Ewelina MAJDA-ZDANCEWICZ, Andrzej P. DOBROWOLSKI[1]

Military University of Technology, Faculty of Electronics (1)

doi:10.15199/48.2015.10.52

Text Independent Automatic Speaker Recognition System using fusion of features

Abstract. *This paper presents a speaker recognition system, which is independent of the linguistic context. The solved task includes: the pre-processing stage, the segmentation of speech signal leading to the extraction of features based on three techniques, selection of the most important features, and the classification stage involving a serial combination of classifiers. Sets of descriptors were obtained using three techniques: cepstral coefficients, mel-cepstral coefficients and original weighted cepstral coefficients. Optimal robust "Voice Print" has been determined using fisher coefficients and PCA analysis. Experiments on the 2002 NIST Speaker Recognition Evaluation corpus show that the proposed system is able to recognise the speaker, regardless on the speech content, even language content with great accuracy.*

Streszczenie. *W pracy przedstawiono system rozpoznawania mówcy niezależny od tekstu wypowiedzi. Rozwiązane problemy obejmują: etap przetwarzania wstępnego, segmentację sygnału mowy prowadzącą do etapu ekstrakcji cech bazującej na trzech technikach analizy sygnału mowy, selekcję najbardziej istotnych cech oraz etap klasyfikacji obejmujący analizę kaskady klasyfikatorów. Zestaw cech uzyskano przy użyciu trzech technik: cepstrum, mel-cepstrum oraz autorskich ważonych cech cesptralnych. Optymalny wektor cech wyekstrahowano przy użyciu współczynników istotności Fishera oraz analizy PCA. Eksperymenty z wykorzystaniem bazy 2002 NIST Speaker Recognition Evaluation pokazują, że przedstawiony system rozpoznaje mówcę niezależnie od ograniczeń lingwistycznych treści, a nawet języka wypowiedzi, z zadowalającą dokładnością.* ***(Automatyczny system rozpoznawania mówcy niezależnie od wypowiadanego tekstu bazujący na fuzji cech)***

Keywords: automatic speaker recognition, features extraction, features selection, PCA.
Słowa kluczowe: automatyczne rozpoznawanie mówcy, ekstrakcja cech, selekcja cech, PCA.

Introduction

Speaker recognition refers to the automated method of identifying or confirming the identity of an individual, based on his/her voice. A very important characteristic of speaker's recognition systems is their dependence on the recognised text spoken by a person, that is, the limitations imposed on the linguistic material. Speaker recognition systems can be divided into text-dependent and text-independent tasks. In text-dependent 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].

In general, the procedure for identification of persons can be divided into three phases. The pre-processing block is responsible for receiving the signal from the microphone and its initial processing. The second stage involves analysis of the speech signal, in order to obtain parameters carrying information about the individual characteristics of the voice of the speaker, regardless on the speech content. The final stage is classification [1]. For any speaker recognition system, the most critical step is to arrange an adequate set of parameters, which would enable carrying out the recognition procedure. The basic requirement for such a set of voices is to ensure discrimination between different individuals based on values and repeatability of the parameters for various phrases expressed by the same person. A better parameter is considered the one, the value of which is exactly reproducible (or very similar) for various expressions of the same speaker and relatively different from expressions of other speakers. In order to extract relevant parameters from a speech signal, the signal must be parameterised, which is critical for effectiveness and reaction rate of the entire speaker recognition system. The result of speech signal parameterisation is a unique features vector, called the author's "voice print" [9]. When dealing with a vast number of various parameters one should seek some method of selecting the optimal (most discriminating) set of parameters describing the signal.

The paper presents a text-independent speaker recognition system. The main objective of the research was to extract features, which would be robust against speech and language content, while ensuring a great identification rate.

Related work

The techniques for text-independent speaker recognition may be divided into two main tasks: features extraction and classification [2].

In the feature extraction a few approaches can be used. The main approach includes such methods as: time domain extracted features [3], spectral features [4], mel-cesptral coefficients (MFCCs) [5], linear predictive cepstral coefficients (LPCSs) [6], and perceptual linear prediction (PLP) [7]. The second involves voice source features, including: fundamental frequency and other parameters related to the glottal flow model, the shape of the glottal pulse. The degree of vocal fold opening and the duration of the closing phase (wavelet analysis, residual phase, cepstral coefficients, and high-order statistics) are included [2]. The next way relates to the prosodic feature including duration (e. g. pause statistics, phone duration), speaking rate and energy distribution/modulations among others. Another approach includes a high-level features attempt to capture conversation-level characteristic of speakers, such as characteristic use of words. The choice of features must be based on their discrimination, robustness and practicality.

Classification may be categorised into three major approaches: *template models* (*vector quantisation* (VQ), *dynamic time wrapping* (DTW)), *stochastics models* (*Gaussian mixture models* (GMM) and *hidden Markow model* (HMM)) and also *parametric methods*: *artificial neural network* (ANNs) and *support vector machines* (SVM) [2].

Methods and systems

The speech database

The speaker recognition experiments were conducted using our own database, with 50 enrolled speakers (38 men and 12 women). The total length of all recordings registered in Polish was about 4 minutes. The signals were sampled at a frequency of 22 050 Hz with 16-bit amplitude resolution and recording of a single channel (mono). *This database has been created only to build the system and optimise*

associated parameters. To appraise the system's work, the authors used a well-known dataset (*NIST database*).

Pre-processing method

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 - lowpass type II Chebyshew filter: (4.6 kHz, -3dB), (5kHz, -6 dB), and normalisation is performed to eliminate differences between different frequency characteristics and the measurement circuits. The return loss, noise and disturbance were bypassed by assuming no distortion and signal noise issues. These issues are active subject, often described in a separate research [8]. However, these issues will be taken into account in further research.

Frames selection

Speech signals have a variable frequency structure in time. Thus, the parameterisation is subject to successive signal fragments and not the signal as a whole. Sections of the divided signal are called frames (where frame length is Δt and the shift – leap – τ). Framing of a signal causes discontinuities in the processed signal, which are associated with frequency leakage. To minimise this effect, the signal of each frame must be windowed by multiplication with an appropriate window function (the Hamming window has been applied).

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. 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. By choosing representative frames for each speaker, an additional constraint was applied by the authors – the detection of speaker activity. Use of another parameter responsible for the rejection of frames without speech is intended to eliminate the silence of the recording and the rejection of frames that are potential noise, which can cause erroneous feature extraction. The power of the variable component (the variance of the signal) has been chosen. The establishment of an additional parameter, the power level p_p, is the next task to optimise.

Another restriction is associated with the determination of the fundamental frequency (one of the features included in the "Voice Print"). According to the literature, calculating the fundamental frequency by the cepstral method (F_{0c}) is less accurate, but more robust, than the autocorrelation method (F_{0ac}), especially for an extremely noisy speech signal. To achieve greater stability for the "Voice Print", an additional constraint (p_f threshold) has been used. The formula is as follows

$$|F_{0c} - F_{0ac}| \le p_f \min(F_{0c}, F_{0ac}) \tag{1}$$

Studies on the optimisation of the individual parameters (Δt, τ, p_v, p_p, p_f) are presented in the *Multicriteria system optimization*.

Feature extraction

The features utilised by these systems must describe the human voice as a means of distinguishing between different speakers. After appropriate feature selection, a feature vector will be created and used as the basis for classification (identification and verification). The authors decided to search for distinctive features by considering phenomena related to the internal structure of the source of the speech signal [12].The feature generation is based on three cepstral analysis techniques. In each method, a set of preliminary pre-selection characteristics is created, and then all generated descriptors are fused.

Cepstral features

The primary and basic form in which the speech signal is registered is its temporal form. The time domain is not the most appropriate to perform further operations because the speech signal is characterised by significant redundancy therefore the homomorphic processing methods, in particular to the concept of cepstrum are being used. A thorough analysis led to the conclusion that the characteristic descriptors include the fundamental frequency F_{0av} (*Descriptor 1*), corresponding to the inverse of the first maximum of the cepstrum, and the values of the 4 successive maxima of the cepstrum normalised by the value of the first maximum [9].

Mel-cepstral features

The most popular method of parameterising speech signals is to use the *MFCCs* (*Mel-Frequency Cepstrum Coefficients*) – Fig.1.

Fig. 1. Diagram of the calculation of coefficients MFCC

A major feature of this transformation is the conversion of the spectrum to a linear scale, which accounts for the nonlinear perception of sound frequency by humans and significantly reduces the size of the data. The mel scale was determined empirically by the following process [5, 10].

$$f[\text{mel}] = 1127 \ln\left(1 + \frac{f[\text{Hz}]}{700}\right) \tag{2}$$

Thirty filters were applied during the MFCC generation, providing 30 distinct coefficients, i.e. 30 filters used in the band from zero to half the sampling frequency. Determining which of the MFCC features are representative of the pronounced sound and which are representative of the speaker is a difficult task. Features that are related to the linguistic content of the speech should not be considered and, as described above, the cepstral reconvolution technique should only consider features above a certain threshold. The authors applied an initial pre-selection of relevant features and reduced the length of the MFCC vector to 7 while minimizing any loss in the vector's representativeness. The results were checked using *Principal Component Analysis* (*PCA*). This method was used because of the large initial dimension of the preliminary vector of *MFCC* features. Display 30 – dimensional vector *of MFCC* features on plane, enabled the efficient initial pre-selection of features relevant to the modelled of feature generator.

Original "weighted cepstral features"

The authors, inspired by the idea of the MFCC method, attempted to extend the features vector to include other original features defined in the cepstrum by using sub-band bleeder filters. The proposed algorithm does not produce the same peaks at their expected positions; rather, it sums the amplitudes of all of the relevant bands with certain weights (Fig.2). To optimise the system, the optimal characteristics of the filter (weighting function) and the optimal widths of the bands must be selected. The rectangular function was found to be optimal. The second algorithm through the fifth cepstral maxima represents the 4 relevant weighted cepstral features and is normalised to the

sum of the amplitudes received in the first band, which corresponds to the fundamental frequency.

At the feature generation step, 16 numerical descriptors are defined to differentiate speakers c_1-c_{16}. These descriptors include the fundamental frequency F_{0av} (c_1), corresponding to the inverse of the first maximum of the cepstrum; four weighted cepstral features (c_2-c_5); the four successive normed maxima of the ordered cepstrum (c_6-c_9) and seven mel-cepstral features (c_{10}-c_{16}). Each set of features for each speaker was averaged over a set of representative frames.

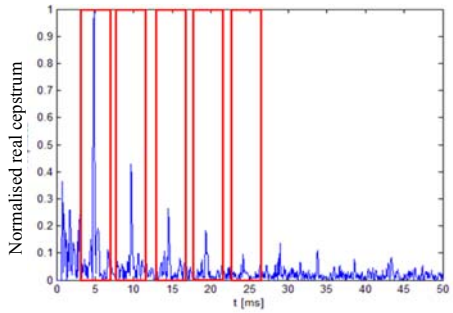


Fig. 2. The idea of Weighted cepstral features (for the normalised real cepstrum module, rectangular weighting function has been applied)

Multicriteria system optimisation

The authors had the task to optimise 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) [9]. Due to the wide ranges of changes of all the optimised parameters, the authors decided arbitrarily to make an initial choice of the value of the parameters based on the coefficient of significance that Fisher defined in the following function

$$F_{ij}(f) = \frac{\left|c_{av(i)} - c_{av(j)}\right|}{\sigma_i + \sigma_j} \tag{3}$$

The quantities $c_{av(i)}$, $c_{av(i)}$ and σ_i, σ_j denote the sample average values and the sample standard deviations of features for classes *i* and *j*, respectively [11].

The Fisher coefficients were determined for sixteen descriptors based on the fifty classes consisting of women and men, because value of the descriptor may have high discriminative power between women but much less for men. Thus, the Fisher coefficient was categorised into three subclasses: *Women*, *Men* and the subclass of *All*. Because the number of classes is more than two, the Fisher coefficient was calculated for all pairs and was subsequently summed (the total Fisher coefficient). In the first stage, the parameter to optimise was the frame length Δt (Fig. 3.).

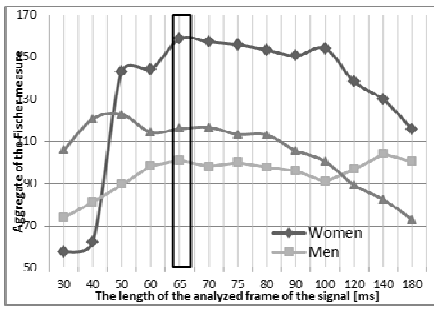


Fig. 3. Aggregate of Fisher measure for each subclass depending on the length of the analysed frame of the signal

Note that there is no such frame length for which the Fisher coefficient reaches a maximum in all three subclasses. Thus, a compromise was attempted. Initially the authors decided on the value of 65 ms, but about the final value of this parameter, as well as all others, *the parallel selection process has been decided on*. It is worth noting that the selection of the features affects the optimisation, so the two processes must be repeated to obtain the optimal solution.

Another parameter to optimise was the shift with which the frame will move along the analysed speech signal. It was attempted to seek the shift value of the frame run in parallel with the optimisation of the three other parameters (p_v, p_p and p_f). Parallel analyses have been based using two methods: the Fisher coefficients and the *PCA analysis* (*Principal Component Analysis – PCA*). 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 and on PCA_3/PCA_4. The research was based on the three separable sets of speakers (each set includes 8 speakers), treated as a representative group of 50 personal databases of speakers. The main problem has been related to the fact, that selecting the optimal parameter values for a set of parameters to ensure a perfect distinction in one set of speakers has not worked best in the case of another set. In the experiments it is necessary to make a compromise considering all the persons involved in the experiment. The set of optimised parameters for the features generator of 15-second segments of voice are shown in Tab. 1. Final optimisation of the parameters has been made after the final choice of the classification method.

Table 1. Optimised parameters of the feature generator

Parameter		Value
Frame length	Δt	65 ms
Shift frame	τ	16 ms
Voiced level	p_v	10%
Power level	p_p	20%
Level of differences in the fundamental frequency	p_f	20%

Feature 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 impacts on the pattern recognition. Two strategies can be used to study the quality of these features. The first strategy is to test each feature regardless of the method of classification (the so-called filtering features) and assess their ability to differentiate the speakers without considering the specific classifier. The second strategy is to select the features based on the characteristics of the classifier [11]. The authors decided to filter the features, because a final decision regarding the specific classifier has not yet been made. The serial model of selection has been used for achieve better results. It was a combination of two methods: the *Fisher's method* supplemented by the analysis of *Principal Component Analysis*. The total Fisher coefficients of each descriptor are shown in Fig. 4. The Fisher coefficient calculated by (3) can be directly applied to two-class problems. To deal with the problem of many classes it is necessary to use an approach *one vs. rest*. In this approach, a set of coefficients for each should be added together to obtain the value of the total Fisher coefficient.

The Fisher method as an example of ranking methods does not take into account the dependencies between

features. For this reason, *PCA analysis* has been used. 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 employyed. However, it is known that the feature discriminative ability may change when used in co-operation with the others [11]. PCA takes into account the characteristics of competition. Fig. 5 shows two examples of this distribution.

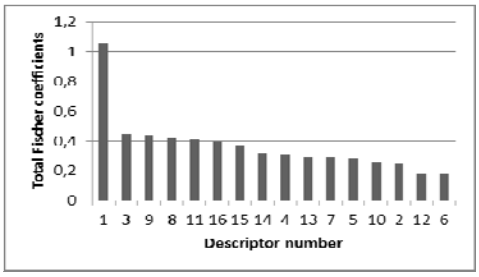
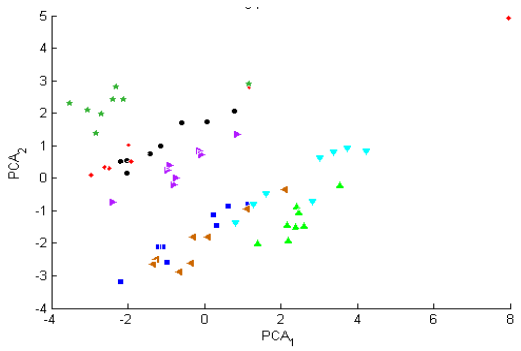


Fig. 4.Total Fisher coefficients of each descriptor

a)



b)

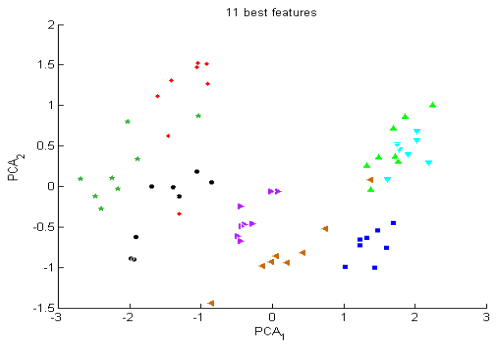


Fig. 5. The distribution of the data focused on the two principal components for 8 speakers (8 different symbols); single speaker represented by 8 “Voice Print” included; a) all features, b) the 11 best features;

Based on the Fisher coefficients of each descriptor and of the observed changes in the feature vectors based on the PCA transformation, the optimal *11-dimensional feature vector VP*, called the *Voice Print*, has been determined. For each speaker averaging was made of selected set of features based on the 15-second excerpts recorded speech, taking into account only the correct frame selected at the pre-processing stage.

$$F_{0av} = \frac{1}{N}\sum_{j=1}^{N} F_{0j},$$

$$S_i = \frac{1}{N}\sum_{j=1}^{N} s_{i,j}, \quad M_i = \frac{1}{N}\sum_{j=1}^{N} m_{i,j}, \quad MC_i = \frac{1}{N}\sum_{j=1}^{N} mc_{i,j}, \tag{4}$$

$$\mathbf{VP} = \left[F_{0av}, \frac{S_2}{S_1}, \frac{S_5}{S_1}, \frac{M_2}{M_1}, \frac{M_3}{M_1}, \frac{M_5}{M_1}, MC_1, MC_2, MC_3, MC_5, MC_7 \right]$$

where: N – number of correct frames, F_{oj} – fundamental frequency of the j-th frame, $s_{i,j}$ – sum of value of the real cepstrum surrounded by the i-th maximum for the j-th frames (equivalent to the mean value in sub-band), $m_{i,j}$ – value of the i-th maximum of the real cepstrum for the j-th frames, $mc_{i,j}$ – i-th coefficients of mel-frequency cepstrum (*MFCC*) for j-th frames.

Detection of individual peaks was carried out on a search around the maximum values predicted peaks set on the fundamental frequency.

Classification

In the speaker recognition system it was decided to attempt the cascade of classifiers. In the first stage, due to the needed low computational demands, two nonparametric analysis classification methods: the *k nearest neighbours method* and the *method of near average,* have been used. In the second stage *Support Vector Machine SVM* is applied. It is known as an excellent classifier of good generalisation [11].

The data for 50 speakers is divided into training set (75 % all data) and a test set (25 % all data). The purpose of the study was to select the optimal value of *k* (*k* nearest neighbours method) and *a* parameter (near average method). Parameter *a* determines the extent of the class Table 2-3 presents the results of these experiments.

Table 2. Number of misclassifications (the *k* nearest neighbours)

k	1	2	3	**4**	5	6	7	8
Number of errors	12	12	12	**11**	14	16	18	20

Table 3. Number of misclassifications (method of near average)

a	1	**1.5**	2	2.5	3	4	5
Number of errors	38	**37**	40	40	40	40	40

The best result of recognition of all classes has been achieved for 4 of the nearest neighbours method - misclassification rate 2.2% (misclassification of 11 of the 500 tested vectors). This result is undoubtedly a very satisfactory outcome for this type of system.

In the second stage of research an additional classifier – *SVM* has been analysed to reduce number of misclassification. To deal with problem of many classes the "one against all" approach has been used in a limited set of classes (linear and non-linear SVM). In the linear SVM the regularisation constant *C* has been adjusted. The non-linear SVM of Gaussian kernel has been used. The hyperparameters σ of the Gaussian function and the regularisation constant *C* have been adjusted. Unfortunately, the use of both the linear and non-linear SVM network did not yield the expected results (increased calculation time, unacceptable number of misclassifications).

Perhaps a better solution could be to use the approach: "one against one". This solution was rejected because it is not acceptable due to the fact that the addition of a new person to the base requires new classifiers. These results led the authors to reject the SVM classifier as an additional classifier.

Results of experiments

The text-independent speaker recognition system has been presented in the previous sections. It was built using a hand-created database (50 speakers). For a reliable assessment of the proposed system, experiments should be performed using an independent voice database. The *2002 NIST Speaker Recognition Evaluation database* has been used. In recent years the National Institute of

Standards and Technology (NIST) has promoted research in the context of text-independent speaker recognition [1, 2]. Data includes cellular telephony speech data registered in English (women and men). The sampling rate was 8 kHz and amplitude resolution of 8 bits. The quality of NIST database is much lower than our own database. This approach gave a unique opportunity to test the robustness of the system in various conditions. The total speech length for each speaker was about 120 seconds.

According to the assumptions, the highest efficiencies have been achieved by the length of testing data -15 s (error rate - 2.2%). In the first part of the research, experiments were limited to this approach. The longest training speech segment (90-seconds) has been used. The 2 % - misclassification ratio - on the testing data has been obtained. (1 false identification for 50 speakers).

To evaluate the influence of the training size, the experiments were run over three sets with decreasing training size (90 s., 60 s., 30 s.). As it can be seen from Table 4, the best identification rate for speakers was one of 98 % (1 false identification), provided by the maximum length of training data – 90 s. Reduction size of learning data causes increase in the level of misidentification only in one case. There is no difference in misclassification ratios during decreasing learning size from 60 to 30. The results are very promising. The conclusion is clear: three times reducing the size of learning data does not so significantly decrease the identification rate (from 98% to 94%).

To evaluate the influence of the testing size, the experiments look very similar. The length of testing data has been reduced (from 15 s to 5 s), while the training size has remained constant (90 s). Identification rate was reduced from 98% to 94%. However it should be noted, that the effectiveness of the system is still high (94% correct identification). The results for testing are given in Table 4 and Table 5.

Table 4. Correct identification as a function of the length of learning data

	The length of learning data		
	90 s	60 s	30 s
correct identification	98%	94%	94%

Table 5. Correct identification as a function of the length of testing data

	The length of testing data		
	15 s	10 s	5 s
correct identification	98%	96%	94%

Conclusions

The paper has presented a speaker recognition system, which is independent of the linguistic content. The most important problems solved in the work include: the pre-processing stage, the segmentation of speech signal leading to the extraction of features based on three techniques, selection of most important features, and finally the recognition of the speaker using non-parametric 4 nearest neighbours methods. The robustness of the proposed system has been checked on the *2002 NIST Speaker Recognition Evaluation database.* These experiments have led us to build the feature extractor, which is characterised as robust to the spoken text, and the accompanying classifier, which provides the minimum number of false identifications. In all, the approach identification rate is above 90%. It is worth emphasising that system was designed based on recordings in the Polish language, but final researches were made using recordings in the English language. This system may be used for different conditions. An analysis led us to the conclusion that the language of expression has no significant impact on the operation of the entire system.

REFERENCES

[1] Furui S. *Recent advantages in speaker recognition*, Pattern Recognition Letters, 18 (1997), no. 9, pp. 859-872.

[2] Kinnunen T., Li H., *An overview of text-independent speaker recognition: from features to supervectors*, Speech Communications 52, 2010, pp. 12-40

[3] Orman D., Arslan L., *Frequency analysis of speaker recognition*, Proc. Speaker Odyssey: the Speaker Recognition Workshop, Greece, 2001, pp.219-222

[4] Lupu E., Emerich S., *Speaker identification approach based on time domain extracted features*, 52 nd International Symposium EMLAR 2010, Croata, pp. 355-358

[5] Zheng F., Zhang G., Song Z., *Comparison of Different Implementations of MFCC*, J. Computer Science &Technology 16(6),pp. 582-589, 2001

[6] Huang X., Acero A., Hon H.W., *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice Hall PTR, 2001

[7] Hermansky H., *Perceptual linear prediction analysis for speech*, J. Acoustic Soc. Amer., 87, 1990, pp. 1738-1752

[8] Ming J., Hazen T., Glass J. R., Reynolds D. A., Robust Speaker Recognition In Noisy Conditions, *IEEE Transactions on Audio, Speech, and Language Processing*, 15, (2007), no. 5, pp. 1711-1723

[9] Majda E., Dobrowolski A. P., *Modeling and optimization of the feature generator for speaker recognition systems*, Przegląd Elektrotechniczny, 88, (2012), no. 12a, pp. 131-136.

[10] Kopparapu S.K., Laxminarayana M., *Choice of Mel Filter Bank in Computing MFCC a resamples Speech*, 10 th International Conference on Information Science, Signal processing and their Applications, 2010, pp. 121-124, Malaysia

[11] Kruk M., Osowski S., Koktysz R. , *Recognition of Colon Cells Using Ensemble of Classifiers*, International Conference on Neural networks, 2007, pp. 345-349, Orlando

[12] Dobrowolski A. P, Majda E. *Application of homomorphic methods of speech signal processing in speakers recognition system*, Przegląd Elektrotechniczny, 88, (2012), no.6, pp. 12-16

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