Piotr LENARCZYK, Zbigniew PIOTROWSKI

Military University of Technology, Faculty of Electronics

Speaker Recognition System Based on GMM Multivariate Probability Distributions built-in a Digital Watermarking Token

***Streszczenie**. Przedstawiony poniżej artykuł opisuje system rozpoznawania mówcy na podstawie mowy ciągłej, wykorzystując wielowariancyjne rozkłady prawdopodobieństwa GMM. Opisane zostały procesy ekstrakcji cech dystynktywnych głosu oraz tworzenia modeli statystycznych. Algorytm został zaimplementowany w systemie Linux w celu poprawy funkcjonalności identyfikacji użytkownika Zaufanego Osobistego Terminalu PTT. (**System rozpoznawania mówcy na podstawie wielowariancyjnych rozkładów prawdopodobieństwa zaimplementowany w tokenie znaku wodnego**).*

Abstract. The article describes a speaker recognition system based on continuous speech using GMM multivariate probability distributions. A theoretical model of the system including the extraction of distinctive features and statistical modeling is described. The efficiency of the system implemented in the Linux operating system was determined. The system is designed to support the functionality of the Personal Trusted Terminal PTT in order to uniquely identify a subscriber using the device.

Słowa kluczowe: GMM, rozpoznawanie mówcy, PTT, biometria.
Keywords: GMM, speaker recognition, PTT, biometrics.

Introduction

Systems for identification of speaker biometric identity allow for much more efficient information security management since many services, such as authorization, may be performed by means of individual and unique data that is specific to a single person. A Trusted Personal Terminal, PTT [1], [2] is a hardware digital watermarking token developed for the purpose of identifying the subscriber over open (unencrypted) communication links. The digital watermark sent in speech signals transmitted over the telecommunication link is inaudible, and represents the subscriber's PIN. In order to enable the use of PTT only to authenticated users a method of voice identification was developed and described in this article. The subscriber authentication model for open telephone links using the PTT is a two-stage process. In the first stage a subscriber is assigned to the PTT by voice authentication. In the second stage a subscriber authenticated as above sends his or her PIN using the PTT watermark token to the other side of the link in order to identify their identity.

In general it can be said that the speech signal carries two different types of information [3] - essentially it is syntax, semantic and pragmatic in sense, but at the same time, it is a rich source of information that clearly identifies the speaker. A voice is a rich source of biometric data [4] [5], due to the fact that it differs for each person in terms of frequency, amplitude and phase in a sufficiently unique way to distinguish various speakers. Both the vocal tract unique in its construction and the signal stimulating this tract generate the distinctive speech features assigned to a particular person.

User recognition

Speaker recognition is the process of determining which one of the registered users spoke a fragment of an utterance. In these types of systems information about the identity of the speaker is not given a 'priori and certain voice characteristics are sought in the test utterance. Then they are compared with a database of statistical parameters stored for all users and the ones that best identify the speaker are indicated. The implemented speaker recognition system is of an open type, where the test utterance belongs to one of the registered users. During testing an adaptation coefficient is calculated for the set of distinctive speech features extracted during the testing process in relation to those calculated during learning (training) and to the previously calculated UBM impostor model. Calculated on the basis of statistical model approximation of infinite number of users, it estimates a model of a person from outside the set of authorized users. The necessary condition is to calculate the UBM in conditions of a set similar to the set of authorized users (ratio of male to female, age, accent, etc.) where none of the people used for estimating the UBM belongs to the set of authorized users. Then the user with the highest adaptation coefficient is recognized as the author of the test utterance, or if all users have obtained similar results and the testing of the impostor model on the basis of the test utterance has achieved the highest score, the user will be recognized as a new user (or impostor - for authentication systems). A system of this type requires N+1 comparisons for N users; well-designed speaker recognition algorithms have a wide scope of application, thanks to low error rates achieved [6], [7]. Speaker recognition algorithms can operate on parameters calculated based on continuous speech or isolated words (phrases). The described algorithm uses MFCC continuous speech coefficients [8] for testing the user model.

Theoretical model of the speaker recognition system

Two approaches to the problem of statistical modeling are mainly used in the process of designing algorithms for speaker recognition:

Hidden Markov Models (HMM) - this is the oldest approach to this problem and it employs the modeling of each user as a single Markov model.

Gaussian Mixture Models (GMM) - in this case the dependence of modeling for each user is employed, in the form of a sum of Gaussian probability distributions for extracted distinctive speech feature vectors [7]

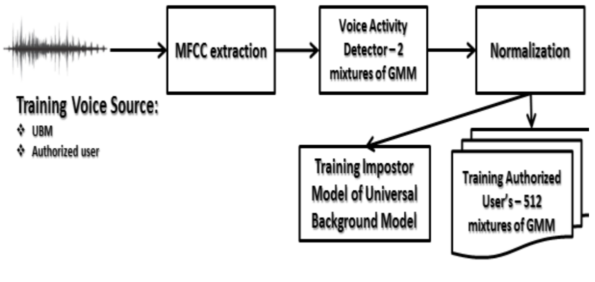
MFCC extraction
Voice Activity Detector – 2 mixtures of GMM
Normalization
Training Voice Source:
UBM
Authorized user
Training Impostor Model of Universal Background Model
Training Authorized User's – 512 mixtures of GMM

Fig. 1. Diagram of the training phase of the designed system

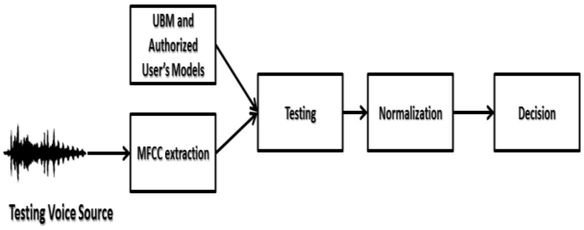

Fig. 2. Diagram of the testing phase of the designed system

Speaker recognition system diagram

Figures 1 and 2 show diagrams of the two life cycle phases of the system - training and testing. The training phase is used to develop the model that statistically estimates the extracted set of vectors identifying the speaker identity, while the testing phase is used to indicate the most probable attribution of the test utterance to one of the available (authorized) user models, or to qualify it as an impostor's utterance (or a new user).

In the training phase, following the extraction of vectors of distinctive speech features, VAD - Voice Activity Detector is used. Its employment stems from the need to reject such feature coefficients that contain unnecessary information. In the real world, speech includes a lot of pauses, respiratory breaks, environmental noise, etc. which are undesirable in the process of GMM learning. For this purpose, a two - distributive mixture model of multivariate Gaussian probability function of the input multidimensional set of extracted distinctive speech features is used. After the two distributive classification process, there follows a rejection of low-energy coefficients with corresponding speech signal segments marked as unsuitable.

The next step consists in the Normalization of the set of coefficients of distinctive speech features aimed at reducing the environmental influence on the speech signal. In order to remove the effect of using different microphones, A/D and D/A converters, the telecommunication channel characteristics, the coefficients are normalized in relation to both the expectation and variance. A simple and effective way is to calculate the average μ and expectation σ^2, taking into account all the coefficients (assuming that for the duration of the recording those are approximately constant). The next step consists in the normalization of each coefficient according to the formula:

$$x' = \frac{x-\mu}{\sigma} \tag{1}$$

Normalization relative to the mean is called Cepstral Mean Substraction - CMS and historically it is the one used for the longest period [9].

One of the basic ideas of speaker recognition systems is the transformation of the vectors of the set of distinctive speech feature coefficients into a more general model. This process is called training or adaptation to training coefficients. At least two models are developed in the algorithm - that of a user and that of an impostor. They are used in the testing process, where a statistic comparison is used for the maximized probability of a statistical model of the test utterance belonging to the model of particular users or an impostor - on this basis a decision is made regarding the classification of the user.

Figure 2 shows a simplified diagram of the testing stage of the designed system. Initially distinctive features are extracted from continuous speech of the test utterance and then a statistical comparison of log likelihood coefficients is performed, after calculation of testing speech samples model adaptation to the user models and the impostor model. The results obtained are normalized due to the fact that, in spite of calculations performed earlier (sampling, extraction and formation of a set of distinctive speech feature vectors, GMM modeling, testing), which are sufficient to develop a good system for user recognition, studies show that the effectiveness of the algorithms can be increased by normalizing the resulting coefficients. The last - final stage of the testing phase is to decide on the basis of the test utterance on the selection of the most similar user model, or its rejection, if it will be most similar to impostor model.

Extraction of distinctive speech features in speaker recognition systems

An appropriate design of the extraction algorithm is one of the most important elements of the speaker-recognition system. The purpose of the extraction of distinctive speech features is to achieve appropriate transformation of the input speech signal into a sequence of speech features vectors, where each of them represents information contained in a frame. In other words, features extraction transforms multielemental speech discrete signal into low-level features vector. The creation of the set of distinctive speech feature vectors is performed based on MFCC coefficients [8].

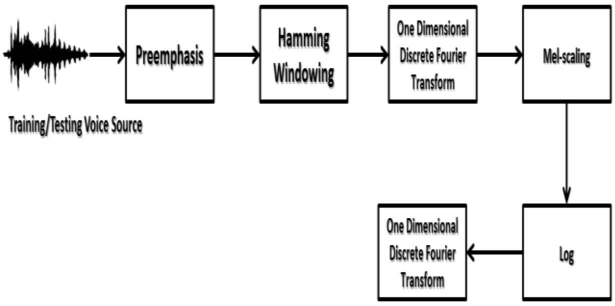
Preemphasis
Hamming Windowing
One Dimensional Discrete Fourier Transform
Mel-scaling
Training/Testing Voice Source
One Dimensional Discrete Fourier Transform
Log

Fig. 3. Diagram of MFCC coefficient creation

Statistical modeling of users

In the case of multivariate Gaussian probability functions the approach to speaker recognition is purely statistical. On the basis of a set number of extracted distinctive speech feature vectors (based on MFCC, calculated on the basis of audio samples) a GMM model is created. This process requires a set number of model parameters to be defined (such as the amount of mixtures, the range of estimated speech feature vectors, etc.).

The algorithm uses the fact that Gaussian distribution can be extended to any dimensional variable x in a domain R^{n}, then it is called the multivariate Gaussian probability function, for which it is possible to write:

$$p(x) = \lim_{\delta\to 0}\int_{x_n-\delta}^{x_n+\delta}\cdots\int_{x_n-\delta}^{x_n+\delta} f_N(z,\mu,\Sigma^2)dz \tag{2}$$

$p(x)$ is the probability of encountering an infinitely small quality coefficient around a multidimensional vector x. Twodimensional distribution is illustrated in figure 4 [10].

It is easy to notice that in the case of a single mixture the set of distribution means is a vector, while the set of variances is described as a $\mathrm{n}\times\mathrm{n}$ dimensional covariance matrix Σ^2. The use of the full covariance matrix makes it possible to create a rotated distribution relative to the expectations. Restricting Σ only to a diagonal matrix (this type of matrix is called the variance matrix) prevents the creation of rotated distributions, simplifying and shortening the calculations performed. GMM models usually use only diagonal variance matrices, as opposed to full covariance matrices, for describing multivariate Gaussian probability density functions [6], [7], [11], [12], [13].

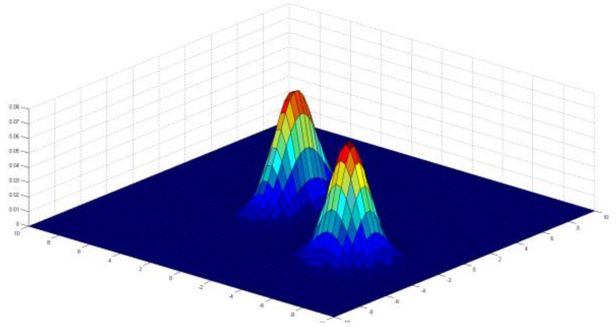

Fig. 4. Two - dimensional GMM distribution

The mixture is a finite sum of N distributions:

(3) $$p_M(x) = \sum_i^N w_i p_i(x)$$

where $\mathrm{p_i}$ signifies single distributions and in the case of GMM - Gaussian distributions. Weights $\mathrm{w_i}$ - determine the impact of a particular single distribution on the whole multivariate distribution. In accordance with probability theory for $0 \leq \mathrm{p_M}, \mathrm{p_i} \leq 1$ the sum of the weights must be 1 (otherwise $\mathrm{p_M(x)}$ it is not a probability distribution). The use of GMM in speaker recognition algorithms requires at least two algorithms.

First, it is necessary to define the way of obtaining each values of the set of distributions first moments, variances and weights to statistically model multi-dimensional value distributions of distinctive speech feature vectors. They are obtained through the EM algorithm (Expectation Maximization - described later in this article) which calculates $3\mathrm{N}$ parameters for the mixture of N distribution, where each distribution is described by three parameters: expectation μ, variance σ^2 and weight w_i.

Secondly, it is necessary to define the method for evaluating distinctive speech feature vectors for particular audio segments in order to identify a particular speaker. Due to the fact that the output result of GMM is a probability value, it is necessary to determine the decision-making method. The simplest and most widely used method is to employ thresholding which for applications in speaker recognition can be described as a process of assigning any value to one of two classes.

Thresholding is a single estimated value independent of the number of users. Usually, [7], [11], [13] a constant threshold is used, user independent, due to the fact that in this case there is no need to compare the statistical assessment of the distinctive speech feature vectors for each user (from a set of authorized users) with all other estimated models for other users.

Likelihood and A Posteriori Probability

During the GMM training and testing process two kinds of probabilities are employed, which represent two different data properties. The first is *Likelihood* $p,(x|H_M)$ - the probability that the data x was generated under the conditions of the Hypothesis H_M. This hypothesis states that the tested user is an authorized user and assumes in the query that the user is authorized through an assigned authorized user model M from among the available authorized user models. In speaker recognition algorithms this probability is described through an inverse relationship $p(H_M|x)$ **-** the resulting *A Posteriori Probability* checking whether the tested user's identity is the true identity (from set of authorized users and impostor model), taking into account the tested segment of audio data (extracted distinctive speech feature vectors). The Bayes Learning Theory defines the relationship between *Likelihood* and *A Posteriori Probability:*

(4) $$p(H_M|x) = \frac{p(x|H_M)p(H_M)}{p(x)}$$

Probability $\mathrm{p(H_M)}$ is *A Priori Probability* which determines the probability of the speaker's tested identity being the true identity. This means that the system is to set higher probability values for "typical" users and lower for users with more sparse values distribution of distinctive speech feature vectors. Due to the fact that speaker recognition systems are designed for a large number of users, it is assumed that this probability is constant for all users, and as it is easy to notice, with the multiplicity of the set of authorized users going towards infinity the *A Priori Probability* goes towards zero. Therefore probability $p(H_M)$ is ignored - taking it into account for a finite set of users is employed merely for scaling. Likelihood $\mathrm{p(x)}$ is the probability of finding features x in a set of all features. That is why typical values of distinctive speech features have an insignificant influence over GMM, contrasted with the rare ones. The most commonly used Neyman - Pearson Lemm's [14] probability $p(x)$ is described later in this article.

EM algorithm

The most often used method for training GMM models is the EM algorithm [15]. The issue of training consists in using a finite number of samples in order to determine optimal parameters for modeling. In the case of speaker recognition the samples are usually segmented into parts in which the user utters certain word sequences. The training of the model is done through iteration within a segment and finding gradually better values of the parameters (sets: weights, averages and variances). Due to the fact that the EM algorithm is not able to achieve final result, a Maximization step must be performed. In speaker recognition applications this is done through Likelihood, or A Posteriori Probability. Therefore, the EM algorithm is one way to find *Maximum Likelihood – ML* or *Maximum A Posteriori Probability - MAP.*

Let us assume that θ describes the parameters of the model, θ^* describes the optimal parameters, θ' describes the new value θ after a single optimizing iteration. Therefore, the EM algorithm iteratively finds θ' which maximizes the probability of finding the analyzed variables values, taking into account the current values of model parameters, so:

(5) $$\theta' = {}^{argmax}{}_{\Theta} E(p(x, Y|\Theta)|x, \Theta)$$

where x is the available data, Y is a random variable describing unknown (hidden) data. The expectation is usually described as $\mathrm{Q(\theta)}$ and calculating its value is the first step in each iteration.

EM algorithm used in multivariate Gaussian probability functions.

Calculating the optimal GMM model parameters through maximization ML [16], as well as MAP [7] is complicated, but can be summarized as the most significant formulas described below.

Calculating the optimal GMM model parameters with Probability Maximization ML is limited to three basic formulas for updating parameters of the mixtures model [16]:

(6) $$\theta = (w_1, \ldots., w_n; \mu_1, \ldots, \mu_n; \sigma_1, \ldots, \sigma_n)$$

(7) $$w_i' = \frac{1}{n}\sum_j^n p(i|x_j, \theta)$$

(8) $$\mu_i' = \frac{\sum_j^n x_j p(i|x_j,\theta)}{\sum_j^n p(i|x_j,\theta)}$$

(9) $$\sigma'^2_i = \frac{\sum_j^n (x_j - \mu_i')^2 p(i|x_j,\theta)}{\sum_j^n p(i|x_j,\theta)}$$

where $\mathrm{w_i}$ is the weight for the i- distribution and $\mathrm{p(i|x_j, \theta_i)}$ is the probability that $\mathrm{x_j}$ was generated from this i-distribution. Therefore, the weighted average is:

(10) $$p\left(i\middle|x_j, \theta\right) = \frac{w_i p_i(x_t|\theta_i)}{\sum_k^M w_k p_k(x_t|,\theta_k)}$$

Calculating the optimal GMM model parameters with Maximum A Posteriori Probability - MAP is done using the formulas [5]:

(11) $$w_i' = \frac{r^w + N_i}{n + N r^w}$$

(12) $$\mu_i' = \alpha_i^\mu E_i(x) + (1 - \alpha_i^\mu)\mu_i$$

(13) $$\sigma_i^{2'} = \alpha_i^\sigma E_i(x^2) + (1 - \alpha_i^\sigma)(\sigma_i^2 + \mu_i^2) - \mu'^2_i$$

(14) $$N_i = \sum_j^n p\left(i\middle|x_j, \theta\right)$$

(15) $$E_i(x) = \frac{1}{N_i} \sum_j^n p(i|x_j, \theta) x_j$$

(16) $$E_i(x^2) = \frac{1}{N_i} \sum_j^n p(i|x_j, \theta) x_j^2$$

where N - is the number of distributions, $\mathrm{r^w}$ and α are constants, described later in this article. The algorithm consists in a simple repetition of three formulas, taking into account the condition of convergence, or a defined number of repetitions. The algorithm of Maximization of A Posteriori Probability is especially sensitive to model overtraining, where during statistical modeling, when multivariate Gaussian probability functions are used, overly detailed data will be taken into account and the model will not fulfill its role properly.

Testing the speaker's set of distinctive speech feature vectors

Once GMM modeling is finished for both the impostor and authorized user, the next step is the testing stage described in figure 2. In this stage the system must be able to recognize the speaker identity and reach a decision at the beginning. In the case of testing the simplest answer could be a single test coefficient for a single voice and model segment. A more sophisticated response is a vector containing test coefficients responsible for different voice characteristics. The most often used are the Likelihood Ratio - LR and the Log Likelihood Ratio - LLR as defined in [17]. This is the optimum test coefficient minimizing the number of errors of the FN type - False Negatives, that is, taking wrong decision on a negative decision given for the tested utterance. It is described by the formula:

(17) $$\Lambda_M = \frac{p(H_M|x,\theta)}{p(\neg H_M|x,\theta)}$$

(18) $$\log \Lambda_M = \log p(H_M|x, \theta) - \log p(\neg H_M|x, \theta)$$

where $p(\neg H_M|x, \theta)$ is the probability that $\mathrm{H_M}$ is not the valid hypothesis for x (M is not an authorized user for his tested utterance). A useful approximation for the LLR testing coefficient: $\mathrm{p(H_M|x)}$ is the most widely used, due to the fact that it is possible to skip all users except the tested M in calculations for the GMM model, which in practice means skipping consideration for all possible human voices. In this case the fact that $\mathrm{H_M}$ is not part of $\neg H_M$ is used. Thanks to this approximation calculations are significantly simplified:

(19) $$\log \Lambda_M = logp(H_M|x, \theta) - logp(H_W|x, \theta)$$

A Posteriori Probability is described by the formula:

(20) $$p(H_M|x) = \frac{p(x|H_M)p(H_M)}{p(x)}$$

where it was mentioned that $p(x)$ in this form is rarely used due to the fact that it requires significantly more information than needed (information about all possible human voices). However, based on the fact that the appropriate number of users (non-authorized) is used as part of training the Impostor Model it is possible to use a simple probability calculation $\mathrm{logp(H_W|x, \theta)}$ instead of $\mathrm{\log p(\neg H_M|x, \theta)}$. Formalizing: $\mathrm{p(H_W) = 1}$ and the final form of the probability:

(21) $$p(x) = p(x|H_W)p(H_W) = p(x|H_W)$$

On this basis the decision process is performed which consists in making a decision regarding the choice of the most similar user model, or its rejection, if it will be most similar to the impostor model.

Description of a practical implementation of the speaker recognition system

The theoretical considerations described earlier were implemented in practice in the form of a C++ application, in the QT environment (version 4.7.4), on the Linux operating system (Mandriva, kernel version 2.6.39.4-5.1) with the purpose of porting the algorithm to the PTT hardware platform. Algorithm was based on opensource ALIZE library [18]. A simplified diagram of the implemented speaker recognition system is illustrated in figure 5.

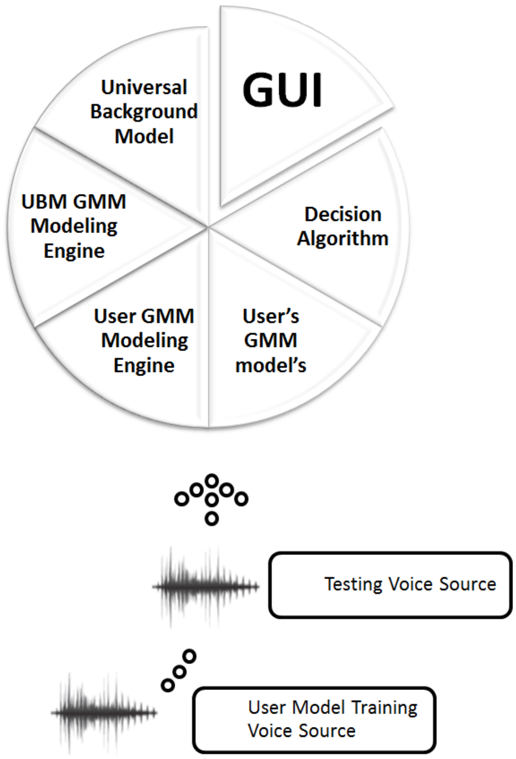

Fig. 5. Simplified diagram of the implemented speaker recognition system

On the basis of the practically implemented speaker recognition system, the effectiveness of the algorithm was tested (using FAR - False Accept Rate and FRR - False Rejection Rate coefficients in order to calculate the EER - Equal Error Rate coefficient of the implemented system), under the following conditions: 50 users (40 men and 10 women) who trained the UBM model, 40 users tested (30 men and 10 women), number of mixtures of GMM UBM mixtures 512, number of mixtures of GMM user models

512, total duration of utterances of users training the UBM model 2.5 hours, duration of training utterances of a single user ~ 180s (numbers repeated in sequence), testing utterance time ~5s (random words found in a dictionary), number of testing utterances 1480, recording sample rate 44100Hz, bits per sample 16, number of channels 2 , number of microphones 7, size of the training and testing dictionary 10 words (digits), resulting efficiency of the algorithm is: EER = 2.13%.

The high accuracy of the speaker recognition results stems mainly from a small number of people tested and high quality recordings and from the fact that the set of words in dictionary was low (10 elements).

Summary

The article presents an algorithm for biometric speaker identity recognition based on continuous speech using multivariate probability distributions. The system is designed to support the functionality of the Personal Trusted Terminal PTT in order to uniquely identify a subscriber using the device. The theoretical model of the statistical modeling algorithm has been accurately described, its practical implementation and test results were shown.

Acknowledgements

This paper has been financed from science funds granted within the years 2010-2012 as a research project of the Polish National Centre for Research and Development of the Republic of Poland No. 0181/R/T00/2010/12.

REFERENCES

[1] Piotrowski Z., Zagoździński L., Gajewski P., Nowosielski L.,Handset with hidden authorization function, *European DSP Education & Research Symposium EDERS (2008)*, Proceedings, 201-205, Texas Instruments

[2] Piotrowski Z., The NNC System and its components in the age of Information Warfare, Safety and Security Engineering III, SAFE III, WIT Press, Southampton, Boston,(2009), 301-309

[3] Davis, S., Merlmestein, P., Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Trans. on ASSP (1980), 357-366

[4] Neiberg D., Text Independent Speaker Verication Using Adapted Gaussian Mixture Models CTT (2001)

[5] Joseph P., Campbell, Jr., Speaker recognition, PII: S 0018-9219(97)06947-8

[6] Reynolds D.A., Rose R.C., Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Process,* 3 (1995), 72–83

[7] Reynolds D.A., Speaker identification and verification using Gaussian mixture speaker models, *Speech Commun.* 17 (1995), 91–108

[8] Rabiner L., Juang B.H., Fundamentals of Speech Recognition, *Prentice-Hall (1993)*

[9] Markel J., Oshika B., Gray Jr. A., Long-term feature averaging for speaker reognition. *ZEEE Transactions on Acoustics, Speech, and Signal Processing,* August (1977),54-61

[10] McLachlan G., Peel D., Finite Mixture Models. Hoboken, NJ: John Wiley & Sons, Inc., (2000)

[11] Reynolds D.A., Automatic speaker recognition using Gaussian mixture speaker models, *Lincoln Lab. J,* 8 (1996), 173–192

[12] Reynolds D.A., Comparison of Background Normalization Methods for Text-Independent Speaker Verification, Proceedings of Eurospeech, (1997), 963–966

[13] Reynolds D.A., Quatieri T.F., Dunn R.B, Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing*, (2000)

[14] Jiang H., Confidence Measures For Speech Recognition*,* Survey A., Speech Communication, *Volume* 45 No. 4 (2005), 455–470

[15] Dempster A.P., Laird N.M., Rubin D.B., Maximum-Likelihood From Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B., 39 (1977)

[16] Bilmes, J., *Gentle A.,* Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, International Computer Science Institute, (1998)

[17] Jiang, H., Confidence Measures For Speech Recognition*,* Survey A., *Speech Communication*, Volume 45 No. 4(2005), 455–470

[18] Bonastre J., Wils F., Meignier S., ALIZE, a free toolkit for speaker recognition, Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, 2005, 737-740

***Authors**: M.Sc. Piotr Lenarczyk, Military University of Technology, Faculty of Electronics, Warsaw, Poland, Tel. +48512127726, Fax. +4822-683-90-3, E-mail piotr.lenarczyk@interia.pl; PhD. Zbigniew Piotrowski, Military University of Technology Faculty of Electronics, Warsaw, Poland, Tel. +4822-683-97-99, Fax. +4822-683-90-3, E-mail zbigniew.piotrowski@wat.edu.pl.*