Zbigniew PIOTROWSKI, Jarosław WOJTUŃ, Kamil KAMIŃSKI

Wojskowa Akademia Techniczna, Wydział Elektroniki

Subscriber authentication using GMM and TMS320C6713DSP

Abstract. *The article presents the theoretical basis for the implementation of Gaussian Mixture Models and implementation of a word recognition system on the basis of DSK TMS302C6713 DSP from Texas Instruments. The effectiveness of the algorithm based on Gaussian Mixture Model has been demonstrated. The system was developed as a software module for voice authentication of a subscriber in a Personal Trusted Terminal (PTT). The PIN of a subscriber is verified through an utterance in the Personal Trusted Terminal.*

Streszczenie. *W artykule zaprezentowano teoretyczne podstawy realizacji Modeli Mikstur Gausowskich oraz implementację systemu rozpoznawania słów z wykorzystaniem zastawu uruchomieniowego DSK TMS302C6713 DSP firmy Texas Instruments. Zobrazowano skuteczność działania algorytmu opartego na Modelach Mikstur Gausowskich. System został opracowany jako moduł programowy na potrzeby głosowego uwierzytelniania abonenta w Osobistym Zaufanym Terminalu (PTT). Poprzez wypowiedzenie głosem swojego PIN-u abonent jest weryfikowany w Osobistym Zaufanym Terminalu.* ***(Uwierzytelnianie abonenta z wykorzystaniem GMM oraz TMS320C6713DSP)***

Keywords: speaker recognition, mel-frequency cepstral coefficient, gaussian mixture model, DSK C6713 DSP, Personal Trusted Terminal.
Słowa kluczowe: rozpoznawanie mowy, współczynniki mel-cepstralne, modele mikstur gaussowskich, DSK C6713 DSP, Osobisty Zaufany Terminal, uwierzytelnianie abonenta.

Introduction

Subscriber authentication by voice in a telecommunications system means a process of assigning a voice to the subscriber whose voice profile was determined in the system beforehand. In the described system it is not the subscriber's voice profile that is recognized, but a set of short utterances of digits. The set of digits represents the subscriber's Personal Identification Number. Therefore, the Subscriber in this case is authenticated through digits spoken by him or her.

Subscriber authentication in a telephone terminal usually consists in entering into the terminal, using the keyboard, an appropriate sequence of digits (PIN). The entered PIN is compared in the terminal with the previously memorized PIN number and in case of correspondence of the two PINs, resources of the terminal are made available.

Evaluation board kits of the DSK C6713 series are dedicated to processing a speech signal in real time. For example, a solution for the problem of speech signal segmentation using an evaluation board kit using digital signal processing, where the speech signal of the subscriber is determined in real time [1] is known. In the described authentication system prototyping was used with the DSK C6713 set due to the speed of performed operations and the ease of prototyping of DSP algorithms. The developed system was eventually used as a software module in the Personal Trusted Terminal [2]. The PTT acts as a digital watermark token on radio links. A verification module of the subscriber willing to use the terminal was added to the PTT on the basis of recognizing digits uttered by the subscriber.

Work conducted up to date on the recognition of digits for Polish speakers [3] and recognition of short phrases [4] indicates the high efficiency of algorithms based on Hidden Markov Models as well as Gaussian Mixture Models. Mixed methods are also known, e.g. combining GMM models with the Support Vector Data Description (SVDD) [5] method. Representative speech signal parameters are isolated for the classification of common features of voice utterances, such as in the scope of the cepstrum [6]. The described authentication system uses mel-frequency cepstral coefficients, signal energy coefficients, as well as derivatives of these coefficients.

Extraction of speech features

A direct comparison of signals in the case of speech recognition has proved in practice not to be overly effective. The solution to this problem is to compare the characteristic features of the signal. There are many approaches to determine the characteristics of the speech signal from the point of view of digital signal processing. The most popular way to determine the characteristic features of a speech signal is to calculate the Mel-Frequency Cepstral Coefficients (MFCC) [7]. Another frequent solution is to determine Linear Predictive Coefficients (LPCC) and coefficients determined on their basis of - Partial Correlation Coefficient (PARCOR), Linear Predictive Cepstral Coefficients (LPCC) [8]. Moreover, literature on the subject mentions many other proposals for determining the characteristic features of the speech signal, for example Mel-Frequency Discrete Wavelet Coefficients (MFDWC) [9, 10], Wavelet Octave Coefficients Of Residues (WOCOR) [11], Mel Cepstrum Modulation Spectrum (MCMS) [12]. From among the demonstrated features, some are applicable in recognizing words, others in identifying speakers. In this paper we will confine ourselves to the description of the method of deriving Mel-Frequency Cepstral Coefficients (MFCC) and differential coefficients designated on their basis, known from literature as Delta and Delta – Delta Cepstral Coefficients [7].

First, the speech signal being analyzed is processed to remove silent fragments at the beginning and end of the recording. Thanks to this operation it is possible to reduce the number of signal samples analyzed, which improves the speed of voice recognition of phrases. The elimination of silent fragments also increases the effectiveness of recognition, as this makes particular phrases more audibly different from each other, which facilitates identification by the system. The proposed algorithm for elimination of silent fragments is based on the criterion of signal energy in a couple of initial and final signal frames [13]. Then, the signal is divided into frames with a length of 16 ms with a 50% overlay (for a sampling rate of 8000 Hz, a frame contains 128 samples). In the next step the discrete Fourier transform is calculated for the signal frame multiplied with a Hamming window [14]:

$$F(k,\tau)=\frac{1}{\sqrt{M}}\sum_{t=1}^{M-1}\left[x(t+\tau)e^{-j\frac{2\pi kt}{M}}\cdot w_\tau(t)\right], \quad k=0,\ldots,M\text{-}1 \tag{1}$$

$$w_\tau(t)=0,54-0,46\cos\left(\frac{2\pi t}{M-1}\right), \quad t=\tau+0,\tau+1,\ldots\tau+M-1 \tag{2}$$

The next step is to determine the signal spectrum modulus:

$$FC(k,\tau)=\sqrt{F(k,\tau)\cdot F^*(k,\tau)},\ k=0,\ldots,M\text{-}1 \tag{3}$$

After the above operations are performed, the transition from a frequency expressed in hertz into a frequency expressed in mels is performed. The mel scale was experimentally defined by Stevens S.S., Valkman J.E. and Newnam E.B. in the 1930s [7]. The mel scale is used because the human ear responds in a non-linear way to frequencies of an audio signal - the differences in sound levels are more easily discernible in the case of lower frequencies. The relationship between the mel scale and frequency expressed in hertz is described (4) and shown in Figure 1.

$$f_{Mel} = 2595 \log\left(1 + \frac{f[Hz]}{700}\right) \tag{4}$$

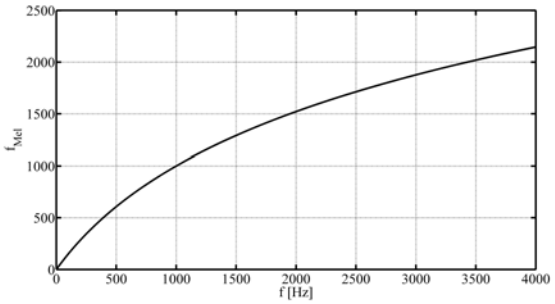

Fig. 1. Mapping frequencies in hertz to a mel scale

Using the mel scale we create a bank of 26 triangular band-pass filters. Within this scale these filters are identical, symmetrical triangles with a base width of 160 mels, shifted with a 50% overlap. After applying an inverse relation to (4), i.e. transitioning from a mel scale to a frequency expressed in hertz we obtain triangular (already assymetrical) basis functions depicted in Figure 2.

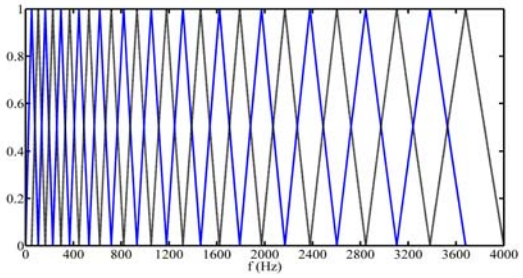

Fig. 2. Triangular filter bank

The next step is to determine Mel-Frequency Coefficients (MFC), which is done by multiplying the triangular basis functions with the estimate of spectral density of the signal:

$$MFC(l,\tau) = \sum_{k=0}^{M-1} \left[D(l,k) \cdot FC(k,\tau) \right], \; l = 0,\ldots,L-1 \tag{5}$$

where: L - number of Mel-Frequency Coefficients.

The last stage in determining Mel-Frequency Cepstral Coefficients (MFCC) is to perform an inverse discrete Fourier transform (in practice this is performed by executing a discrete cosine transform DCT):

$$MFCC(k,\tau) = \sum_{l=0}^{L-1} \left[\log\left(MFC(l,\tau)\right) \cdot \cos\left(\frac{k\pi(2l+1)}{2L}\right) \right], \tag{6}$$
$$k = 0,\ldots,K-1$$

where: K - number of Mel-Frequency Cepstral Coefficients.

In order to increase the efficiency of phone recognition, additional post processing is applied called liftering (filtering in the cepstrum), which involves removing the negative effects of laryngeal frequency and its harmonics on the set of features. This filtering is performed through the use of a sinusoidal window [7]:

$$w_k = 1 + \frac{K}{2} \sin\left(\frac{k\pi}{K}\right) \tag{7}$$

For speech recognition systems the first 12-13 MFCC coefficients are analyzed first, as all higher coefficients are closely dependent of laryngeal frequency of the speaker and they are used primarily to identify speakers.

In the next step the set of features is increased by an energy factor:

$$E(\tau) = \sum_{t=1}^{M-1} x^2(t+\tau) \tag{8}$$

Differential coefficients are used in the form of dynamic information, adding to the feature set of a frame. A regression is used covering 5 consecutive frames for approximation of each of the Mel-Frequency Cepstral Coefficients over time:

$$\Delta C(k,\tau) = \frac{2C(k,\tau-2) + C(k,\tau-1) + C(k,\tau+1) + 2C(k,\tau+2)}{6} \tag{9}$$

where $C(k, \tau)$ means $MFCC(k, \tau)$.

It is also possible to perform the same action for the coefficient ΔC. We then obtain $\Delta\Delta C$, as well as for the energy coefficient by obtaining, respectively, ΔE and $\Delta\Delta E$. Differential coefficients help in distinguishing individual vowels (monophthongs) from double vowels (diphthongs). Besides, they are characteristic for transitions between successive phones.

As a result of the extraction of speech signal features, each frame is assigned a 39-element set $\{12xMFCC, 12x\Delta MFCC, 12x\Delta\Delta MFCC, E, \Delta E, \Delta\Delta E\}$.

Gaussian Mixture Models

Recognition of individual words will currently mean comparing associated sets of cepstral coefficients with sets of cepstral coefficients for model words contained in the database [13]. The simplest and historically oldest method of recognizing words is Dynamic Time Wrapping (DWT) [15]. Currently, algorithms based on Gaussian Mixture Models (GMM) [16, 17], Hidden Markov Models (HMM) [18], Support Vector Machines (SVM) [19] or Artificial Neural Networks (ANN) [20] are more commonly used. Literature knows also examples of hybrid models [5, 21]. The paper will present the method for use of Gaussian Mixture Models for speech recognition of Polish.

The Gaussian Mixture Model is a parametric probability density function which is represented by the weighted sums of Gaussian distributions. Let $X=\{x_1, x_2, \ldots, x_T\}$ be D – a dimensional set of T – cells containing data matrices (in our case these will be the earlier designated 39-element sets of speech signal features). Then the probability density function appears as follows [22]:

$$p(x_t \mid \lambda) = \sum_{i=1}^{M} w_i g(x_t \mid \mu_i, \Sigma_i) \tag{10}$$

where λ represents a set of parameters consisting of $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i meaning the weight of individual distributions, while $g(x_t|\mu_i,\Sigma_i)$ is a multi-dimensional Gaussian distribution with the expected value μ_i and a covariance matrix Σ_i. The function describing the Gaussian distribution function is defined in the following way [22]:

$$g(x_t \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} \sqrt{|\Sigma_i|}} \exp\left(\frac{-(x_t - \mu_i)^2}{2\Sigma_i}\right) \tag{11}$$

In addition it is required that [22]:

$$\sum_{i=1}^{M} w_i = 1 \tag{12}$$

Having training vectors (in our case this will be the prepared earlier database of model phrases – digits from 0 to 9) and wishing to use GMM, it is necessary to estimate the parameter set $\lambda = \{w, \mu, \Sigma\}$. Values of model parameters can be determined through various methods [22]. Most often they are determined in accordance with the principle of the Maximum Likelihood (ML) estimator. For a set of learning samples X the GMM likelihood is [22]:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda) \tag{13}$$

The objective of ML estimation is to define such new parameters of the model $\overline{\lambda}$ that there occurs the relationship:

$$p(X \mid \overline{\lambda}) \geq p(X \mid \lambda) \tag{14}$$

This stems from the fact that expression (14) is a nonlinear function of parameters λ and direct maximization is not possible. However, parameters can be obtained iteratively. An example of a simple iterative algorithm for estimating parameters is the algorithm of Expectation Maximization (EM) which aims at maximizing the likelihood function of the model with a given set of learning data [23]. In each iteration, new parameters for the model are designated in order to increase the likelihood of the model. They are determined on the basis of the relationship [22]:

$$\overline{w_i} = \frac{1}{T}\sum_{i=1}^{T} P_r(i \mid x_t, \lambda) \tag{15}$$

$$\overline{\mu_i} = \frac{\sum_{i=1}^{T} P_r(i \mid x_t, \lambda) x_t}{\sum_{i=1}^{T} P_r(i \mid x_t, \lambda)} \tag{16}$$

$$\overline{\sigma_i^2} = \frac{\sum_{i=1}^{T} P_r(i \mid x_t, \lambda) x_t}{\sum_{i=1}^{T} P_r(i \mid x_t, \lambda)} - \overline{\mu_i^2} \tag{17}$$

where $P_r(i|x_t,\lambda)$ is the a posteriori probability of the component i (in the case of recognition of digits, the individual digits are the components) and is expressed by the formula:

$$P_r(i \mid x_t, \lambda) = \frac{w_i g(x_t \mid \mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k g(x_t \mid \mu_k, \Sigma_k)} \tag{18}$$

Let λ_k for $k=1,...,N$ indicate models of words, where N is the number of different words in the dictionary (in our case this will be 10 digits). The classifier is designed in such a way as to find the appurtenance of vector X (vector of features of the recognized word) to one of the N models of words, using the discrimination function $g_k(X)$. For this purpose, likelihood is calculated between the unknown vector X and each of the models of words λ_k, and then the model λ_k^* is selected that meets the criterion [24]:

$$k^* = \arg\max_{1 \leq k \leq N} \left(g_k(X) \right) \tag{19}$$

The discrimination function is determined from the following relationship *a posteriori*:

$$g_k(X) = p(\lambda_k \mid X) \tag{20}$$

using Bayes' rule it can be written [25]:

$$p(\lambda_k \mid X) = \frac{p(\lambda_k)\, p(X \mid \lambda_k)}{p(X)} \tag{21}$$

Assuming that every word is equally likely, i.e.: $p(\lambda_k)=1/N$ and that $p(X)$ is the same for all models of words, it can be written [24]:

$$g_k(X) = p(X \mid \lambda_k) \tag{22}$$

Finally, a decision is made regarding the identification of a word base on log-likelihood [24]:

$$k^* = \arg\max_{1 \leq k \leq N} \left(\sum_{t=1}^{T} \log\left(p(x_t \mid \lambda_k) \right) \right) \tag{23}$$

where $p(x_t|\lambda_k)$ is determined in accordance with (10).
The value k^* for the maximum of the last expression corresponds to the model representing the recognized word.

Figure 3 shows a graphically described method for classification of the feature vector X in relation to the model λ_k. The value of the likelihood logarithm between several realizations of zero and each word model λ_k was illustrated.

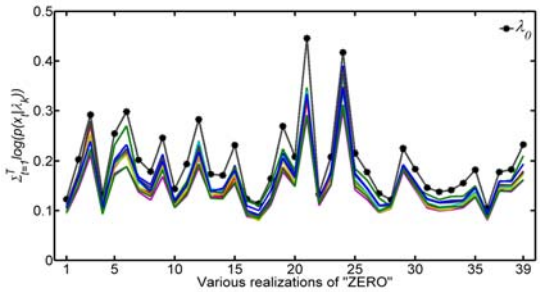

Fig. 3. Likelihood logarithm between realizations of zero and each word model λ_k

Implementation and results

The above algorithms were implemented in the CCSv4 (Code Composer Studio) environment on the basis of the TMS320C6713 signal processor using the DSK 6713 evaluation board kit. Figure 3 shows the DSP test bed used for research in word recognition.

Fig. 4. DSP test bed for word recognition

Figure 5 demonstrates the effectiveness of recognition of individual voice phrases. The study was conducted on a group of 20 people. Training vectors were obtained on the basis of a database of recordings with 300 different realizations for each of the digits. At the GMM training stage, vectors of features for individual numbers were

estimated using 8 Gaussian distributions. An increase in the number of distributions results in an improvement of the effectiveness of recognition, but this entails a significant increase in the number of calculations and translates into an increase in waiting time for recognition results.

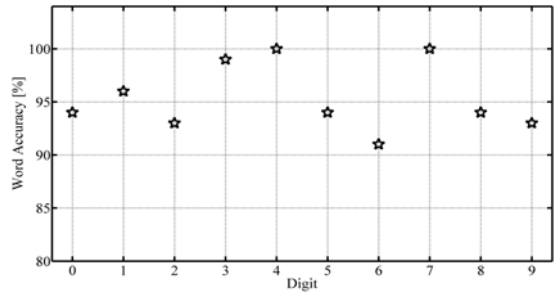

Fig. 5. Effectiveness of word recognition

In user verification systems there are four variants for a decision: correct acceptance, correct rejection, false acceptance and false rejection [26]. The last two variants describe erroneous system functioning. False acceptance occurs when spoken digits that are not the PIN code are validated by the system as the correct PIN. False rejection is, on the other hand, the opposite situation - a correct PIN is rejected. Figure 6 shows the relation of the false acceptance rate and false rejection rate as a function of the likelihood logarithm normalized to the maximum of this likelihood. The graph shows that the optimal detection threshold is the maximum value of the likelihood logarithm, which is the justification for the choice of such a criterion in equation 23. In addition, it is worth noting the steepness of the characteristics in the area of the maximum of the likelihood logarithm, which indicates the high selectivity of the presented algorithm - for a likelihood threshold determined in this way, the probability of FAR and FRR errors is minimized.

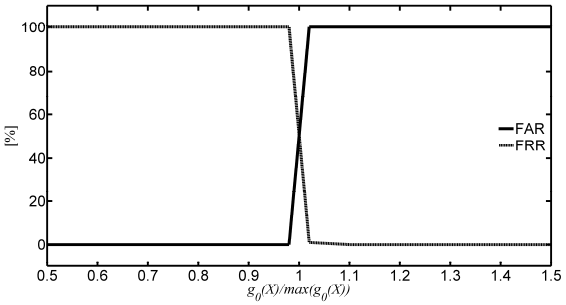

Fig. 6. FAR and FRR as a function of the normalized detection threshold

Summary

The paper presents an algorithm for word recognition using Gaussian Mixture Models. The results of the effectiveness of word recognition and a method for using the algorithm for user authentication have been shown. Further work should be aimed at developing an algorithm for speaker identification. Then both algorithms could be combined and two-step user identification be performed: identifying the speaker and recognizing words spoken by him or her.

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 No. 0181/R/T00/2010/12.

REFERENCES

[1] Dąbrowski A., Pawłowski P., Weychan R., Mayer A., Portalski M., Chmielewska A, Janiak T.: Real-time watermarking of one side of telephone conversation for speaker segmentation, *Electrical Review*, vol. 88, no. 6/2012

[2] Piotrowski Z.: The National Network-Centric System and its components in the age of Information Warfare, Safety and Security Engineering III, SAFE III, WIT Press 2009, Southampton, 301-309

[3] Dulas J., Automatyczna identyfikacja cyfr dla mówców polskojęzycznych, *Przeglad Elektrotechniczny*, vol. 86 no. 5/2010

[4] Marciniak T., Krzykowska A., Weychan R.: Speaker recognition based on telephone quality short Polish sequences with removed silence, *Przeglad Elektrotechniczny*, vol. 88, no. 6/2012

[5] Zhou Y., Zhang X., Wang J., Gong Y., Zhou Y.: Speaker recognition based on the combination of GMM and SVDD, *Przeglad Elektrotechniczny*, vol. 87, no. 3/2011

[6] Dobrowolski A., Majda E.: Application of homomorphic methods of speech signal processing in speakers recognition system, *Przeglad Elektrotechniczny*, vol. 88, no. 6/2012

[7] Beigi H., Fundamentals of Speaker Recognition, Springer, 2011

[8] Rabiner L., Juang B. H.: Fundamentals of Speech Recognition, PTR Prentice Hall, New Jersey, 1990

[9] Gowdy J. N., Tufekci Z.: Mel-scaled discrete wavelet coefficients for speech recognition, *Acoustics, Speech, and Signal Processing, 2000, Proceedings. (ICASSP '00)*. vol. 3, (2000) 1351-1354

[10] Tufekci, Z., Gurbuz S.: Noise Robust Speaker Verification Using Mel-Frequency Discrete Wavelet Coefficients and Parallel Model Compensation, *Acoustics, Speech, and Signal Processing, 2005, Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, (2005) 657-660

[11] Chan, W. N., Zheng, N., Lee T.: Discrimination Power of Vocal Source and Vocal Tract Related Features for Speaker Segmentation, *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15(6), (2007), 1884-1892

[12] Tyagi V., Mccowan L., Misra H., Bourlard H.: Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR, *Automatic Speech Recognition and Understanding (ASRU '03)*, IEEE Workshop (2003)

[13] Zieliński T.: Cyfrowe przetwarzanie sygnałów – od teorii do zastosowań, Wydawnictwa Komunikacji i Łączności, 2009

[14] Kasprzak W.: Rozpoznawanie obrazów i sygnałów mowy, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2009

[15] Myers C., Rabiner L. R., Rosenberg A. E.: Performance Tradeoffs in Dynamic Time Wraping Algorithms for Isolated Word Recognition, *Acoustic, Speech and Signal Processing, IEEE Transactions on*, vol. 6, (1980), 623-635

[16] Reynolds D. A., Quatieri T. F., Dunn R. B.: Speaker Verification Using Adapteg Gaussian Mixture Models, *Digital Signal Processing*, vol. 10, Elsevier, (2000)

[17] Kumar G. S., Raju K. A., Cpvjn M. R., Satheest P.: Speaker Recognition Using GMM, *International Journal of Engineering Science and Technology*, vol. 2, (2010), 2428-2436

[18] Juang B. H., Rabiner L. R.: Hidden Markov Models for Speech Recognition, *Technometrics*, vol. 33, (1991), 251-272

[19] Ganapathiraju A., Hamaker J. E., Picone J.: Applications of Support Vector Machines to Speech Recognition, *IEEE Transaction on Signal Processing*, vol. 52, (2004)

[20] Othman A. M., Riadh M. H.: Speech Recognition Using Scaly Neural Networks, *World Academy of Science, Engineering and Technology*, vol. 38, (2008)

[21] Bansal P., Kant A., Kumar S., Sharda A., Gupta S.: Improved Hybrid Model of HMM/GMM for Speech Recognition, Intelligent Information and Engineering System INFOS 2008, (2008)

[22] Reynolds D. A.: Gaussian Mixture Models, *Encyclopedia of Biometric Recognition*, Springer, (2008)

[23] 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.*, vol. 39, (1977), 1-38

[24] Reynolds D. A., Rose R. C.: Robust Text-Independent Speaker Identification Usign Gaussian Mixture Speaker Model, *Speech and Audio Processing, IEEE Transactions on*, vol. 3, (1995)

[25] Bronshtein I. N., Semendyayev K. A., Musiol G., Muehlig H.: Handbook of Mathematics, Springer, Berlin, 2007

[26] Jain A. K., Ross A., Prabhakar S.: An introduction to biometric recognition, *IEEE Transactions on Circuits And Systems For Video Technology*, vol. 14, (2004)

***Autorzy**: dr inż. Zbigniew Piotrowski, E-mail: zpiotrowski@wat.edu.pl, mgr inż. Jarosław Wojtuń E-mail: jwojtun@wat.edu.pl, mgr inż. Kamil Kamiński, E-mail: kkw.kaminski@gmail.com, Wojskowa Akademia Techniczna, Wydział Elektroniki, Instytut Telekomunikacji, ul. Kaliskiego 2, 00-908 Warszawa 49.*