \operatorname{col}\{y_{m,i}, y_{m,1,2}, \ldots, y_{m,i,k}, \ldots, y_{m,i,K}\}, \tag{7}$$

where $m = 1, 2, \ldots, M$ (M denotes the number of the speakers), $i = 1, 2, \ldots, I$ (I is the number of repetitions for a speaker and is the same for each of them), $k = 1, 2, \ldots, K$ (K is the number of parameters, being equal to the number of time channels).

3. Phonetic material — the extraction of a set of parameters

The phonetic material for the experiments described in this paper were statements by 10 male speakers in the age range from 20 to 35. The statements of the speakers were recorded in two sessions A and B spaced by a 3 month time interval. In the course of each session for each speaker about 15-20 min. of text were recorded. In session A a newspaper text was recorded, and in session B the text from a popular scientific paper was recorded. The recordings were effected in an audio-monitoring studio. The speech signal was recorded on magnetic tape AN25 by means of microphone MDU26 and tape recorder MP224, produced by ZRK.

The extraction of parameters $\overrightarrow{y_{m,i}}$ was performed as shown in Fig. 4. In this system a signal from the tape recorder at a level of about 40 dB is fed through a band-pass filter (75-5000 Hz and 50 dB/octave) to the pulse shaper whence

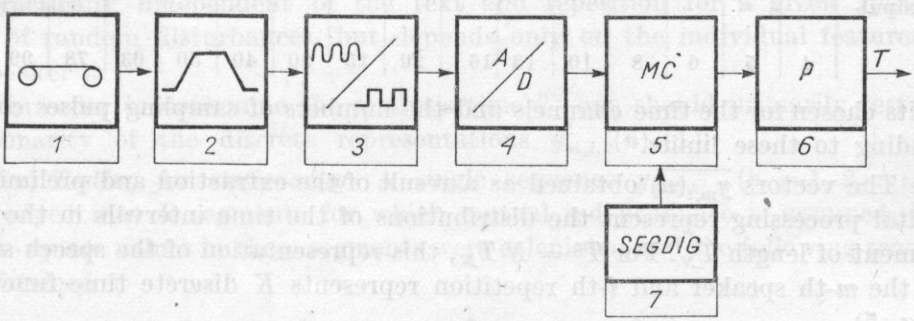

Fig. 4. Block diagram of the system of the vector extraction of parameters $\overrightarrow{y_{m,i}}$
1 — tape recorder, *2* — band-pass filter, *3* — pulse shaper, *4* — analogue to digital convertor with memory, *5* — small-size computer, *6* — tape perforator, *7* — control programme

it is passed to the analogue to digital convertor with a memory (the digital event recorder type 7502 Bruel & Kjaer). Thence it is sampled and fed in portions of 10 240 samples (the maximum storage size of the recorder) for programme processing through a small computer (type 7504 Varian/Bruel & Kjaer). The values of $\overrightarrow{y_{m,i}}$ are read out of the computer onto paper tape (TP).

The calculation of the values of the vectors $\overrightarrow{y_{m,i}}$ is controlled by the program SEGDIG (specially developed for this purpose) which can establish the number K of time channels (parameters) at will, and also their limiting values. The limiting values are given in the form of the numbers of the sampling pulses of the analogue to digital convertor. In the experiments performed, the sampling frequency was $f_p = 20\ 000$ samples/s, permitting the single recording of a signal over a time segment of length $T_N = 0.5$ s. With a view to the need for the extraction of $\overrightarrow{y_{m,i}}$ from time segments that are considerably longer (a subject that will be dealt with in the next section) on automatic system of random readout from the magnetic tape of the signal in a segment of duration $T_N = 0.5$ s was used. The sum of these windows gave the total time of analysis, T^s.

Preliminary results of experimental investigations permitted the determination of the extreme values t_d and t_g for the band 75 to 5000 Hz. This interval was then divided into 16 exponentially divided time channels with an accuracy given by the sampling frequency. Table 1 contains the data concerning the time

Table 1. Parameters of time channels

K	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
t_{k-1}[ms]	0.15	0.2	0.25	0.3	0.4	0.5	0.65	0.8	1.0	1.25	1.55	0.0	2.5	3.15	3.9	4.95
t_k [ms]	0.2	0.25	0.3	0.4	0.5	0.65	0.8	1.0	1.25	1.55	1.0	2.5	3.15	3.9	4.95	6.22
Sampled number of impulses: Bottom	3	4	5	6	8	10	13	16	20	25	30	40	50	63	78	99
Top	4	5	6	8	10	13	16	20	25	30	40	50	63	78	99	124

limits chosen for the time channels and the numbers of sampling pulses corresponding to these limits.

The vectors $\overrightarrow{y_{m,i}(n)}$ obtained as a result of the extraction and preliminary digital processing represent the distributions of the time intervals in the n-th segment of length T_N. For $T^s = N\,T_N$, this representation of the speech signal for the m-th speaker and i-th repetition represents K discrete time functions (Fig. 5).

Fig. 5. Examples of the presentations of the function $y_{m,i,k}(n)$

4. The choice of the time T^s for signal analysis

In many investigations the length of the total analysis time T^s for speaker identification was selected a priori by verifying only the correctness of the choice on the basis of the results of the attempted identification [3, 4, 8]. Such a selection is not, however, an optimal method. The assumption of too long a T^s, i.e. longer than is needed, unnecessarily extends the analysis time, while the selection of too short a segment increases the probability of identification error.

In this paper it is assumed that the distribution of the time intervals between successive zero-crossings of the speech signal measured in the total time of analysis

$$T^s \geqslant \max_{\substack{\wedge \\ m,i,k}} \{T^s_{m,i,k}\}, \tag{8}$$

where $T^s_{m,i,k}$ is the minimum time of the stationarity of $y_{m,i,k}(n)$, $n = 1, 2, \ldots, N$, is practically independent of the text and repetition for a given speaker, and of random disturbances, but depends only on the individual features of a speaker[1]).

In order to determine the analysis time T^s one should primarily test the stationarity of the discrete representations $y_{m,i,k}(n)$ [1].

(a) *Testing for stationarity.* A single sequence $y_{m,i,k(n)}$ $(n = 1, 2, \ldots, N)$ is grouped into R segments for which mutual independence is assumed. The mean-square values in these segments were calculated and the following sequence obtained

$$y^2_{m,i,k}(1), y^2_{m,i,k}(2), \ldots, y^2_{m,i,k}(r), \ldots, y^2_{m,i,k}(R). \tag{9}$$

Then the median of the mean-square values was calculated and the sequence (9) examined for the presence of the basic trend.

If the hypothesis concerning stationarity holds true, then changes in the sequence (9) will be of a random nature and will exhibit no trend.

To verify the stationarity a non-parametric test was performed at a level of significance of $\alpha = 0.05$. For the tested signals $y_{m,i,k}(n)$ it has been assumed that $T_N = 0.5$ s, $N = 100$ and $R = 20$. From the table containing the value of quantities i_p of the order $1-\alpha$ of the distribution of the number of series for the sequence $T = 2\ p = 20$ observations $i_{p,1-\alpha} = 6$ and $i_{p,\alpha} = 15$ [1] were read off. This series is the sequence of the values exceeding the median or possessing the values smaller than it which follows or precedes the other sequence.

For the assumption of the hypothesis of stationarity it is sufficient that

$$i_{p,1-\alpha} < i_{se} < i_{p,\alpha}. \tag{10}$$

---

(1) Refers to the independence, at a certain level of significance, at recording sessions not too distant in time.

In Table 2 examples of the values of the median for each time channel and a number of series for one voice are presented.

In a test carried out for two randomly selected repetitions and for each m and k the hypothesis of stationarity was confirmed.

Table 2. The value of the median and numbers of the series i_{se} for the stationarity test

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Median	250	39	70	300	195	183	58	217	320	300	405	870	590	450	372	19
i_{se}	7	9	7	11	9	13	9	9	12	14	14	10	11	11	8	11

(b) *Definition of* $T^s_{m,i,k}$. In a comparatively simple manner it is possible to define $T^s_{m,i,k}$ for a random stationary ergodic signal. The necessary and sufficient condition of ergodicity (in a broad sense) of a stationary random signal, such as $y_{m,i,k}(n)$, is that [1]

$$\frac{1}{S_{m,i,k}}\sum_{l=1}^{S_{m,i,k}}|C^{yy}_{m,i,k}(l)|\to 0 \quad \text{for } S_{m,i,k}\to\infty, \tag{11}$$

where $l = 1, 2, \ldots, S_{m,i,k}$, and $C^{yy}_{m,i,k}(l)$ is the autocovariance function (see Figs. 6 and 7).

For practical application formula (11) can be re-written in the form

$$\frac{1}{S_{m,i,k}}\sum_{l=1}^{S_{m,i,k}}|C^{yy}_{m,i,k}(l)|\leqslant \delta C^{yy}_{m,i,k}(0), \tag{12}$$

where δ denotes the assumed coefficient of deviation from the value $C^{yy}_{m,i,k}$ (0).

The fulfilment of condition (11) is sufficient for the ergodicity (in a broad sense) of the process $y_{m,i,k}(n)$ and confirms the hypothesis with a level of confidence dependent on the value of the coefficient δ, that is accepted.

Since no experiment can be optionally long, the fulfilment of condition (12) permits definition of the number of samples $S_{m,i,k}$ for practical purposes. This enables the calculation of a minimum time $T^s_{m,i,k}$:

$$T^s_{m,i,k} = S_{m,i,k}T_N. \tag{13}$$

Fig. 6 shows the network of operations of the algorithm PESP for calculating $S_{m,i,k}$ while Fig. 7 gives an example of a discrete autocovariance function.

Table 3 contains maximum values $T^s_{m,i,k}$, selected according to formula (8) from the population of 10 speakers, with 3 repetitions for each speaker and with 16 time channels.

It should be noted that the maximum value, $T^s_{m,i,k}$, exhibits a considerable variation with the different channels. It is highest for channels 12, 13, 14, i.e. the ones corresponding to smaller frequencies.

Fig. 6. The flow-chart of the algorithm PESP

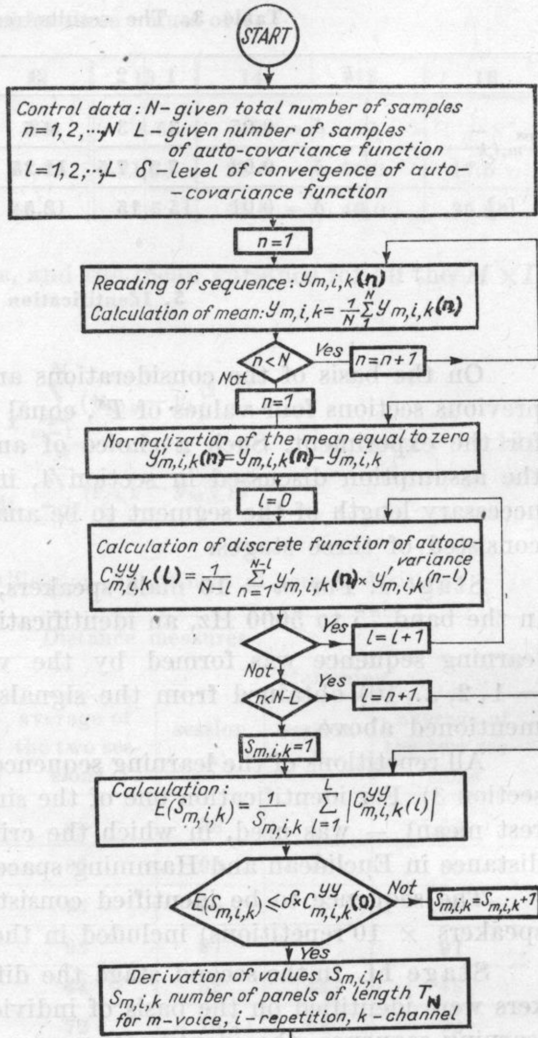

Fig. 7. An example of the presentation of the discrete autocovariance function $C^{yy}_{m,i,k}(l)$

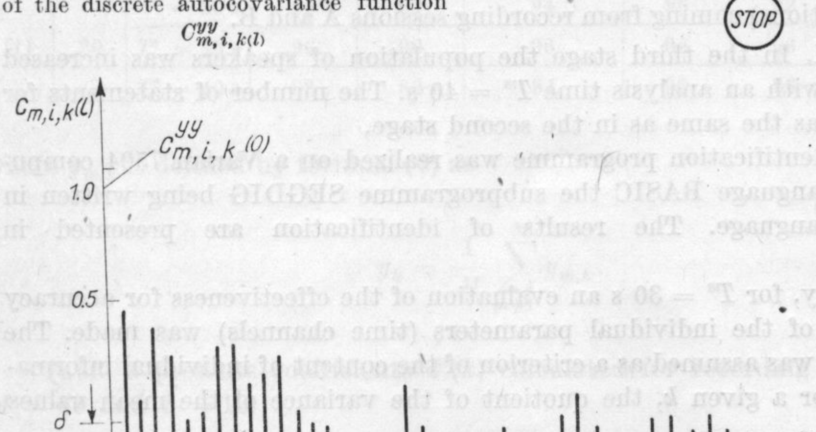

Table 3. The results results of measurements max $\{T^s_{m,i,k}\}$ for

	k	1	2	3	4	5	6	7
$T^s_{m,i,k}$	$\delta = 0.05$	3	3	4.5	6	6	5.5	5
	$\delta = 0.02$	7.5	7.5	11.25	15	15	13.75	12.5
[s]	$\delta = 0.01$	15	15	12.5	30	30	27.5	25.0

5. Identification experiment

On the basis of the considerations and measurement results presented in previous sections four values of T^s, equal to 10, 20, 30 and 40 s, were selected for the experiment. Such a choice of analysis times was intended to check the assumption discussed in section 4, in order to practically determine the necessary length of the segment to be analyzed. The identification experiment consisted of three stages.

Stage I. For $m = 10$ male speakers, $I = 10$ repetitions and for a signal in the band 75 to 5000 Hz, an identification test was carried out in which the learning sequence was formed by the vectors $\overrightarrow{y_{m,i}}$ $(m = 1, 2, \ldots, 10,\ i = 1, 2, \ldots, 10)$ obtained from the signals $V(t)$ for the 4 durations of signal mentioned above.

All repetitions of the learning sequence came from recording session A (see section 3). For identification one of the simplest heuristic algorithms NM (nearest mean) – was used, in which the criterion of decision was the minimum distance in Euclidean and Hamming space [4, 5].

The sequence to be identified consisted of 100 individual statements (10 speakers $\times$ 10 repetitions) included in the learning sequence.

Stage II. In the second stage the difference consisted only in that speakers were identified on the basis of individual statements, not included in the learning sequence, the identification sequences being 50 statements (10 speakers $\times$ 5 repetitions) coming from recording sessions A and B.

Stage III. In the third stage the population of speakers was increased to 20 persons, with an analysis time $T^s = 40$ s. The number of statements for each speaker was the same as in the second stage.

The NM identification programme was realized on a Varian 7504 computer using the language BASIC the subprogramme SEGDIG being written in the internal language. The results of identification are presented in Table 4.

Additionally, for $T^s = 30$ s an evaluation of the effectiveness for accuracy of recognition, of the individual parameters (time channels) was made. The coefficient $F(k)$ was assumed as a criterion of the content of individual information. $F(k)$ is, for a given k, the quotient of the variance of the mean values

$N = 100, L = 25, f_p = 20\,000$ samples/s and for three values of δ

8	9	10	11	12	13	14	15	16
5.5	7	6.5	7	8.5	8.5	9	8	—
13.75	17.5	16.25	17.5	21.25	21.25	22.5	10.0	17.5
27.5	35.0	32.5	35.0	42.5	42.5	45.0	40.0	35.0

obtained for the individual M speakers, and the mean variance for all the $M \times I$ voices and as given by the formula

$$F(k) = \frac{\dfrac{I}{M-1}\displaystyle\sum_{m=1}^{M}(y_{m,k}-y_k)^2}{\dfrac{I}{(I-1)M}\displaystyle\sum_{m=1}^{M}(y_{m,k}-y_{m,i,k})^2}, \tag{14}$$

Table 4. The results of speaker identification (in percent of correct decisions)

Stage	Number of speakers		Distance measures: Euclidean, session A	Euclidean, session B	Euclidean, average of the two sessions	Hamming, session A	Hamming, session B	Hamming, average of the two sessions
		$T^s = 10$ s	60	—	60	58	—	58
I	10	$T^s = 20$ s	89	—	89	90	—	90
		$T^s = 30$ s	97	—	97	96	—	96
		$T^s = 40$ s	98	—	98	97	—	97
II	10	$T^s = 10$ s	48	40	44	52	44	48
		$T^s = 20$ s	76	68	72	76	68	72
		$T^s = 30$ s	96	92	94	96	92	94
III	20	$T^s = 40$ s	96	96	96	98	96	97
		$T^s = 40$ s	88	80	84	92	86	89

where $y_{m,k}$ is defined by formula (7) as

$$y_k = \frac{1}{M}\sum_{m=1}^{M} y_{m,k}.$$

Table 5 presents coefficients $F(k)$ calculated for recording session A for $M = 10$ and $I = 10$.

It can be seen from Table 5 that the differences in the values of $F(k)$ are not too high and this gives evidence, to some extent, of an even distribution of individual information in selected time channels.

Table 5. The values of ability coefficients of parameters k

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$F(k)$	47.3	54.8	112.1	68.0	34.2	28.3	35.0	40.3	24.9	30.6	18.7	36.0	60.4	25.2	43.3	91.9

6. Discussion of results and conclusions

The results obtained of the identification (Table 4), although concerning not too numerous a population of speakers, confirm practicability of long-term analysis of the zero-crossings of a speech signal as a method for speaker identification. Especially promising are the results obtained for analysis times of 30 and 40 s which give considerably better than 90 % correct identification for $M = 10$, and are comparable to the results of other tests which use much more complicated methods.

An increased population of speakers ($M = 20$) brought about some decrease in the probability of correct identification.

Comparison of the results of identification achieved for various lengths of analysis times permits formulation of the thesis that 30 to 40 s of continuous speech signal from a newspaper text cam be accepted as the minimum time of long-term analysis for the calculation of an eventual distribution of time intervals between successive zero-crossings to be used for speaker identification.

In the case of essential differences from the number of time channels assumed in this paper, and the method of their division, differences from the required stated values $T^s = 30\text{-}40$ s may arise.

Comparing the values given in Table 3 with those in Table 4 a practical conclusion can be drawn that to determine the lengths of the time for analysis according to the method described in section 4 it is necessary to accept $\delta \leqslant 0.02$. If the condition for the stationarity of the distribution of time intervals is satisfied then there should be no essential difference in the results of identification of a particular sequence and whether or not it forms part of the learning sequence.

The results of Table 4 are in agreement with this theorem although to some extent the effect of the time lapse between recording sessions A and B can be seen. This is obvious because the individual features of the voice are not constant and change with time.

For a more comprehensive estimation of the practicality of the method of speaker identification presented consideration should be given to the restrictions assumed by the authors:

1. The experiment was carried out in laboratory conditions and thus the results of identification are independent of the influence of the technical conditions of the recording [7].

2. The extraction of parameters was performed over a frequency band of speech signal from 75 to 5000 Hz. Neither the number of the channels nor the limit of the time channels were optimized, an exponential division within the interval t_d to t_g being assumed a priori.

3. The experiment was carried out with the cooperation of the speakers, i.e. the speakers did not try during the recordings to change the manner of their pronunciation or to imitate the voices of other speakers.

References

[1] J.S. Bendat, A.G. Piersol, *Random data: Analysis and measurement procedures,* Wiley, 1971.

[2] R. Gubrynowicz, *The zero-crossing method in speech signal analysis, and the automatic identification of a limited set of words,* Reports IPPT PAN, No 37/1974 [in Polish].

[3] K-P. Li, G.W. Hughes, *Talker differences as they appear in correlation matrices of continuous speech spectra,* Journ.Acoust.Soc.Am., **55**, 4, 833-837 (1974).

[4] W. Majewski, H. Hollien, *Euclidean distance between long-term speech spectra as a ctriterion for speaker identification,* Proc. Speech Communication Seminar, Stockholm 1974, Vol. 3, pp. 303-310.

[5] W. Majewski, Cz. Basztura, H. Hollien, *Speaker identification by means of statistical distributions of zero-crossing rates,* 8 Int.Con.Pho.Sci., Abstracts of papers, Leeds 1975 (186).

[6] R. Niederjohn, *A mathematical formulation and comparison of zero-crossing analysis techniques which have been applied to automatic speech recognition,* IEEE Trans. on Acoustic Speech and Signal Proc., Vol. ASSP-23, 4, 373-380 (1975).

[7] E. Ozimek, *The effect of the external conditions of a recording on the possibility of identifying the individual features of a voice,* Archiwum Akustyki, 4, **5**, 389-409 (1970) [in Polish].

[8] G.S. Ramiszwili, *Ob awtomaticzeskom uznawanii golosow,* Techniczeskaja Kibiernietika, 4, 87-93 (1966).

[9] M. Sambur, *Selection of acoustic features for speaker identification,* IEEE Trans. on Acoustics, Speech and Signal Proc., ASSP, **23**, 2, 176-182 (1975).

[10] A. Wasson, R. Donaldson, *Speech amplitude and zero-crossing for automated identification of human speakers,* IEEE Trans. on Acoustics Speech and Proc., ASSP, **23**, 4, 390-392 (1975).

Received on 25th May 1977