Ewelina MAJDA, Andrzej P. DOBROWOLSKI[1]

Military University of Technology, Faculty of Electronics (1)

Modeling and optimization of the feature generator for speaker recognition systems

Abstract. *This paper presents issues related to modeling and optimization of the feature generator for the speaker recognition system (ASR – Automatic Speakers Recognition). The parameterization stage of generating a speech signal (features generation) is fundamental in this type of system because the unique vector of features is crucial in the process of speech recognition. The task is to describe the speech signal using as few descriptors as possible without loss of relevant information for speaker recognition. In addition, the parameterization should be robust to acoustic and technical registration conditions and to the recorded linguistic material. The research presented in this paper focused primarily on the multicriteria optimization of selected parameters of the feature generator based on cepstral analysis that additionally enables the selection of features. Finally, the evaluation of the results was based on the analysis of the main components of a set of descriptors for the samples of voice acquired from 24 speakers.*

Streszczenie. *W referacie przedstawiono zagadnienia związane z modelowaniem i optymalizacją generatora cech dla systemu automatycznego rozpoznawania mówcy (ang. Automatic Speaker Recognition – ASR). Etap generacji cech (parametryzacji sygnału mowy) jest fundamentalny w tego typu systemach, z uwagi na fakt, że unikatowy wektor cech ma decydujące znaczenie w procesie rozpoznawania. Zadaniem generatora cech jest opisanie sygnału mowy za pomocą możliwie mało licznego zbioru deskryptorów, bez utraty informacji istotnych z punktu widzenia rozpoznawania mówcy. Ponadto parametryzacja powinna wykazywać odporność na warunki akustyczne i techniczne rejestracji oraz na zawartość lingwistyczną rejestrowanego materiału. Badania przedstawione w referacie koncentrowały się przede wszystkim na wielokryterialnej optymalizacji wybranych parametrów generatora cech opartego na analizie cepstralnej, uwzgledniającej dodatkowo selekcję cech. Oceny otrzymanych wyników dokonano w oparciu o analizę składników głównych (ang. Principal Component Analysis – PCA) zbioru deskryptorów wyznaczonych dla próbek głosu pochodzących od 24 mówców. (**Modelowanie i optymalizacja generatora cech dla systemu rozpoznawania mówcy**).*

Keywords: automatic speaker recognition, feature extraction, features selection, principal component analysis.
Słowa kluczowe: automatyczne rozpoznawanie mówcy, ekstrakcja cech, selekcja cech, analiza składników głównych.

Introduction

Speech is a natural and one of the most effective means of communication between humans. Automatic speech recognition has a variety of technical solutions. The common feature of these solutions is the processing of the speech signal using a digital device to extract the required information for specific applications. This paper presents a procedure for processing a speech signal to identify the speaker.

Automatic recognition-of-voice/sensing-of-voice is divided into two fundamentally different procedures: identification and verification of the speaker. *Identification* of a speaker is a process of decision-making, which involves confirmation of the identity of the speaker and is based only upon the characteristics of the speech (without declaring his/her identity). On the other hand, *verification* of a speaker is a process of decision-making, using the characteristics of the speech signal to determine whether the speaker is, in fact, the person whose identity he/she declares. The outcome of the verification of a speaker is confirmation or refusal of the claimed identity.

A very important characteristic of speakers recognition systems is their dependence on the recognized text spoken by a person, that is, the limitations imposed on the linguistic material. Speaker recognition systems can be further divided into text-dependent and text-independent tasks. In text-independent systems, the linguistic content of the training and test material is generally the same. Text-independent systems do not require the use of specific words to perform recognition tasks. Sentence tests can be differentiated from sentence learners, at least in the order of words. In this case the speaker can be identified regardless of the language of expression [1].

The automatic recognition of speakers identifies/verifies a person based on the comparison between the *attribute vector* and the database of registered voices models. Figure 1 shows an exemplary diagram of such a system. Analysis of the speech signal to obtain the features vector carrying information about the individual characteristic of the voice of the speakers (voice model) can be performed in two modes: learning or identification.

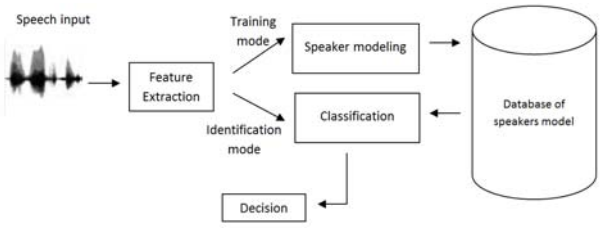

Fig. 1. Diagram of the speakers identification procedure

In the learning mode, a new speaker with a known identity is enrolled in the database of the system. In the identification mode, a speaker is identified based on the comparison of the extracted unique features of the unknown speaker's voice with the samples from the database of system (classification). Both the learning and the identification phase use the same algorithms of speech signal parameterization defining a unique features vector, known as the "*voice print*".

Characteristics of the phenomena associated with generating the speech

The communications process using speech involves the generation and reception of acoustic stimuli. The organ of speech is a specialized system that enables the generation of a wide range of sounds. The speech system controls the air stream flowing out of the lungs, allowing the encoding of useful information in the form of instantaneous changes in pressure. Aside from information about the content of speech, any speech signal also carries information related to the internal structure of its source. These inter-individual differences reflect the individual characteristics of the speaker's voice. The characteristics result from the differences in the construction of the organ of articulation (voice path) in different people, the habits acquired when learning to speak and the degree of mastery of a given language.

Speech sounds are produced in the organ of speech, of which the essential elements are the lungs, trachea, larynx, throat, nose, nasal cavity and mouth. The part of the organ of speech lying above the larynx is called the voice channel. The shape of its cross-section may vary considerably due to movements of the tongue, lips and jaw (the organs of articulation), allowing the pronunciation (articulation) of different sounds. An essential element of the larynx for generating speech is the vocal folds (cords). The space between these vocal folds is called the glottis. Vocal folds can open and close, affecting the air flow to the lungs. The sound produced during the escape of air from the lungs through the vocal folds that perform quick movements (periodic or quasi-periodic), e.g., the closing and opening of the glottis, is called laryngeal sound [2]. Sounds produced with the participation of vocal folds vibrations are called voice. The pitch of voice and, more specifically, its fundamental frequency changes during speech due to natural intonation. A male voice has an average frequency between 100-130 Hz, while a female voice has an average frequency between 200-260 Hz. The fundamental frequency in speech varies from 60 to 200 Hz for males and from 180 to 400 Hz for females.

Laryngeal sound is an input to the voice channel in which the spectrum is subject to significant modifications. The voice channel acts as a filter circuit (resonators) for specific resonant frequencies, which produces local maxima in the spectrum of the larynx tone after passing through the filter system that are called *controls*.

The primary and basic form in which the speech signal is registered is its temporal form. Assuming that for quasi-stationary fragments of speech the voice path is a linear sys-tem that is constant in time, the speech signal $s(t)$ is represented as a combination of pulse stimulation generated in the glottis $e(t)$ and the impulse response of the voice path $h(t)$.

$$s(t) = e(t) * h(t) \tag{1}$$

The time domain is not the most appropriate to perform further operations because the speech signal is characterized by significant redundancy. Further analysis of the speech signal is more efficiently performed in the frequency domain. A primary reason for analyzing speech in the frequency domain is an attempt to imitate nature; in the course of millions of years, an organ has evolved for the generation of human speech, in which the speech signal is generated, and the ear has evolved for the detection and analysis of human speech in the frequency domain. A significant number of computer methods are based on spectral analysis, which replaces the convolution in the time domain with the product of the spectrum of stimulation (laryngeal) and the spectrum of transmittance of the voice track (variable in step of articulation) in the frequency domain [3].

However, as the amplitude of the speech signal is modulated by the passing through the voice path, it is preferable to calculate in the first phase the logarithm of the spectrum. This way, the multiplicative relationship between the stimulation and the voice path is re-placed by an additive relationship, which greatly simplifies the subsequent separation of the two components. The reasoning presented above leads directly to *homomorphic processing methods*, in particular to the concept of *cepstrum* [3].

Because the calculation of the complex logarithm is associated with complications arising from the necessity of ensuring the continuity of phases, but the basic information in the speech signal is contained in the amplitude of its spectrum, in practice, the real cepstrum, is defined as follows:

$$c(t) = \mathcal{F}^{-1}\left\{ \ln\left(\left| \mathcal{F}\{s(t)\} \right| \right) \right\} \tag{2}$$

which for discrete signals, may be reduced to the following form

$$c(n) = IDFT\left(\ln\left(\left| DFT\left(s(n) \cdot w(n) \right) \right| \right) \right) \tag{3}$$

and finally

$$\begin{aligned} c(n) &= \frac{1}{N}\sum_{m=0}^{N-1} C(m) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} = \\ &= \frac{1}{N}\sum_{m=0}^{N-1} \ln\left(\left| \sum_{n=0}^{N-1} s(n)\, w(n) \mathrm{e}^{-\mathrm{j}2\pi\frac{mn}{N}} \right| \right) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} \end{aligned} \tag{4}$$

Due to the periodicity of the Fourier transform kernel, the logarithm of the amplitude spectrum module $C(m)$ is periodic and simultaneously it meets the equation

$$C(-m) = C(N-m) \tag{5}$$

Hence, it is an even function (symmetry with respect to the axis of ordinates), and there-fore, only cosine (even) functions appear in its expansion. As a result, it does not matter whether in the last step one uses a simple or inverse Fourier transformation or simply uses a cosine transformation. This allows for easy interpretation of the real cepstrum as a spectral logarithm-scale amplitude [3].

The amplitude spectrum of the speech is usually determined using a Fast Fourier Transform. The signal is composed of a rapidly changing factor (arising from the stimulation) and a slowly changing one (arising from the current construction of the organ of articulation) that modulates the amplitude of successive pulses resulting from the stimulation. Interpretation of the spectrum amplitude logarithm is similar, but the slowly changing component is not multiplied by the amplitudes of individual pulses from stimulation. Instead, the slowly changing component is added to the amplitude of the individual pulses. The calculation of the spectrum of such signals shows that the low-frequency waveforms associated with the transmittance of the voice path are close to zero on the pseudo-time axis, and pulses associated with laryngeal sound begin approximately at the laryngeal signal period and repeat periodically. Information related to the voice path transmittance is focused near zero time, and therefore, one should look for concise information on *what is being said* in this area. On the other hand, for the time period above the laryngeal sound, information about what is being said is minimized, and the only legible information is that concerning the laryngeal sound. Because the laryngeal sound is closely connected to anatomy of the larynx and glottis, it is a good carrier of individual information.

Parameterization of the speech signal

The most important step performed by a speaker recognition system is to generate an appropriate set of numerical descriptors that best characterize the recognized speakers. The purpose of the parameterization of the speech signal for ASR is transformation of the temporary input process to obtain the smallest possible number of descriptors containing in-formation relevant to the speaker,

while minimizing their sensitivity to variation in the signal that is irrelevant for ASR. The selection of these descriptors was guided by the analysis presented above, the process of speech generation and the searching for elements related to the individual characteristics.

Research method

The recordings of the acoustic signal were made at the *Institute of Electronic Systems Faculty of Electronics WAT* using a *Monacor DM-500* dynamic microphone, a computer sound card and *Matlab software*. The spatial variations of the acoustic pressure generated by the speaker are recorded at some point in space using a microphone, the task of which is to convert sound pressure to voltage. The recording conditions are determined by the characteristics of the recording microphone and the A/C. It was desirable that the equipment was of sufficient quality and that these elements had no significant effect on the structure of the recorded signal. During the test, the distance from speaker's mouth to the micro-phone was approximately 10 cm. Additionally, the microphone was equipped with a shield, which prevents distortion associated with whistling sounds (s-, s-, cz-, ć-) and explosive sounds (p-, b-, t). The phonetic material included a variety of phonetic text produced from fragments of a typical dialogue with a sublime and fun expression. The study group consisted of 16 men and 8 women.

The description is presented in the literature contains various strategies for selecting the sampling frequency. A smaller sample rate means less data to process, but at the expense of a loss of information. A higher sample rate means more data but not necessarily better recognition quality. In designing the speaker recognition system, we should find a compromise between the fidelity of the audio recordings in the context of saving individual characteristics, as well as the amount of data in computer memory and its effect on the speed of calculation. Pilot research was performed with signals sampled at frequencies of 44100 Hz, 22050 Hz and 11025 Hz. The best results were achieved by sampling at 22050 Hz with 16-bit amplitude resolution and recording of a single channel (mono). A database containing the speaker identifier has been created from recorded expressions and the corresponding samples of the acoustic signal.

Pre-processing

Pre-processing of the speech signal is a very important step in data processing be-cause it precedes the introduction of the signal to the feature generator and has a fundamental effect on the quality of the speaker identification process. The main purpose of pre-processing the speech signal is to ensure the greatest independence of the acoustic signals from the settings of the recording equipment. In the pre-processing stage, the filtration and normalization is performed to eliminate the differences between different frequency characteristics and the measurement circuits. This application uses a digital bandpass finite impulse response filter. The return loss, noise and disturbance were bypassed by assuming no distortion and signal noise issues. However, these issues will be taken into account in further research.

Feature generator

Speech signals have a variable frequency structure in time. Thus, the parameterization is subject to successive signal fragments and not the signal as a whole. Sections of the divided signal are called frames (Fig. 2). Generally, the frame length Δt is related to the shift (leap) τ, as follows:

$$\tau = \frac{1}{3}\Delta t \tag{6}$$

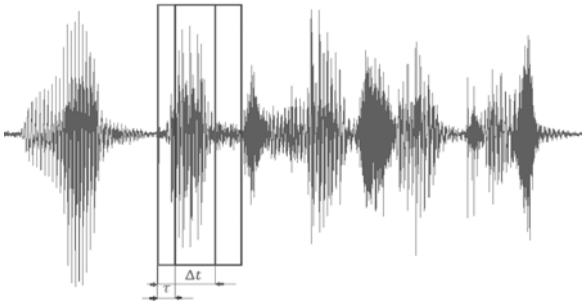

Fig. 2. Illustration of a frame shift - showing two successive frazes

One of the first tasks for the authors was to establish the basic parameters of the feature generator, which is the frame length. Durations of the individual phonetic units are different and depend on the speaker. Units consisting of voiced sounds are characterized by a duration ranging from 10 ms to even 200 ms. [3]. The range of variation is substantial, so the decision concerning the choice of the frame length is extremely important in the ASR. Studies on the optimization of the individual parameters feature generator are presented in the next chapter.

Framing of a signal causes discontinuities in the processed signal, which is associated with frequency leakage. To minimize this effect, the signal of each frame must be windowed by multiplication with an appropriate window function. Windowing results in the smoothing of the discontinuity and the removal of false spectral components. The *Hamming* window with good properties has been applied by the authors.

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right) \tag{7}$$

Because important information related to the speaker is contained only in the voiced parts of speech, only the "voiced frame" should be considered during the analysis. Voiced fragments are characterized by the occurrence of regular peaks (with the period of a basic tone). The voiceless parts are similar to an aperiodic signal. In the system, the classification of the speech signal into voiced or unvoiced parts is performed using the autocorrelation function. To verify if a sound is voiced, the second global maximum is determined and checks one level (the first maximum is in zero). If the level is higher than a reference value p_v, then this part is considered to be voiced; otherwise, it is deemed voiceless. Determining the optimal level of p_v is another step in the optimization described in Chapter 4.

By choosing representative frames for each speaker, an additional constraint was applied by the authors - the detection of speaker activity. During the registration, parts of the signal in which the speaker is not active occurred. Use of another parameter responsible for the rejection of frames without speech is aimed at eliminating the silence of the recording and the rejection of frames that are potential noise, which can cause erroneous feature ex-traction. In this approach, the signal statistics are first determined, and based on the selection, $P(n)$ will be determined and the decision criterion applied. Usually, the reference $P(n)$ is determined by a threshold. Depending on the value, algorithm and threshold upon which the selection is based, the results will be different. The power of the variable

component (the variance of the signal) has been chosen. The establishment of an additional parameter, the power level, is the next task to optimize. This task is described in the next chapter.

For analysis aimed at speaker recognition, the classical method for cepstral reconvolution is to remove the undesired ingredient by resetting the cepstrum samples for pseudo-time near zero. However, we should search for unique features of the speaker for the time period above the laryngeal sound.

The suitability of the real cepstrum for the purpose of speaker recognition can be easily noted by visually analyzing the waveforms shown in Figure 3; the information regarding the pronounced phone is blurred, while differences between different speakers are clearly visible [4].

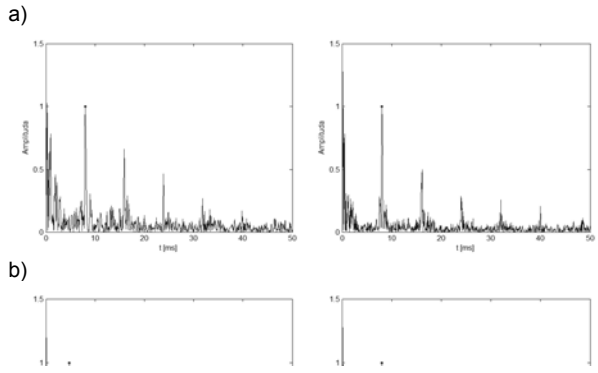

Fig. 3. Real cepstrum modules of *a*, *e* phones a) a male voice, b) female voice

Nine numerical descriptors differentiating the speakers were defined at the features generation stage. These descriptors include the fundamental frequency F_{av} (descriptor 1), corresponding to the reversal of the second maximum of the cepstrum, and the value of 7 successive maxima of orderly cepstrum c_1-c_7 (descriptors No. 3-9). The sets of cepstral features have been averaged based on representative frames. In addition, the sets have been completed by the standard deviation of the fundamental frequency σ (descriptor 2).

Multicriteria system optimization

The previous section showed a general diagram of the designed system. Depending on what function the system is to fulfill (recognizing the content of speech or the identity of the speaker), the optimal parameters of the system must be chosen with a consideration of the procedure of features extraction and the registration mode. The authors had the task to optimize the system based on four basic parameters: the length of the frame (Δt) and its shift (τ), the threshold of voiced frame (p_v) and the level of power (p_p).

Due to the wide ranges of changes of all the optimized parameters, the authors first decided to make an initial choice of the value of the parameters based on the coefficient of significance that Fisher defined in the following function:

$$F_{ij}(f) = \frac{|c_i - c_j|}{\sigma_i + \sigma_j} \tag{8}$$

The quantities c_i, c_j and σ_i, σ_j denote the mean values and the standard deviations of features for classes *i* and *j*, respectively.

The Fisher coefficients of significance were determined for nine descriptors based on the eight classes consisting of four women and four men. The even partition of men and women was not accidental. Note that the value of the descriptor may have high discriminative power between women but much less for men. Thus, the Fisher coefficient of significance was categorized into three subclasses: *Women*, *Men* and the subclass of *All*. Because the number of classes is more than two, the Fisher coefficient of significance was calculated for all pairs and was subsequently summed (the total Fisher coefficient of significance). In the first stage, the parameter to optimize was the frame length (Δt). The results are illustrated in Figure 4.

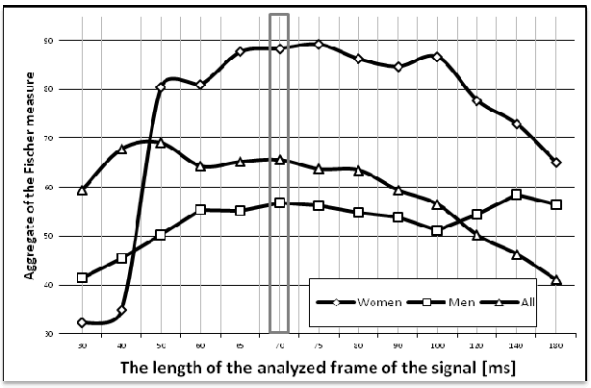

Fig. 4. Aggregate of the Fisher measure for each subclass depending on the length of the analyzed frame of the signal

It is clear from the graph that for a small frame length (30-40 ms) the Fisher coefficients are small. The strong growth begins approximately 50 ms, and for the frame lengths exceeding 90 ms, the value of coefficients in subclasses Women and All are significantly decreased. Thus, a frame length ranging from 60 to 80 ms was chosen. Note that there is no such frame length for which the Fisher coefficient of significance reaches a maximum in all three subclasses. Thus, we attempted to compromise. Finally, a frame length of 70 ms was chosen. To validate the choice, several series of detailed studies have been performed. These studies confirmed that the optimal frame length was Δt = 70 ms.

Another parameter to optimize was the shift with which the frame will move along the analyzed speech signal. To solve the problem, note that a smaller step value yields a larger number of frames, which translates into a longer calculation time. We attempted to seek the shift value of the frame run in parallel with the optimization of the two other parameters (p_v, and p_p). Due to the large amount of information contained in the input data (9-dimensional vectors of features), we decided to optimize based on the analysis of the main components (*Principal Component Analysis – PCA*). The essence of the PCA method is the transformation of a large amount of information contained in the mutually correlated input data into a set of statistically independent components arranged according to their validity. The PCA step was one of the most laborious research stages. The work relied on the observation of the change of position of the feature vectors for a speaker on the PCA_1/PCA_2 plane. The research was based on three sets of 8 speakers. The problem of choosing optimal parameter values is repeated for each set. A set of parameters that pro-vides a perfect distinction in one set of speakers is not best in another set. It was therefore necessary to choose a compromise, considering all 24 speakers who participated in the experiment.

According to the literature, calculating the fundamental frequency by the cepstral method is less accurate but more robust than the autocorrelation method, especially for an extremely noisy speech signal. To achieve greater stability of the descriptors, we decided to introduce an additional

constraint used in selecting the correct frames. The fundamental frequency will be compared based on the autocorrelation function and the cepstrum. Finally, we decided that if the differences between the values of the fundamental frequency frame by these two methods differ by more than 15%, the frame will be rejected automatically and will not be involved in the generation of descriptors.
The set of optimized parameters for the feature generator of 30-second segments of voice are shown in Tab. 1.

Tab. 1. Optimized parameters of the feature generator

Parameter		Value
Frame length	Δt	70 ms
Shift frame	τ	18 ms
Voiced level	p_v	10%
Power level	p_p	20%
Level of differences in the fundamental frequency	p_f	15%

Cepstral features selection

The set of descriptors defined at the stage of features generation are the maximum set of distinctive features. These descriptors can be used in automatic pattern recognition systems that represent the tested object. The maximum set of features has been shown to often not lead to the best results because they may have different impact on the pattern recognition. Some of the features resemble noise, thereby reducing the recognition efficacy. Some of the features are strongly correlated with the others, thereby adversely impacting the quality of classification by dominating over others and dampening their beneficial effects [5]. The important element is thus the assessment of the quality of each feature and selection of the best set of features on which the classification will be performed (identification, verification).

Initially the authors decided to apply a selection based on the Fisher method. According to the assumptions of the Fisher method, a large aggregate value coefficient of significance indicates a good potential separation between classes. In contrast, a small value means that the feature values belonging to both classes are scattered and potentially intermingled with each other, thereby disqualifying one as a diagnostic feature. The total Fisher coefficients of significance of each descriptor are shown in Figure 5.

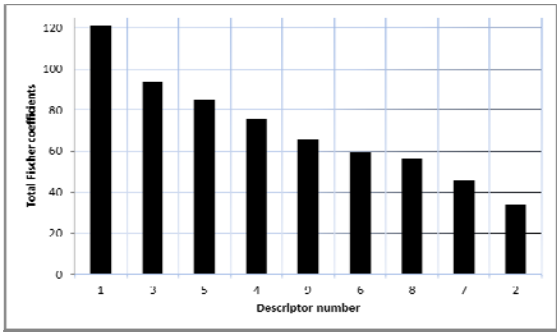

Fig. 5 Total Fisher coefficients of significance of each descriptor

The results indicate that the best descriptors are with numbers 1, 3, 5, 4, 9 and 6. The lowest value was obtained for descriptor 2 (variance of the fundamental frequency). Regardless of the total discriminant value of each feature, when building the automatic classification system, it is worth checking the discriminative power of the descriptors employed. However, it is known that the feature discriminative ability may change when used in co-operation with the others. Some features (even the worst ones) can be mutually enhanced, thereby raising their discriminative ability. We performed that analysis by following changes in the positions of the each vectors defining the speaker on the PCA_1/PCA_2 plane.

Based on the Fisher coefficient and the observation of position of feature vectors on the PCA, the optimal 5-dimensional feature vector v (*Voice Print* - *VP*) was determined. It consists of the fundamental frequency and 4 cepstral features.

$$\begin{cases} vp_1 = F_{av} = \dfrac{1}{N}\sum_{j=1}^{N} F_j , \\ vp_{i-1} = \dfrac{1}{N}\sum_{j=1}^{N} c_{i,j} , \qquad i = 3,4,5,6 \end{cases} \tag{9}$$

where: N – number of correct frames
F_j – fundamental frequency of the j-th frame, determined from the real cepstrum,
$c_{i,j}$ – value of the i-th maximum of the real cepstrum for the j-th frame.

Result of the study

As a result of the multi-criteria optimization of parameters and selection of descriptors, we have obtained the final model of the feature generator for the proposed speaker recognition system. Examples of the results of the PCA transformation are shown in Figure 6. Individual results relate to the three separable sets of speakers (each set includes 8 speakers). Each speaker is represented by eight separate 5-dimensional *Voice Prints*. Note the analysis presented in Figure 6c. It shows only the *Men*, as contrasted with the previous two, where the presented sets consist of equal numbers of women and men (women are grouped on the right side of the plane, while the men are grouped on the left-hand side of it).

The main advantage of the PCA transformation is an opportunity to observe the distribution of individual feature vectors on the plane despite the fact that the original feature vector is 5-dimensional. The vectors were obtained for each speaker using the optimized feature generator. This generator enables an initial classification of the different speakers. Note that for each speaker we obtained reproducible results, despite the large diversity of recorded speech (dialogue, voice serious and humorous speech). This validates the fulfillment of the basic condition of the proposed system, which is the decoupling of the generated features vector from the content and character of the speech.

Note that the observation of the obtained feature vectors was performed only based on two major components. Further experiments show that additional components are con-firmed by an even better separation of each speaker. This improvement is clearly visible when two speakers slightly overlap in the function of the two major components. Additional analysis of these vectors as a function of PCA_3 and PCA_4 leads to the conclusion that these classes are clearly separate.

Note that a mother and daughter were among the participants of the study. A slight overlap of the two speakers may be expected due to the close relationship between the speakers. However, even these two classes are clearly separable. This separation can be observed in Figure 6a, where the vectors denoting the characteristics of the mother and daughter are the points marked in blue and green. This result demonstrates the good discrimination property of proposed feature generator.

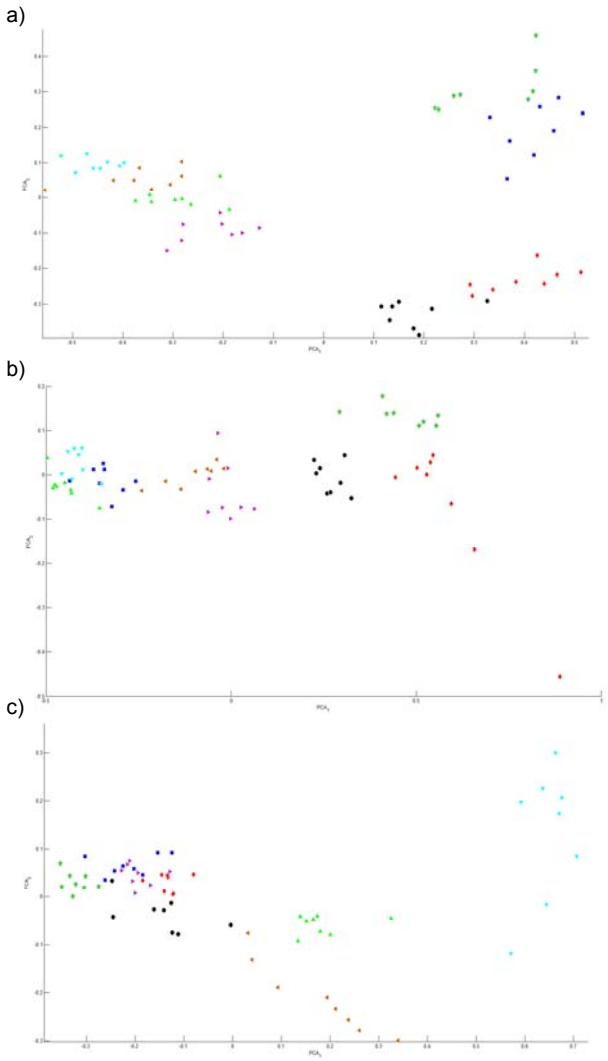

Fig. 6. The results of the data projected on the two major components of PCA; for sets 1, 2 and 3 of the speakers

Conclusion

The stage of parameterization of the signal is very important because the incorrect results of this stage cannot be corrected in further stages. The conducted experiments have allowed for the optimization of the model of feature generator in the proposed speaker recognition system. A multi-criteria optimization of the selected parameters and the selection of the descriptors was performed. The results presented by the PCA transformation look very promising. Each of the speakers is concentrated in a separate area. Additional research performed that accounted for the greater number of PCA components confirmed the presented conclusions. Therefore, the generated vectors are expected to be primarily unique to each speaker as well as robust to the spoken text.

The final stage of the process of speaker recognition is a classification; presently, the authors are undertaking this issue. The current analysis shows that a non-linear SVM will be applied in the classification, and the primary goal for the authors will be the selection of the optimal parameters of the network that ensure minimal errors of classification.

"*This work is supported by the Polish Ministry of Science and Higher Education in the years 2010-2012 as a development project*."

REFERENCES

[1] S. Furui, "Recent advantages in speaker recognition," Pattern Recognition Letters 18, pp. 1859-1872, 1997.

[2] T. Kinnunen, H. Li, " An overview of text-independent speaker recognition: From feature to supervectors", Speech Communication, pp. 12-40, 2010.

[3] A. Dobrowolski, E. Majda, "Cepstral analysis in the speakers recognition systems," 15th IEEE SPA Conference, Poznan, Poland, pp. 85-90, 2011.

[4] A. P. Dobrowolski, E. Majda, "Application of homomorphic methods of speech signal processing in speakers recognition system", Electrical Review, pp. 12-16, 2012.

[5] M. Kruk, S. Osowski, R. Koktysz, "Recognition of Colon Cells Using Ensemble of Classifiers", International Conference on Neural networks, Orlando, Florida, USA, 2007, pp. 288 – 293.

***Authors**: Andrzej P. Dobrowolski, Ph.D, D.Sc., Ewelina Majda, M.Sc. Military University of Technology, Faculty of Electronics, Institute of Electronic System, 2 Kaliskiego street, 00-908 Warsaw, tel. +48 22 6837534, E-mail: Andrzej.Dobrowolski@wat.edu.pl, Ewelina.Majda@wat.edu.pl.*