ARCHIVES OF ACOUSTICS
23, 4, 463–479 (1998)

PARAMETRICAL PATTERNS OF VOICES IN FORENSIC APPLICATIONS

W. MAJEWSKI and CZ. BASZTURA

Institute of Telecommunications and Acoustics
Wrocław University of Technology
(50-370 Wrocław, Wybrzeże Wyspiańskiego 27, Poland)

A method of speaker recognition based on visual comparisons of sets of voice patterns is presented. These sets contain graphical patterns obtained from parameter extraction in the time-domain, frequency-domain and LPC (Linear Predictive Coding). In comparison to classical spectrograms, the parametrical patterns permit a better adaptation to specific speech signals that are available in forensic applications. Since parametrical patterns are more explicit than spectrograms, the decision reached by an expert may be more objective. The laboratory experiments on visual comparison of selected sets of patterns permitted to observe that in parametrical patterns of voices the intraspeaker variations are generally smaller than the interspeaker ones. The presented method of speaker recognition has been verified in a real-life case yielding valuable information on particular speakers. The results of laboratory experiments and real-life applications indicate that the visual method of speaker recognition based on various parametrical patterns of selected speech segments constitutes a very useful tool supporting the evidence proceeding with a much higher confidence level than it could be obtained on the basis of the classical spectrograms.

1. Introduction

Individual voice features contained in a speech signal permit to recognise speakers from their utterances. The speaker recognition may be carried out by means of subjective or objective methods. The subjective methods, that may be either aural-perceptual or visual, are based on subjective judgements, whereas the objective methods encompass all automatic approaches to the speaker recognition where the decisions are made by a computer.

Recognition of speakers, based on their utterances, have many different applications, from among which the most important ones are the forensic applications. Forensic applications of speaker recognition require a high level of reliability that is difficult to achieve by means of a single method of speaker recognition because of the unfavourable conditions (different environmental and transmission conditions for the evidence and reference speech samples, uncooperative speakers, voice disguise, etc.) typical of this type of applications. To solve this problem a system consisting of a parallel application of an aural-perceptual procedure and a visual and automatic one has been proposed [3]. The present paper is focused on the visual procedure of the speaker recognition.

The visual method of speaker recognition is based on comparisons of sets of patterns found in the utterances produced by the examined speakers. Usually spectrograms (sonagrams) are utilised for this purpose. Since a speech pattern depends on both the speaker and the speech content, the patterns under comparison have to be obtained from the same utterance for all the examined speakers. It has to be remembered that in forensic applications the patterns are influenced also by different transmission and recording conditions for the evidence and reference speech samples and by the voice disguise.

The effectiveness of this method of speaker recognition is strongly influenced by the qualification and experience of the expert. Moreover, a considerable confusion exists with respect to the effectiveness of visual speaker recognition based on spectrograms [1, 2, 4]. One of the possible reasons of this controversy is the fact that the spectrograms contain too many details that are hard to follow. Our idea is to include to the visual inspections sets of patterns obtained from parameter extraction in the frequency-domain, time-domain or LPC (Linear Predictive Coding) and presented as parameters averaged for particular utterances or presented in the function of time. Since parametrical patterns are more explicit than classical spectrograms, the decisions reached by the expert should be more reliable and objective. To test the usefulness of parametrical patterns of voices for visual speaker recognition in forensic applications was the main purpose of the present study.

2. Parametrical patterns of voices

The parameters that are to be applied for speaker recognition should be efficient in representing the individual voice features (i.e. they should vary maximally for different speakers and minimally for a given speaker), should be easy to measure and stable over time, occur frequently in speech, should be resistant to speaking and transmission conditions and not susceptible to mimicry [5]. It is not possible, however, to find parameters that fulfil simultaneously all those conditions, and – in practice – a large variety of speech parameters is utilized. The parameters most widely used include: the speech spectrum, formant frequencies, the fundamental frequency, vocal intensity, LPC parameters, cepstrum coefficients and zero-crossing rates (ZCR). For the purpose of the present study the parameters presented below have been chosen.

For recording the speakers' voices a laboratory recording system consisting of an IBM PC computer, microphone and an acoustic input/output with AD/DA card was used.

The preliminary speech signal coming from a microphone has been low-pass filtered with a 4.5 kHz cut-off frequency, sampled at a rate of 10 kHz and digitized with a 12-bit resolution. A Hamming window of the coefficients $\alpha = 0.54$ and $\beta = 0.46$ was superimposed onto the signal. The window length was $N = 256$ samples with a 128 samples shift. Thus, the parameter vectors were taken from 25.6 ms frames.

2.1. *Spectral parameters*

Description and analysis of the speech in the frequency domain permits to present the utterances in a very useful and often applied form. It results from the fact that the changes in the excitation function and the shape of the vocal tract occurring during speech production are significantly reflected in the spectral parameters' values which are very important for speech perception because the human ear acts like a frequency analyser.

The relation between the frequency domain and time representation of a signal is given by the Fourier transform. FFT algorithms given, by G.E. BERGLAND and M.T. DOLAN [6], were used to calculate the short-term speech spectra. These algorithms satisfy the equation:

$$F(k) = \sum_{n=0}^{N-1} u(n) e^{-j\frac{2\pi}{N}nk}, \tag{1}$$

where $u(n)$ – digital input sequence of the speech signal, $F(k)$ – complex coefficients of FFT, $k = 0, 1, 2, ..., N-1$.

Before FFT was calculated, the signals were standardized to the energy equal to unity.

A commonly applied method of spectral parametrization is the mean spectrum in one-third octave bands measured for stationary speech segments or particular key words. The vectors with components representing the mean amplitude spectra in $P = 16$ one-third octave bands were used as spectral parameters of the utterances under investigation. For each frequency band the components $F(p)$ were calculated according to the formula:

$$F(p) = \frac{1}{K_p - P_p} \sum_{j=P_p}^{K_p} \left\{ [\operatorname{Re} F(j)]^2 + [\operatorname{Im} F(j)]^2 \right\}, \tag{2}$$

where $P_p = f_p \cdot N/f_s$, $K_p = f_{p+1} \cdot N/f_s$, f_p – boundary frequency between $(p-1)$ and p frequency band, f_s – signal sampling frequency.

2.2. *The density of zero crossings (ZCR-density)*

Directly from the time course of the speech signal the parameters useful for speaker recognition may be extracted. They concern the analysis of the speech signal zero crossings. The results of this analysis may be expressed either as the density of zero crossings (ZCR-density) or as the distribution of time intervals between successive zero crossings (ZCR-distribution).

The mean density of zero crossings ϱ_0 of the speech signal $u(t)$ in the time interval T_p is expressed by the formula:

$$\varrho_0[u(t), T_p] = \frac{1}{T_p} \frac{\int\limits_{-\infty}^{\infty} f^2 P(f)\, df}{\int\limits_{-\infty}^{\infty} P(f)\, df}, \tag{3}$$

where f – frequency, $P(f)$ – power density spectrum.

The measurement of the density $\varrho_0[u(t), T_p] = \varrho_0(p)$ from a speech signal in the digital form $u(n)$ is made according to the relation:

$$\varrho_0(p) = \frac{1}{T_p} \sum_{n=1}^{N} C\left\{u(n), (k-1)T_p + \frac{n}{f_s}\right\}, \tag{4}$$

where

$C(u,t) = 1$, if there exists signal $u(t) = u(n)$ satisfying the conditions (a)–(c), and

$C(u,t) = 0$, if there is no signal $u(t) = u(n)$ satisfying the conditions (a)–(c).

(a) $u(n)u(n-k) < 0$,

(b) $|u(n)| \geq \alpha$ and $|u(n-k)| \geq \alpha$,

(c) $|u(l)| < \alpha$ for $n-k < l < n$,

and α is a threshold level ($\alpha \neq 0$) which prevents from counting additional zero crossings caused by disturbances, p is the index of a signal segment of duration $T_p = N/f_s$.

2.3. Distribution of time intervals between zero crossings (ZCR-distribution)

The moments of successive zero crossings in the speech signal are detected and the lengths of time intervals between these crossings calculated. The intervals are then grouped and ordered with respect to their lengths. They are classified to successive groups according to previously determined time channels. The number of intervals in the individual channels may be determined:

$$x(p) = \{x(t_0, t_1), x(t_1, t_2), ..., x(t_p, t_{p+1}), ..., x(t_{P-1}, t_P)\}, \tag{5}$$

where t_p – threshold values between the channels, P – number of time channels.

An interval with length t_j is classified to the time channel p in accordance with the dependence:

$$x(p) = x(t_{p-1}, t_p) = \begin{cases} x(t_{p-1}, t_p) + 1 & \text{for} \quad t_j \in (t_{p-1}, t_p], \\ x(t_{p-1}, t_p) & \text{for} \quad t_j \notin (t_{p-1}, t_p], \end{cases} \tag{6}$$

where $x(t_{p-1}, t_p)$ – number of time intervals in a given time channel.

The cumulative distribution of time intervals in the P time channels can be presented as:

$$x(p) = \sum_{p=1}^{P} x(t_{p-1}, t_p) \cdot [\mathbf{1}(t - t_{p-1}) - \mathbf{1}(t - t_p)], \tag{7}$$

where

$$\mathbf{1}(t) = \begin{cases} 0 & \text{for} \quad t \leq 0, \\ 1 & \text{for} \quad t > 0. \end{cases}$$

The boundary values of the time channels for the speech signals are $t_{\min} = 0.2$ ms and $t_{\max} = 6.2$ ms. If in such a range $P-1$ intermediate threshold values are placed, P time channels will be obtained. In the experiments $P = 16$ time channels with logarithmic distribution of length were applied.

2.4. Linear predictive coding (LPC) parameters

The algorithms given by J.D. Markel and A.H. Gray [6] have been used to calculate the LPC parameters. Linear prediction models the signal spectrum by means of an all-pole filter with the transfer function:

$$H(z) = \frac{G}{1 - \sum\limits_{p=1}^{P} a_p z^{-p}}, \tag{8}$$

where G – gain factor, a_p – prediction coefficient, P – prediction order, z – operator of the **Z** transform.

The Levinson–Durbin [7] recursion was utilized to calculate the prediction and reflection coefficients. It is a recursive-in-model-order solution for the autocorrelation equations applied to the window $n \in (t - N + 1, t)$.

Initialization: $p = 0$

$$E^{(0)(t)} = R(0;t) \tag{9}$$

scaled total energy in the "error" from an "order 0" predictor = average energy in the speech frame $h(n)$ $\{u(n)h(t-n)\}$.

Recursion: For $r = 1, 2, 3, ..., P$,

1. Compute the reflection coefficient,

$$k(r;t) = \frac{R(r;t) - \sum\limits_{p=1}^{r-1} a^{(r-1)}(r;t) \cdot R(r-p;t)}{E^{(r-1)}(t)}. \tag{10}$$

2. Generate the order-r set of LPC parameters,

$$a_{r(r;t)} = k(r;t), \tag{11}$$

$$a_p(p;t) = a^{(r-1)}(p;t) - k(r;t) \cdot a^{(r-1)}(r-p;t), \qquad 1 \le p \le r-1. \tag{12}$$

3. Compute the error energy associated with the order-r solution,

$$E^{(r)}(t) = [1 - k(r;t)]^2 \cdot E^{(r-1)}(t). \tag{13}$$

4. Return to Step 1 with r replaced by $r + 1$ if $r < P$.

In the experiments $P = 12$ prediction coefficients a_p averaged over the utterances were used.

2.5. Amplitude envelope

The amplitude envelope $E(t)$ may be calculated by the following formula:

$$E(t) = \sqrt{u^2(n) + \mathbf{H}^2[u(n)]}, \tag{14}$$

where **H** is the Hilbert transform:

$$\mathbf{H}[u(n)] = \sum_{-N/2}^{N/2} \frac{u(k)}{\Pi(n-k)}. \tag{15}$$

In addition to the parameters presented in points 2.1–2.5, two additional speech signals representations, i.e. the waveforms and the digital spectrograms, were used in the experiments.

3. Experiments under laboratory conditions

The group of speakers consisted of 10 Polish adult males with no speaking defects who uttered two isolated Polish words: "Awek" */avek/* and "logarytm" */logaritm/*. The number of repetitions for each word and each speaker was 20. All the utterances were produced during a single session and recorded under normal laboratory conditions. The recorded speech samples were converted to a digital form and fed into the computer. Each speech sample was labelled by the following code: U0SS01RR, where "U" is the first letter of the utterance, "SS" stands for the speaker number and "RR" for the repetition number.

A specially designed computer program enable to perform and visualize the results of the following speech signal analyses:

- waveforms,
- spectrograms,
- mean spectrum in 16 one-third octave bands,
- distribution of time intervals between zero crossings in 16 time channels,
- linear prediction coefficients of 12-th order,
- amplitude envelope with the printed values of time (first number) and amplitude (second number) of local maxima,
- density of zero crossings with the printed values of local maxima (marking the same as above).

The graphical patterns resulted from the performed analyses were used to evaluate the intra- and interspeaker variations of the speech samples. As an example, the patterns obtained for three repetitions of the word "logarytm" produced by seven speakers are presented in Figs. 1–7. In Fig. 1 the waveforms are presented. As may be seen, the intraspeaker variations are quite small. The interspeaker variations are generally larger, but they depend upon the speakers under comparison – in some cases they are very large (e.g. the waveforms for speaker no 7 do not resemble the waveforms for any of the other six speakers) and in other cases they are comparable to the intraspeaker variations (see e.g. the patterns for the speakers no 3 and 5). Thus, these direct representations of the speech samples provide limited information on intra- and interspeaker variations. In Fig. 2 digital spectrograms are presented. The general conclusion is similar: the intraspeaker variations are rather small, the interspeaker variations are generally larger, but the large number of details cause that the similarities and differences between particular patterns are somewhat obscured. The remaining figures (Fig. 3 – mean spectra, Fig. 4 – distributions of time intervals, Fig. 5 – LPC coefficients, Fig. 6 – amplitude envelopes, Fig. 7 – zero crossing densities) concern parametrical patterns that are much simpler in their graphical form than spectrograms and the comparisons between those patterns may be more easily performed. The mean spectra presented in Fig. 3 are very

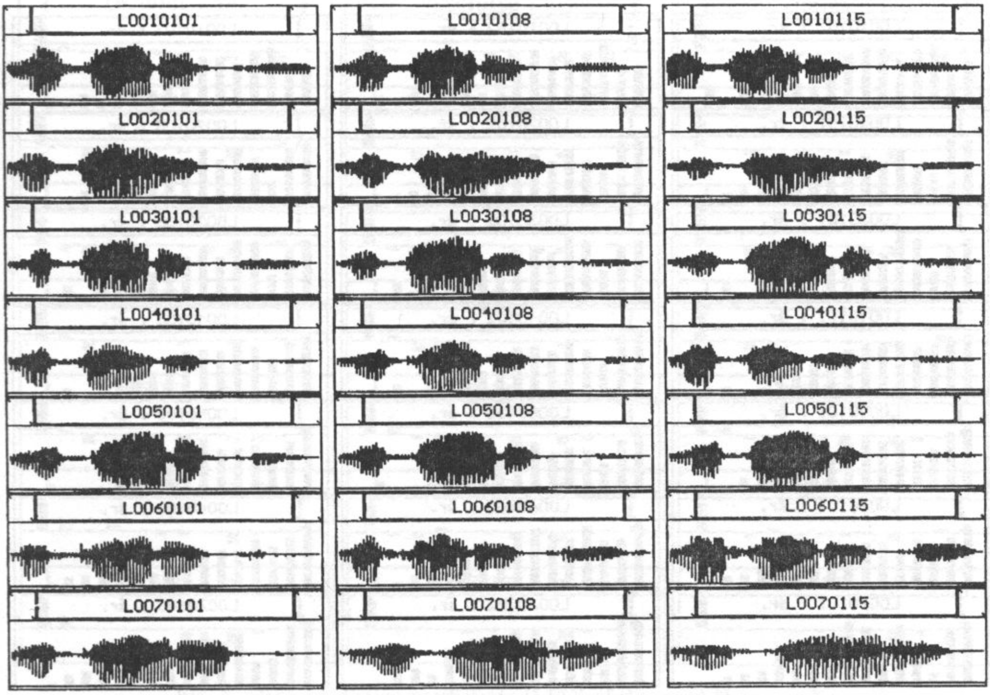

Fig. 1. The waveforms of the word "logarytm" produced three times by seven speakers.

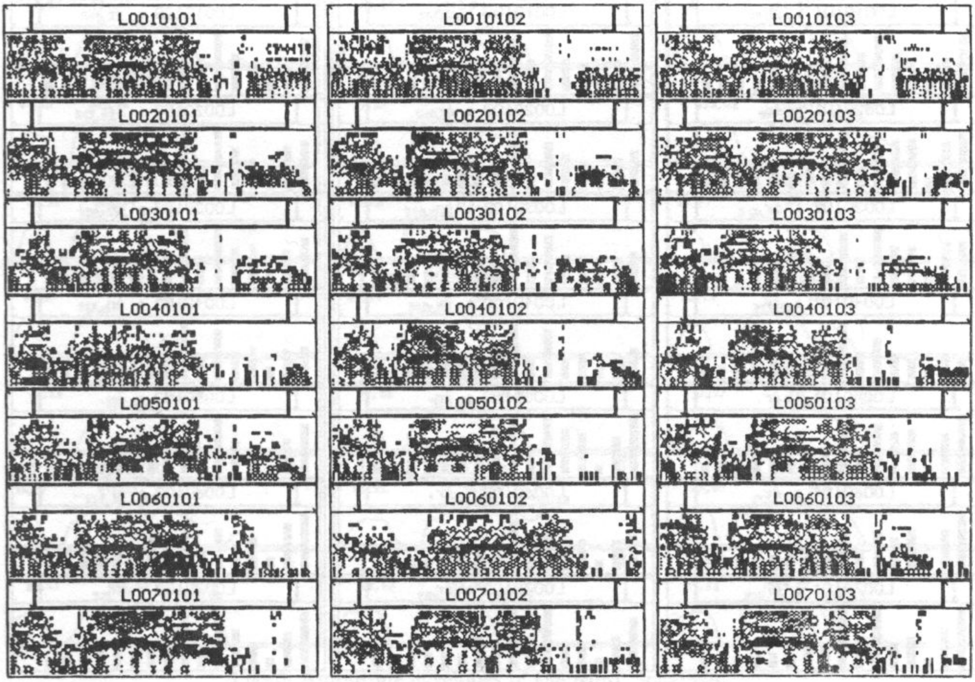

Fig. 2. Digital t-f-a spectrograms of the word "logarytm".

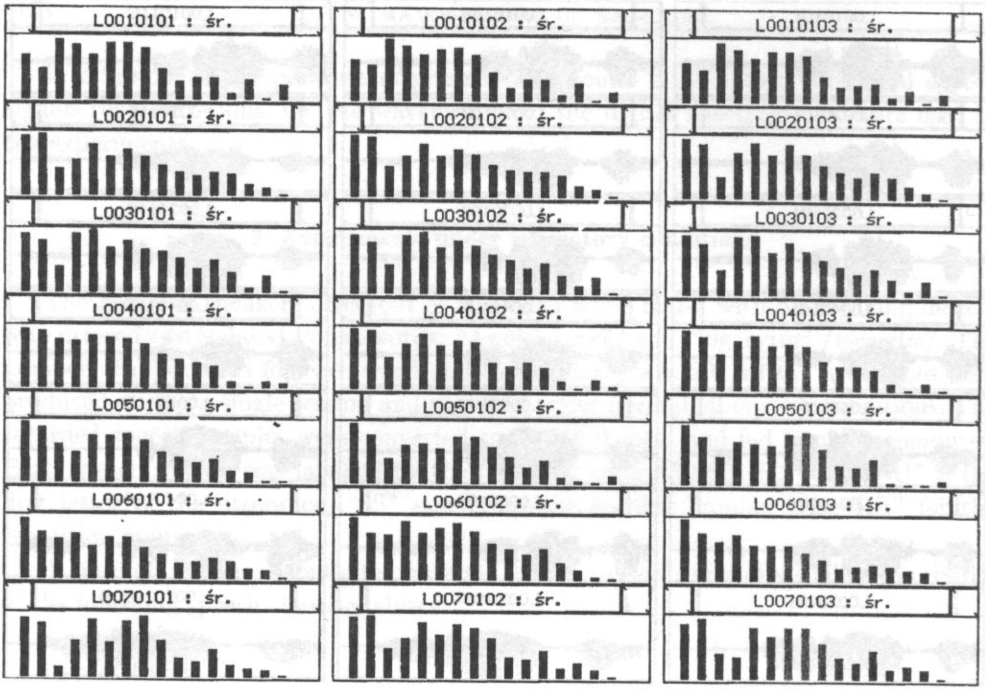

Fig. 3. Mean spectra of the word "logarytm".

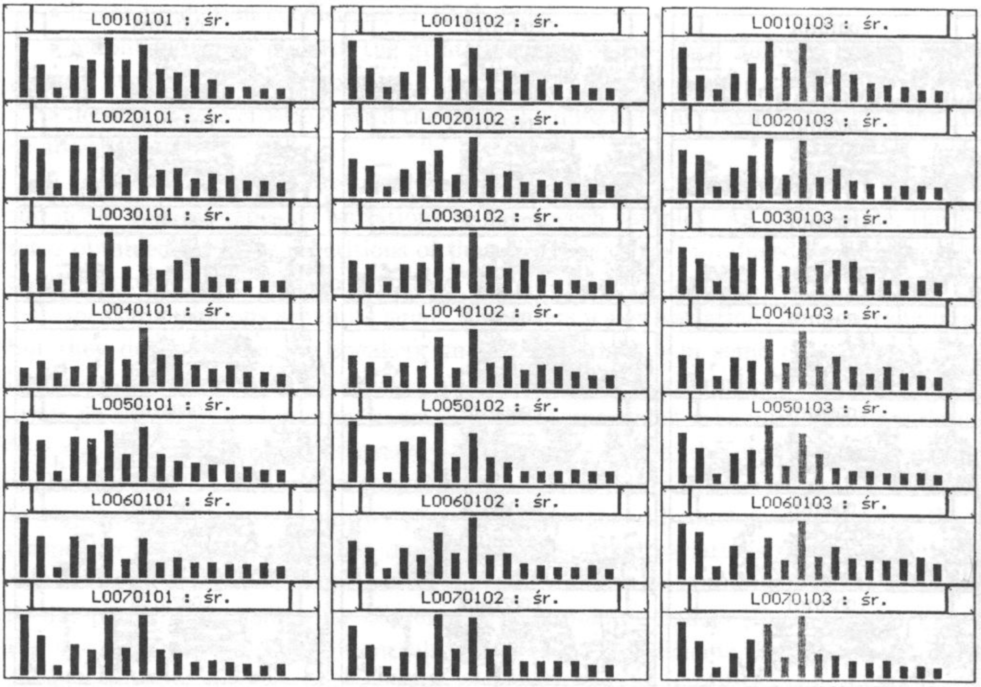

Fig. 4. Distributions of time intervals between zero crossings (ZCR-distributions) of the word "logarytm".

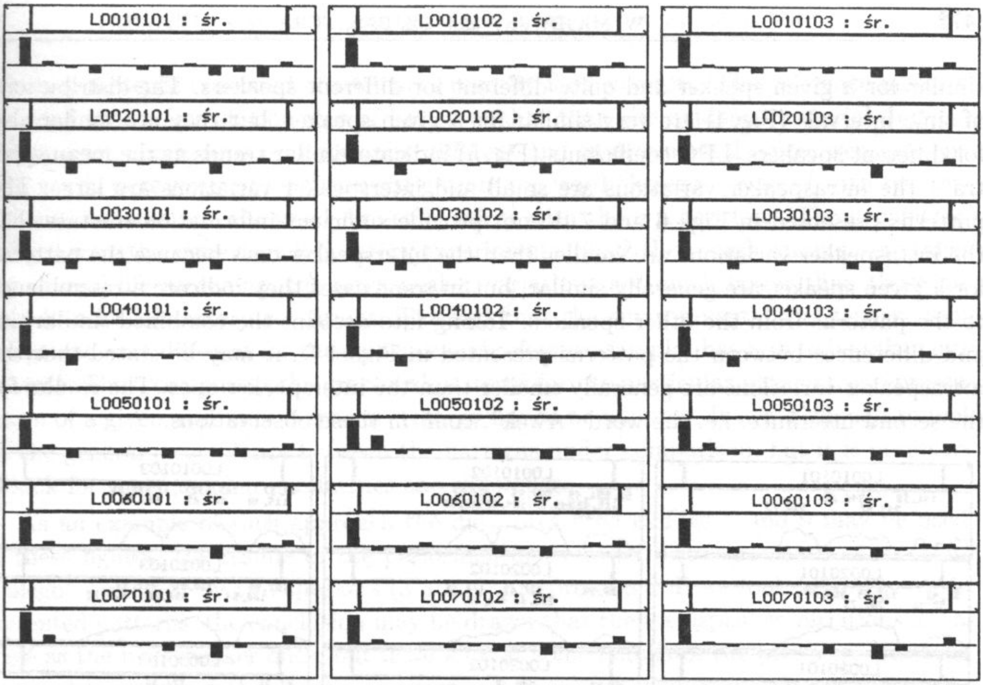

Fig. 5. Mean prediction coefficients (LPC) of the word "logarytm".

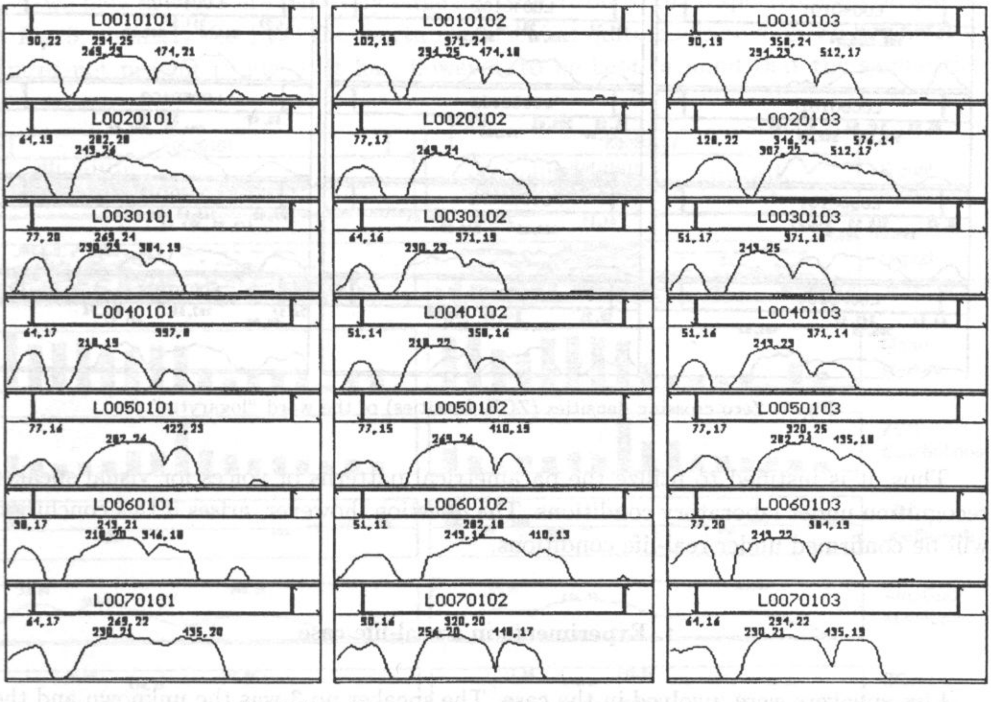

Fig. 6. Amplitude envelopes of the word "logarytm".

similar for a given speaker and quite different for different speakers. The distributions of time intervals (Fig. 4) are very similar for a given speaker, but they are similar also for different speakers. LPC coefficients (Fig. 5) indicate similar trends as the mean spectra – the intraspeaker variations are small and interspeaker variations are large. The patterns presented in Figs. 6 and 7 do not provide sufficient information to state that the intraspeaker variations are smaller than the interspeaker ones because the patterns for a given speaker are generally similar, but in some cases they indicate a resemblance to the patterns from the other speakers. Taking into account the combined similarities and differences between the patterns presented in Figs. 1–7, it may be stated that the intraspeaker variations are generally smaller than the interspeaker ones. The results for the second utterance, i.e. the word "Awek", confirm these observations.

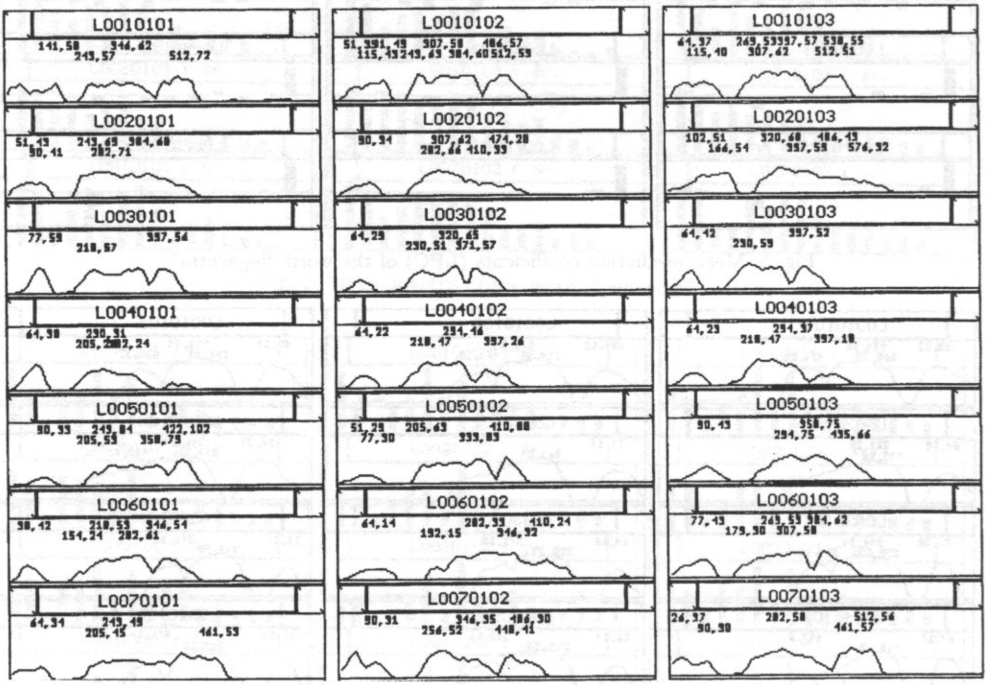

Fig. 7. Zero crossing densities (ZCR-densities) of the word "logarytm".

Thus, it is justified to utilize the parametrical patterns of voices for visual speaker recognition under laboratory conditions. The question, however, arises if this conclusion will be confirmed under real-life conditions.

4. Experiments in a real-life case

Five speakers were involved in the case. The speaker no 3 was the unknown and the speakers marked with numbers 4–7 were the suspects. The evidence material produced by the unknown consisted of telephone calls recorded on a cassette tape recorder. The reference material consisted of sentences read by four suspects. The reference material recorded in a police station was of better quality than the evidence material, but the transmission characteristics were quite different and environmental noises were present. Two words were selected for the investigation: the word "kolego" */colego/* with two repetitions available and the word "waler" */valer/* with three repetitions. The parameters utilized were the same as in laboratory experiments. Since the speech sample were taken from a real-life recording, the intraspeaker variations are much larger than those observed under laboratory conditions. The fact that the environmental and transmission conditions were different for the evidence and reference samples has also to be taken into account. Thus, comparing the graphical patterns for the particular parametric representation of a given utterance a special strategy has to be adopted. It is not enough to look for the general resemblance between the patterns under comparison, but it is necessary to look for some peculiar similarities and differences.

As an example of such approach the data presented in Figs. 8 and 9 may be used. In these figures, the results of the performed analyses for two repetitions of the word "kolego" produced by two speakers (3 and 4) are presented. If we look generally at the presented patterns, the conclusion may be drawn that the intraspeaker variations are as large as the interspeaker ones, but if we look at some details of the presented patterns, it may be confirmed that both repetitions were produced by respective speakers. For example: if we compare all four patterns of the mean spectra, it may be found that the last two bars in Fig. 8 are relatively large in comparison to the respective bars presented in Fig. 9; similarly, for LPC the fourth coefficient is quite distinctive in Fig. 8 and it is almost not present in Fig. 9. It has, however, to be kept in mind that the similarities

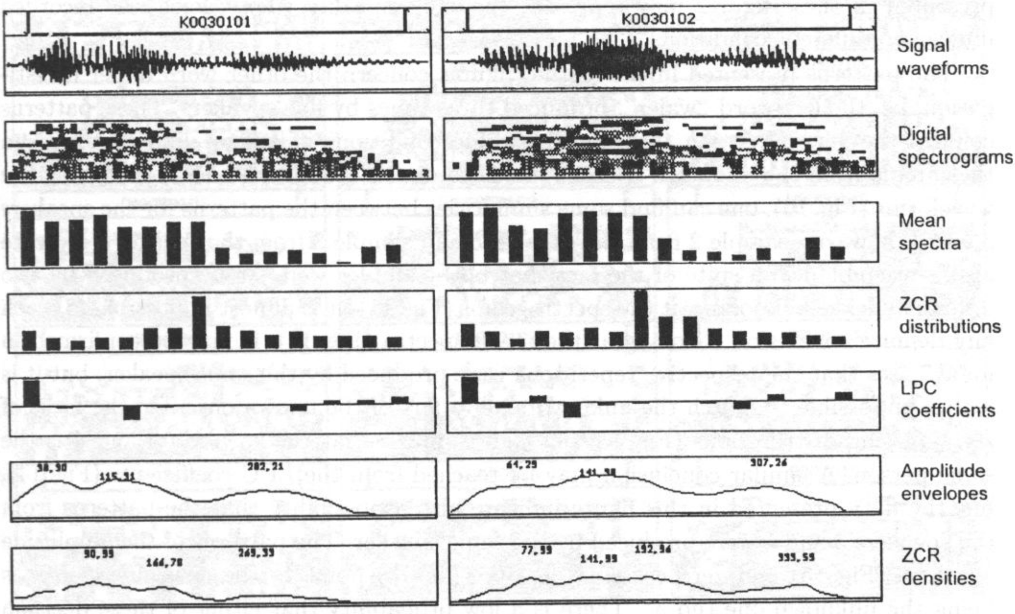

Fig. 8. Seven parametrical representations of the word "kolego" produced two times by speaker no 3.

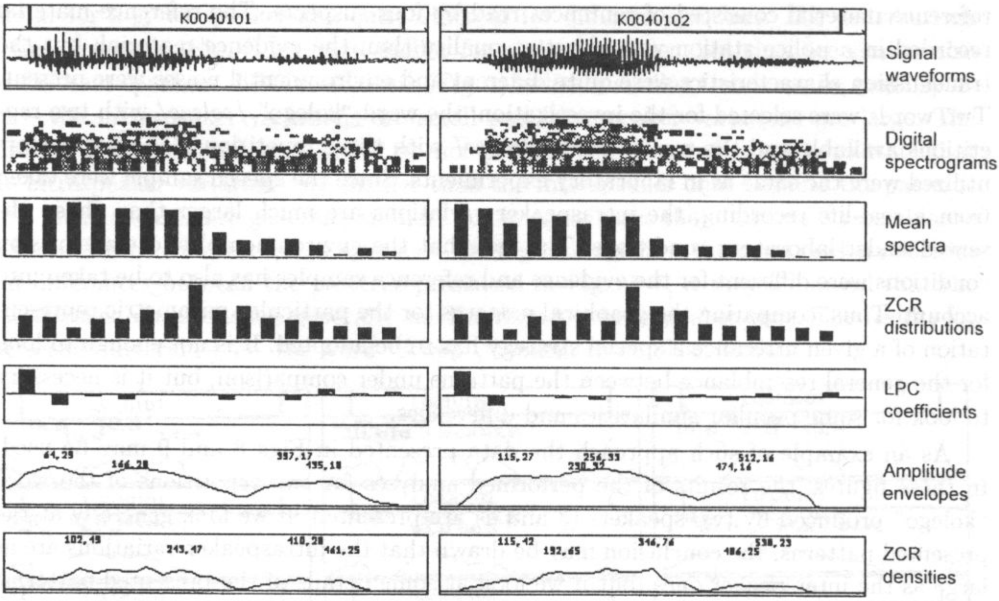

Fig. 9. Seven parametrical representations of the word "kolego" produced two times by speaker no 4.

and differences between the patterns presented in Figs. 8 and 9 may result only from different transmission and recording conditions. Since the speaker no 3 was the unknown and the speaker no 4 was one of the knowns, it is possible that both sets of the patterns presented in these figures may represent the same speaker whose voice was recorded under two different conditions.

The patterns presented in the further figures concern the other word under investigation, i.e. the key word "waler", produced three times by five speakers. These patterns confirm the general notion that under real-life conditions it is very difficult to assign the sample from the unknown speaker (no 3) to one of the known ones. Comparing the waveforms (Fig. 10), one can find some similarities between the patterns for the speakers 3 and 4, however, sample 2 from the speaker 5 and sample 3 from the speaker 6 indicate also a resemblance in spite of the fact that both samples were surely produced by two different speakers. Looking at the spectrograms (Fig. 11), it is almost impossible to reach any definite conclusion. Looking at the mean spectra (Fig. 12), it may be confirmed in many cases that the respective repetitions were produced by the same speaker, but it is rather impossible to match the unknown one to any of the known ones. On the basis of zero crossing distributions (Fig. 13) one may suppose that the speakers 3 and 4 is the same person. A similar conclusion may be reached from the LPC coefficients (Fig. 14), but the data presented in this figure indicate also a possibility that the patterns from the speakers 3 and 5 were produced by the same speaker. The patterns of the amplitude envelope (Fig. 15) and zero crossing density (Fig. 16) point to the speakers 4 or 5 as being the unknown one (no 3). There is a low probability that either of these decision is correct since generally most of the patterns presented in these figures are similar.

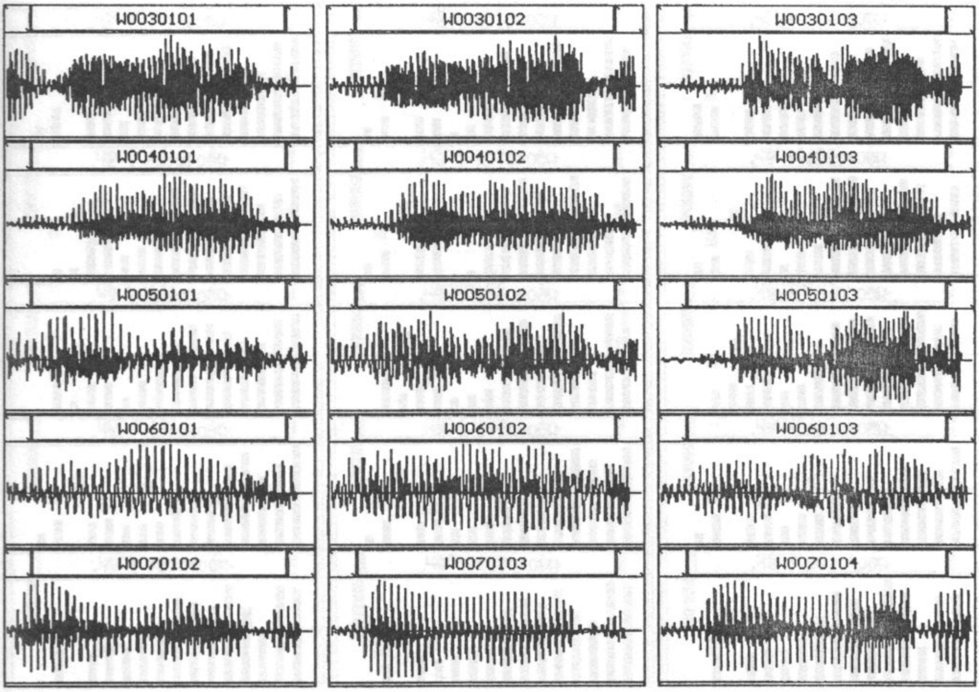

Fig. 10. The waveforms of the word "waler" produced three times by five speakers.

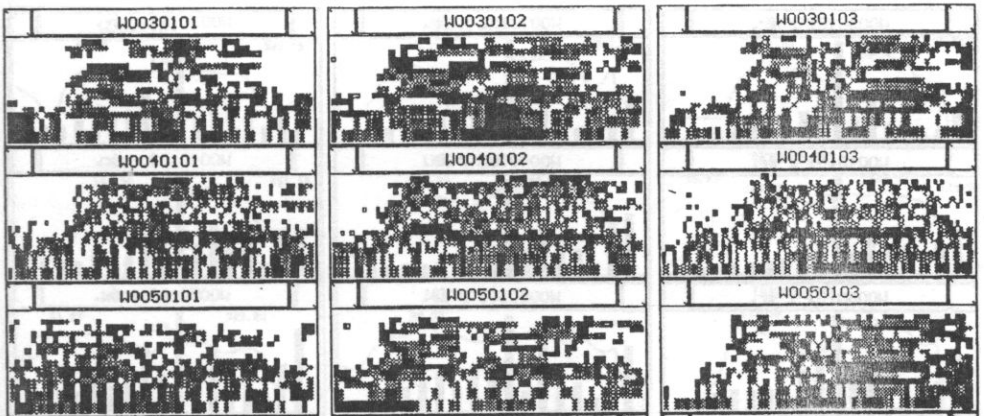

Fig. 11. Digital t-f-a spectrograms of the word "waler".

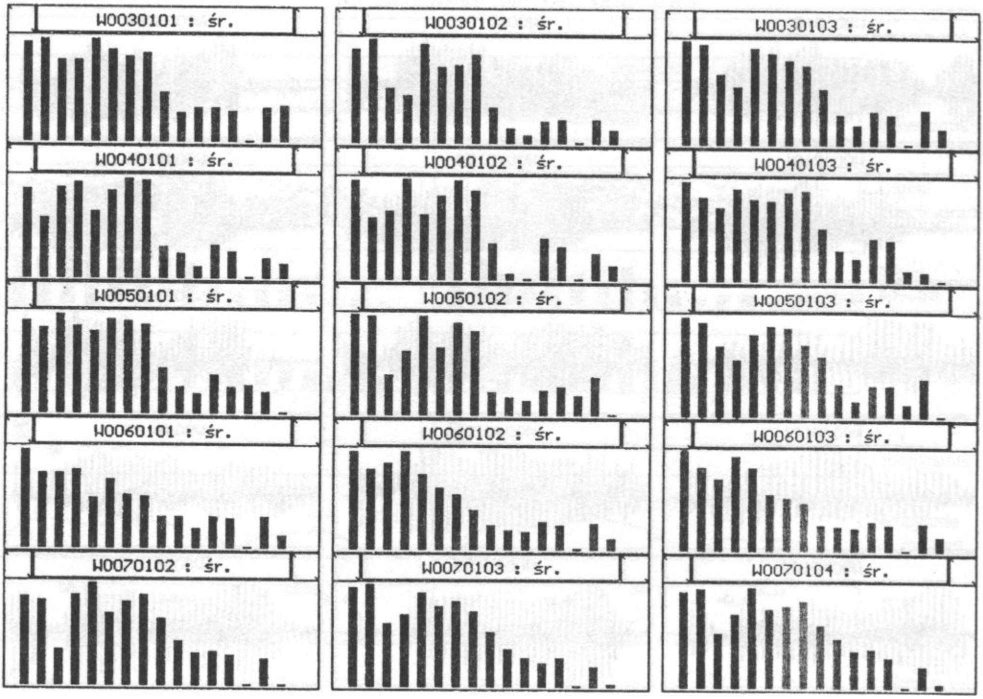

Fig. 12. Mean spectra of the word "waler".

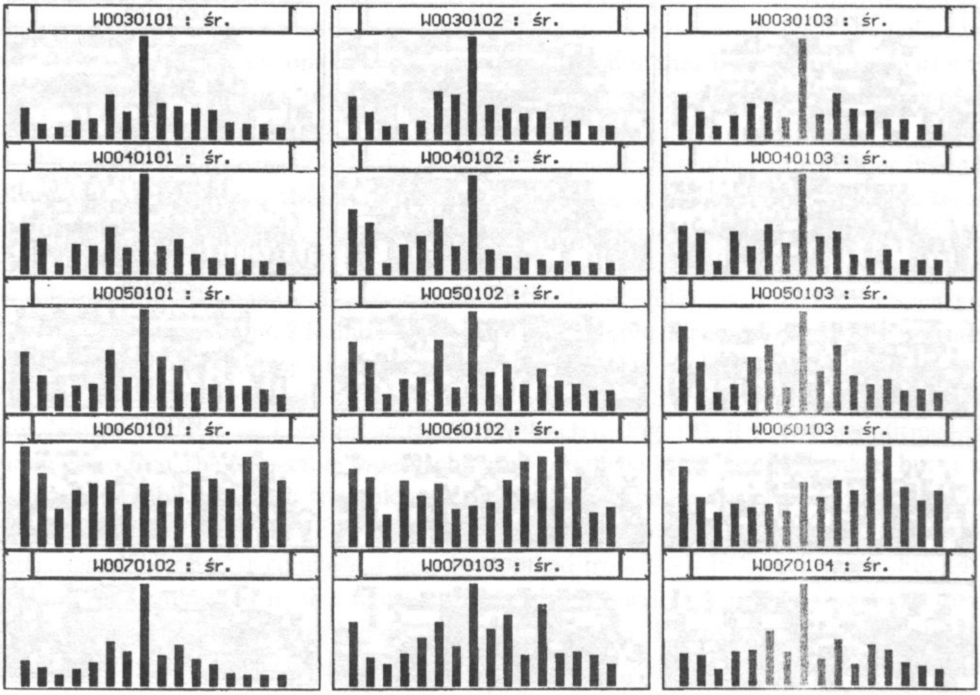

Fig. 13. ZCR-distributions of the word "waler".

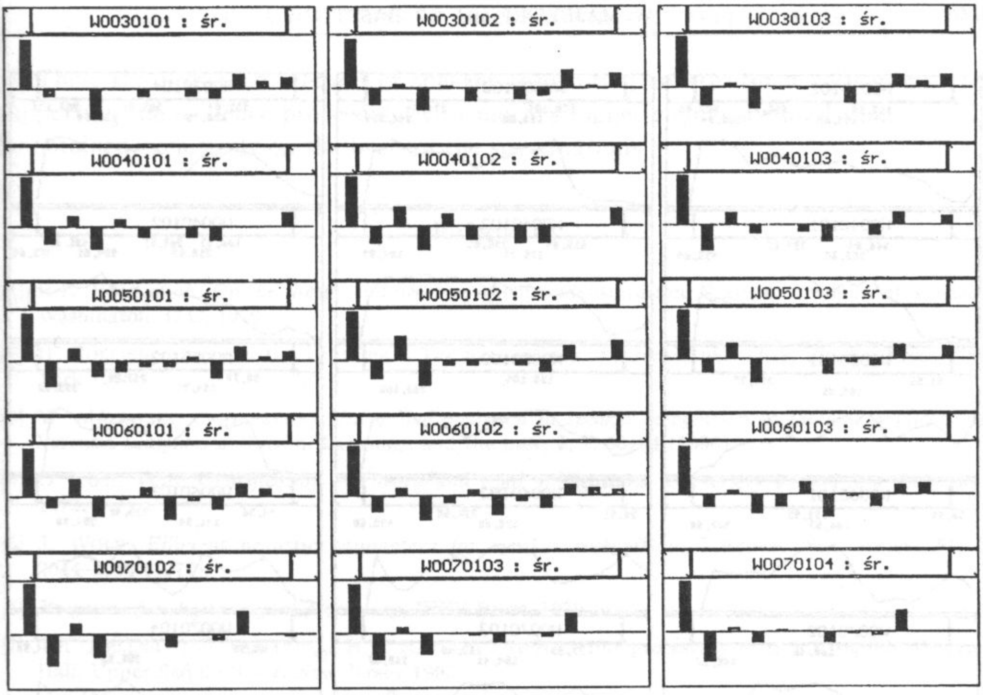

Fig. 14. LPC coefficients of the word "waler".

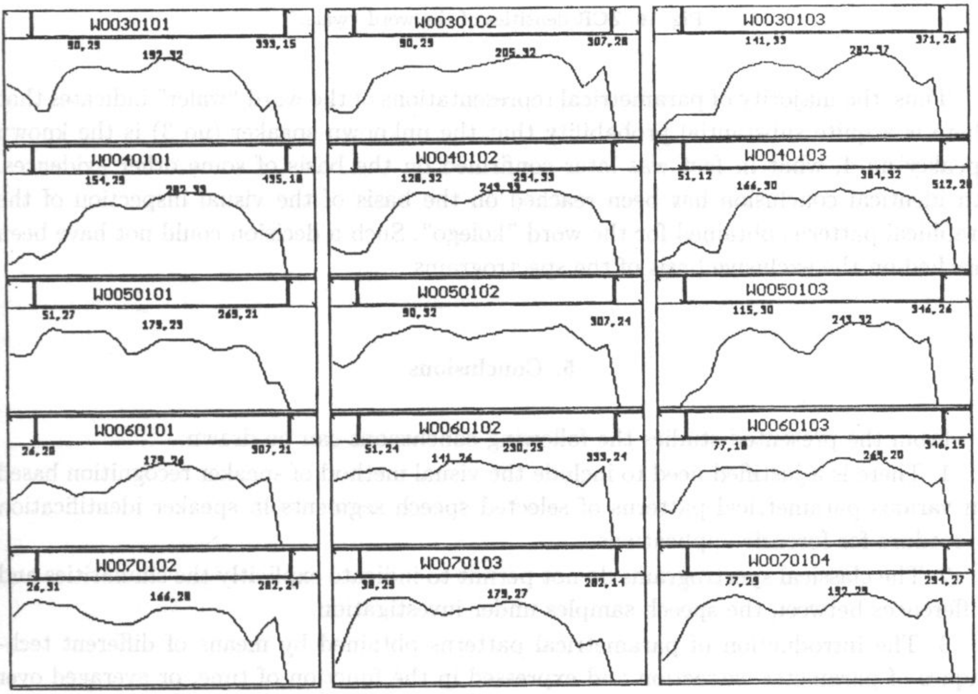

Fig. 15. Amplitude envelopes of the word "waler".

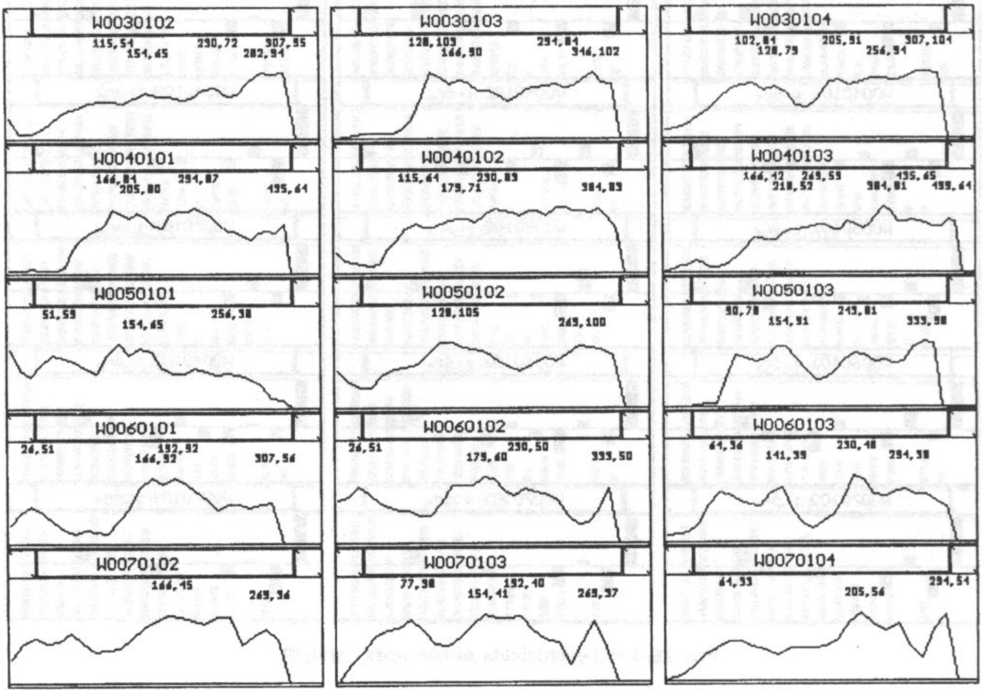

Fig. 16. ZCR-densities of the word "waler".

Thus, the majority of parametrical representations of the word "waler" indicates that there is a quite substantial probability that the unknown speaker (no 3) is the known speaker no 4, what in fact was later confirmed on the basis of some other evidences. An identical conclusion has been reached on the basis of the visual inspection of the graphical patterns obtained for the word "kolego". Such a decision could not have been reached on the exclusive basis of the spectrograms.

5. Conclusions

From the presented studies the following conclusions can be drawn:

1. There is a justified need to include the visual method of speaker recognition based on various parametrical patterns of selected speech segments in speaker identification procedure for forensic applications.

2. The classical spectrograms do not permit to indicate explicitly the similarities and differences between the speech samples under investigation.

3. The introduction of parametrical patterns obtained by means of different techniques of parameter extraction and expressed in the function of time, or averaged over time, permit to visualize substantial similarities and differences in the utterances under investigation that are not visible in the classical spectrograms.

Thus, the presented method of speaker recognition constitutes a very useful tool supporting the evidence proceeding with much a higher confidence level than it could be obtained exclusively on the basis of the spectrograms.

References

[1] R.H. Bolt *et al.*, *On :he theory and practice of voice identification*, National Academy of Sciences, Washington, D.C. 1979.

[2] H. Hollien, *The acoustics of crime – The new science of forensic phonetics*, Plenum Press, New York 1990.

[3] W. Majewski, C. Basztura, *Integrated approach to speaker recognition in forensic applications*, Forensic Linguistics: Speech, Language and the Law, **3**, 1, 50–64 (1996).

[4] O. Tosi, *Voice identification – Theory and legal applications*, University Park Press, Baltimore 1979.

[5] J. Wolf, *Efficient acoustic parameters for speaker recognition*, J. Acoust. Soc. Amer., **51**, 6, 2044–2056 (1972).

[6] *Programs for digital signal processing*, IEEE Press, 1979.

[7] J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete-time processing of speech signals*, Prentice Hall, Upper Saddle River, New Jersey 1993.