Jozef Polacky - Peter Pocta - Roman Jarina *

# AN IMPACT OF NARROWBAND SPEECH CODEC MISMATCH ON A PERFORMANCE OF GMM-UBM SPEAKER RECOGNITION OVER TELECOMMUNICATION CHANNEL

*The automatic identification of person's identity from their voice is a part of modern telecommunication services. In order to execute the identification task, speech signal has to be transmitted to a remote server. So a performance of the recognition/identification system can be influenced by various distortions that occur when transmitting speech signal through a communication channel. This paper studies an effect of telecommunication channel, particularly commonly used narrowband (NB) speech codecs in current telecommunication networks, on a performance of automatic speaker recognition in the context of a channel/codec mismatch between enrollment and test utterances. An influence of speech coding on speaker identification is assessed by using the reference GMM-UBM method. The results show that the partially mismatched scenario offers better results than the fully matched scenario when speaker recognition is done on speech utterances degraded by the different NB codecs. Moreover, deploying EVS and G.711 codecs in a training process of the recognition system provides the best success rate in the fully mismatched scenario. It should be noted here that the both EVS and G.711codecs offer the best speech quality among the codecs deployed in this study. This finding also fully corresponds with the finding presented by Janicki & Staroszczyk in [1] focusing on other speech codecs.*

***Keywords:*** *Speaker identification, GMM-UBM, MFCC features, TIMIT, Speech codecs, Narrowband voice transmission.*

## 1. Introduction

Over the past decades, Automatic Speaker Recognition (ASR) has become a very popular area of research in pattern recognition and machine learning. Scientists from around the world have been constantly working on improving speaker recognition systems and have also been looking for more effective procedures, which increase the actual recognition rate. ASR is a general term for both speaker identification and speaker verification tasks. A principle of a speaker identification and verification is displayed in Figs. 1 and 2 respectively.

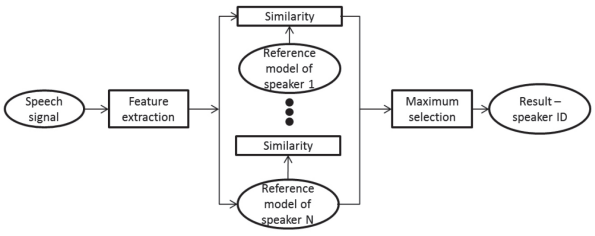


*Fig. 1 Speaker identification*

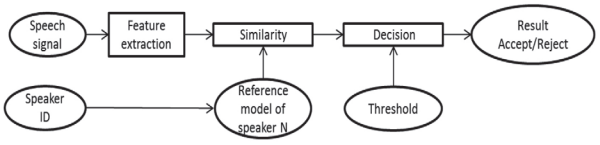


*Fig. 2 Speaker verification*

ASR technique can be used to verify speaker´s identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information and remote access to computers. Most of the above listed applications require a transmission of the user's voice to the remote server that executes an identity validation. Gaussian Mixture Model -Universal Background Model (GMM-UBM) technique represents the currently most popular technique for this task. Its performance has been tested under different conditions in [2 - 4]. ASR research is currently also focused on reducing the within-speaker variations which are caused by a channel mismatch. The channel mismatch occurs when the utterances for enrollment and the utterances for testing are transmitted through channels with different characteristics.

* **Jozef Polacky, Peter Pocta, Roman Jarina**
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia
E-mail: jozef.polacky@fel.uniza.sk

A speech compression plays a significant role in mobile communication, Voice Over Internet Protocol (VoIP), voicemail and gaming communication. In all of the above mentioned cases, lossy speech codecs are deployed. A purpose of speech codec is to compress a speech signal by reducing a number of bits needed for its transmission while maintaining the intelligibility of speech once decoded. The distortions introduced by speech codecs may have a significant impact on a performance of speaker recognition system. An importance of this problem even increases when it comes to the channel mismatch as there is a huge difference between an impact of different codecs deployed in current telecommunication networks on a performance of speaker recognition system. Moreover, it is worth noting here that codec degradations are currently considered as one of the most prominent degradations encountered in current telecommunication networks. Therefore, an analysis of codec-induced degradations in the context of speaker recognition and development of the ASR techniques that are robust against this type of degradations, are of great interest to researchers around the world.

An effect of codec mismatch on a performance of a speaker recognition system has been investigated with different classifiers such as Hidden Markov Models (HMM) [5], GMM-UBM [6], Support Vector Machine (SVM) systems [1] and i-vector techniques [7]. The experiments published in [5] have revealed a significant degradation of performance under mismatched conditions: Pulse-code modulation (PCM) data–trained models vs. Code-excited linera prediction (CELP) coded test data. Finally, two techniques for improving the performance in these situations were examined, namely the maximum a posteriori (MAP) adaptation strategy and the Affine transform strategy. Significant improvements in the performance over mismatched conditions for both cases were achieved. In [6], the authors have demonstrated that a performance of speaker verification system based on GMM-UBM decreases, for all the codecs involved in the study, as the degree of mismatch between training and testing conditions increases. Both the fully matched and mismatched conditions have been investigated in [1] deploying SVM. In the mismatched conditions, Speex codec was shown to perform best for creating robust speaker models. The authors in [7] have focused on the benefits provided by an extended bandwidth used for voice communication in the context of codec mismatch and bandwidth mismatch, using i-vector approach for speaker recognition. Their results show that the performance of ASR on wideband data is significantly better than that employing narrowband (NB) signals for both matched and codec-mismatched conditions.

In this paper, we also focus on the influence of codec mismatch between an enrollment and test utterances. In comparison to the previous studies [1, 5, 6 and 7], we have expanded a set of codecs used for an investigation in terms of their setups, and types of signal degradations introduced by the codecs. The codecs under the tests represent NB speech codecs commonly implemented in current telecommunication networks. We also included a brand new 3GPP EVS codec standardized recently [8]. As a back-end we have applied GMM-UBM classifier as it can provide high recognition rate and is very easy-to-implement [9 and 10]. It overcomes also more sophisticated i-vector approach if only constrained amount of training data is available [11].

The rest of the paper is organized as follows: In Section 2, GMM-UBM approach for speaker's enrollment in speaker identification system is presented. Section 3 describes experiment and experimental results. Finally, Section 4 concludes the paper and suggests a future work.

## 2. GMM-UBM identification approach

GMMs used in combination with MAP adaptation represent the main technology of most of the state-of-the-art text-independent speaker recognition systems [9 and 12]. GMM based speaker-specific models are derived from Universal Background Model - i.e. a generic speaker-independent GMM statistical model, which has been trained on a great amount of multi-speaker speech data.

### 2.1 Gaussian Mixture Model

Gaussian Mixture Model (GMM) [13] is a stochastic model, which can be considered as a reference method for speaker recognition. The Gaussian mixture probability density function of model λ consists of a sum of $K$ weighted component densities, given by the following equation:

$$p(x \mid \lambda) = \sum_{k=1}^{K} P_k N(x \mid \mu_k, \Sigma_k) \tag{1}$$

where $K$ is the number of Gaussian components, $P_k$ is the prior probability (mixture weight) of the $k$-th Gaussian component, and

$$N(x \mid \mu_k, \Sigma_k) = \\ = (2\pi)^{-\frac{d}{2}} \mid \Sigma_k \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\} \tag{2}$$

is the $d$-variate Gaussian density function with mean vector $\mu_k$ and covariance matrix $\sum_k$. The prior probabilities $P_k \geq 0$ are constrained as $\sum_{k=1}^{K} P_k = 1$.

For numerical and computational reasons, the covariance matrices of the GMM are usually diagonal. Training a GMM consists of estimating the parameters $\lambda = \{P_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ from a training sample $X = \{\vec{x}_1, .... \vec{x}_T\}$. The basic approach is to maximize likelihood of $X$ with respect to model $\lambda$ defined as:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) \quad (3)$$

The goal is to obtain Maximum-likelihood (ML) parameter estimation. The process is an iterative calculation called the Expectation-Maximization (EM) algorithm [14]. Note that K-means [15] can be used as an initialization method for EM algorithm.

In the identification process, a set of test utterances and its model is compared with each model of the training database. From each comparison between test and training model is obtained a likelihood and the model with the highest score corresponds to the unknown speaker.

Let us assume a group of speakers $S_p$=*1,2,...S* represented by GMM´s $\lambda_1, \lambda_2, ...\lambda_S$. Unknown speaker model is identified to each model:

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log(p(\vec{x}_t \mid \lambda_k)) \quad (4)$$

## 2.2 Universal Background Model

Universal Background Model (UBM) is an improvement in the field of speaker recognition using GMM. It is typically characterized as a single Gaussian Mixture Model trained with a large set of speakers using the EM algorithm.

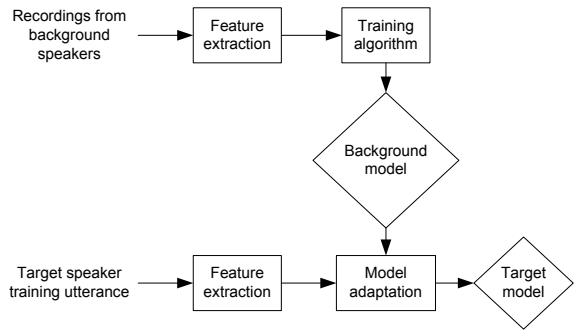


*Fig. 3 UBM adaptation based speaker enrollment*

As shown in Fig. 3 UBM is used as an initial model for training speaker-specific GMM during speaker enrollment. This process prevents from need for estimating the parameters of the speaker model from scratch. There are multiple ways how to adapt the UBM. It is possible to adapt one or more of its parameters as well as all parameters. Adaptation of the parameters is usually done using MAP technique. The background model (UBM) must be built from utterances with common characteristics in the meaning of type and quality of speech, uttered by great number of speakers. For example, an identification/verification system that uses only telephone channel and female speakers must be trained using only telephone speech spoken by female speakers. For a system where the gender composition is an unknown parameter, the model will be trained using both male and female utterances.

The following average log-likelihood formula gives the final score in recognition process [16]:

$$LLR(X, \lambda_{target}, \lambda_{UBM}) = \\ = \frac{1}{T} \sum_{t=1}^{T} \{\log(p(\vec{x}_t \mid \lambda_{target})) - \log(p(\vec{x}_t \mid \lambda_{UBM}))\} \quad (5)$$

Where $X = \{\vec{x}_1, .....\vec{x}_T\}$ corresponds to the set of observation or test feature vectors. The higher the score, the more likely the test features belong to the speaker-model with which they are compared.

## 3. Experiment

The experiment consists of three different scenarios with aim to examine an effect of codec-based speech distortion on a performance of the speaker identification. In the first scenario, speaker enrollment is performed on clean speech while testing is done on degraded speech affected by a codec (named as *partially mismatched* case hereinafter). In the *fully matched* case, both training (speaker enrollment) and testing are carried out on speech data coded by the same codec. In the third scenario, training and testing are carried out also on coded speech but coded by different type of codecs (assigned as *fully mismatched* case hereinafter). In all the scenarios, speaker models were obtained by MAP based adaptation of the UBM model, which was composed of 512 Gaussians. For UBM training, EM algorithm was used, a relevance factor was set to 10, and K-means algorithm with 100 iteration was applied. Note that all of the GMM parameters (i.e. weights, means and covariance matrices) were modified during UBM-GMM.

In order to achieve a high performance of speaker recognition system, the system has to be based on powerful statistical models whose parameters have to be derived by using an adequate amount of training data. Therefore, we have decided to use a widely known and acoustically and phonetically rich TIMIT database [17] containing recordings of phonetically-balanced English speech of 630 speakers of eight major dialects regions of the United States (each reading ten phonetically rich sentences resulting in 6300 sentences). Approximately 70% of the speakers are male and rests are female. It is worth noting here that this database, in spite of its original design intention (designed for a speech recognition), is currently also widely used for a speaker recognition [7 and 18]. As a sampling rate of the recordings is 16 kHz, we have downsampled all of them to 8 kHz via an anti-aliasing low-pass FIR filter with no further processing, with the aim to stick to a sampling rate of NB speech communication. Furthermore, the speech samples were coded by the following codecs at the specified bit rates to

Summary of results for individual conditions Table 1

| | | G.711 | G.729 | AMR_5.9 | AMR_7.95 | AMR_12.2 | EVS_5.9 | EVS_8 | EVS_13.2 |
|---|---|---|---|---|---|---|---|---|---|
| partially mismatched | success rate[%] | 76.00 | 63.00 | 60.11 | 62.95 | 68.53 | 67.53 | 71.37 | 74.47 |
| fully matched | success rate[%] | 78.79 | 49.00 | 39.84 | 44.32 | 36.58 | 69.53 | 67.32 | 71.74 |
| partially mismatched | variance[%] | 6.16 | 10.75 | 27.91 | 19.12 | 12.68 | 17.73 | 10.32 | 8.27 |
| fully matched | variance[%] | 7.93 | 12.60 | 13.80 | 13.23 | 20.51 | 14.43 | 25.07 | 11.30 |

introduce codec specific degradations induced when speech is transmitted over telecommunication network:

- G.711 speech codec [19] (a typical PCM (Pulse-Code Modulation) speech codec) operating at 64 kbps
- G.729 speech codec [20] (a very popular parametric codec dominantly deployed in fixed networks) operating at 8kbps
- AMR-NB [21] speech codec (typically deployed in 3G mobile networks) operating at 5.9kbps, 7.95kbps and 12.2kbps
- EVS speech codec [8] (a brand new 3GPP codec recently standardized by 3GPP and designed to be deployed in 4G (LTE) networks) operating at 5.9 kbps, 8 kbps and 13.2 kbps

Remark that codec degradations are currently considered as one of most prominent degradations encountered in current telecommunication networks. Moreover, the codecs selected for the experiment represent the ones commonly used in current telecommunication networks and also cover all range of degradations currently introduced by NB codecs. The selected bit rates cover the most popular ones.

All speakers in total (630) were used for the UBM training:10 clean (uncoded) recordings and 8 coded speech recordings per speaker were deployed. Note that each of these eight coded recordings was coded by different codec or the same codec operating at different bit rate, as listed above.

For the speaker enrollment (i.e. speaker specific GMM adaptation) phase as well asthe testing phase, 190 speakers from the first three dialect regions of the "training part" of the TIMIT database were used in each scenario. A set of utterances (folder of 10 audio recordings) of each speaker was divided into two non-overlapping parts. One half of the utterances of each speaker were utilized during the enrollment and the rest for the testing and vice versa.

Thus, overall 1900 runs (190 speakers x 10 recordings) were conducted for each test condition (type of codec or its bit rate) for each scenario. For each session, a recognition performance was evaluated by calculating a success rate in percentage. The obtained values were averaged and a mean value and variance were calculated.

As a front-end, speech analysis was performed frame-by-frame with 16 ms frame duration and 50% overlap, 12 Mel Frequency Cepstral Coefficients (MFCCs) (excl. 0-th coefficient) were extracted from each speech frame. The MFCCs are the most common speech features used in both speech and speaker recognition [22].

### 3.1 Experimental results

Figure 4 shows a success rate of the recognition process for the first two scenarios, namely fully matched and partially mismatched scenarios. It can be seen from the graphs that the recognition rate for the partially mismatched conditions is higher in the most of the cases compared to the fully matched conditions, except for G.711 codec, and EVS codec operating at 5.9 kbps. But the difference between the fully matched and partially mismatched case with G.711 and EVS codecs is rather small, less than 3%. The worst performance of the speaker identification system was achieved for AMR codec operating at 5.9 and 7.95 kbps. On the other hand, the best performance was obtained for G.711 codec closely followed by EVS codec operating at 13.2 kbps. A maximum difference between the success rate achieved for partially mismatched and fully matched scenario was attained for AMR codec (all bit rates). As can be also clearly seen from Fig. 4, the success rate grows with increasing bit rate and results are more statistically balanced (see Fig. 5 for statistical variances) for the AMR and EVS codec in the partially mismatched scenario. It is worth noting here that EVS codec has achieved quite good success rate in this experiment despite the fact that this codec represents lossy parametric codecs. This can be considered as a very promising result as this codec is going to be widely deployed in a voice communication over LTE (a successor of 3G mobile communication system). Therefore, a huge amount of the voice communication is going to be coded by this codec in the near future as mobile networks generate a dominant portion of voice communication nowadays. Table 1 summarizes the results for both experimental scenarios and all the test conditions. For comparison, in the case when both training and testing were carried out only on clean uncoded speech, recognition rate reached 96% as resulting from our previous experiments [23].

It is not a trivial task to select a codec for training offering a good performance over wide range of NB codecs currently deployed in telecommunication networks, as a performance of the recognition system is a codec-specific (see for instance results presented above). Moreover, an identification application mostly does not have access to a communication protocol and thus does not have information about the codec used for the voice transmission.

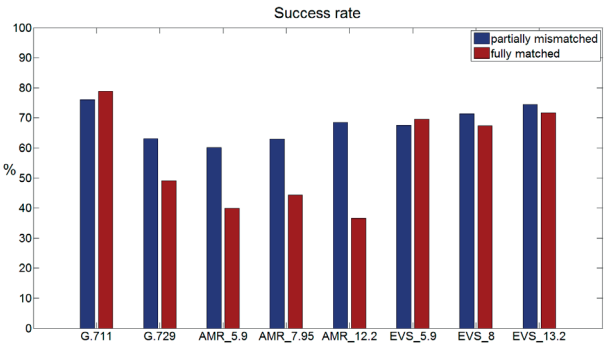


*Fig. 4 Average success rate for the first two scenarios*

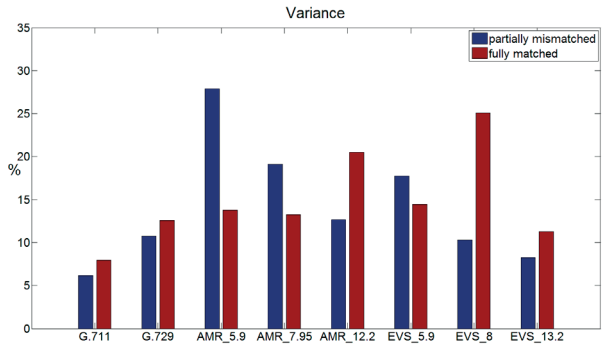


*Fig. 5 Variance of success rate for the first two scenarios*

In order to build the robust system covering all the prospective coding/degradation situations, a good performance in codec mismatch situations is also very important for ASR applications. Such codec mismatch situations are represented by the third scenario used in this study, defined above as the *fully mismatched* scenario, in which training and test materials come from the different codecs. As a fully factorial design would be very demanding in this case, we have decided to use only one bit rate of those offered by AMR and EVS codec for this part of the experiment. We have chosen a bit rate of 5.9 kbps because we believe that this bit rate covers most of the degradations introduced by the particular codecs. Due to clarity reasons, the results are only presented in a tabulated form, see Table 2. The best results were obtained for G.711 and EVS codec deployed for speaker specific model training. It is of a great surprise that the success rate obtained for G.711 and EVS is very similar. The corresponding difference is rather small, less than 1%. Note that EVS codec, as a lossy parametric codec, offers significantly better results than G.729 and AMR codecs. So, we can conclude that EVS and G.711 codecs offering the best speech quality from the codecs involved in this study (based on our expert listening), have achieved the highest success rate. This is in line with the finding presented in [1] involving different codecs.

Speaker recognition accuracy [%] for systems trained and tested with different codecs — Table 2

| enrollment/test | G.711 | G.729 | AMR_5.9 | EVS_5.9 |
|---|---|---|---|---|
| G.711 | - | 61.00 | 63.37 | 67.79 |
| G.729 | 43.26 | - | 44.16 | 39.58 |
| AMR_5.9 | 33.89 | 38.16 | - | 33.53 |
| EVS_5.9 | 67.16 | 64.21 | 61.63 | - |

## 4. Conclusions and future work

In this paper we analyze an impact of digital communication channel, particularly commonly used narrowband (NB) speech codecs in current telecommunication networks, on a performance of automatic speaker recognition in the context of codec mismatch between enrollment and test utterances. We have constructed the speaker identification system (Fig. 1) using the UBM-GMM model for the three different scenarios, namely fully matched, partially mismatched and fully mismatched. Surprising finding is that it is better to use the partially mismatched conditions (the system trained on clean uncoded data) than the fully matched conditions/scenario (the system trained and tested on the coded data using the same coding scheme). In the case of fully mismatched scenario (the system trained and tested on the coded data but using diverse coding schemes), the best recognition rate is achieved if G.711 or EVS codecs are deployed during the speaker enrollment (speaker specific GMM training). Note that both codecs offer the best speech quality from all the codecs deployed in this study. This finding fully correlates with the finding presented in [1] focusing on different speech codecs.

Since GMM training is sensitive to amount of the training data we plan to extend the experiments by using other speech databases (note TIMIT contents less than 20 seconds of speech per speaker), and also apply different types of speech features. In addition to that we would like to improve a performance of the proposed system. Moreover, we also plan to extend a codec set used in this experiment towards wideband speech codecs in order to reflect a current movement in a voice communication over telecommunication networks.

## References


[1] JANICKI, A., STAROSZCZYK, T.: *Speaker Recognition from Coded Speech Using Support Vector Machines,* TSD 2011, LNAI 6836, pp. 291-298, 2011.

[2] BHATTACHARJEE, U., SARMAH, K.: GMM-UBM Based Speaker Verification in Multilingual Environments, *IJCSI Intern. J. of Computer Science Issues*, vol. 9, No. 6, November 2012.

[3] ASBAI, N., AMROUCHE, A., DEBYECHE, M.: *Performances Evaluation of GMM-UBM and GMM-SVM for Speaker Recognition in Realistic World*, ICONIP 2011, Part II, LNCS 7063, pp. 284-291, 2011.

[4] PILLAY, S. G., ARIYAEEINIA, A., PAWLEWSKI, M., SIVAKUMARAN, P. **Speaker Verification under Mismatched Data Conditions,** *IET Signal Processing*, vol. 3, No. 4, 2009, pp. 236-246.

[5] FAKHR, W., ABDELSALAM, A., HAMDY, N.: *Enhancement of Mismatched Conditions in Speaker Recognition for Multimedia Applications,* ICASSP, vol. 1, pp. 377-80, 2004.

[6] QUATIERI, T. F., SINGER, E., DUNN, R. B., REYNOLDS, D. A., CAMPBELL, J. P.: Speaker and Language Recognition Using Speech Codec Parameters, *Eurospeech*, vol. 2, pp. 787-790, 1999.

[7] GALLARDO, L. F., WAGNER, M., MOLLER, S.: *I-vector Speaker Verification for Speech Degraded by Narrowband and Wideband Channels,* ITG-Fachbericht 252: Speech Communication, Erlangen, September 2014.

[8] 3GPP: EVS Codec Detailed Algorithmic Description, Third Generation Partnership Project, 3GPP TS 26.445, 2014.

[9] REYNOLDS, D. A., QUATIERI, T. F., DUNN, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing,* vol. 10, No. 1-3, 2000, pp. 19-41.

[10] BECKER, T., JESSEN, M., GRIGORAS, C.: *Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models*. Interspeech 2008, pp. 1505-1508.

[11] SORDO MARTINEZ, P. L., FAUVE, B., LARCHER, A., MASON, J. S.: *Speaker Verification Performance with Constrained Durations.* Intern. Workshop on Biometrics and Forensics (IWBF), 2014, pp. 1-6.

[12] TOGNERI, R., PULLELLA, D.: An Overview of Speaker Identification: Accuracy and Robustness Issues. *Circuits and Systems Magazine*, IEEE, 11(2), 2011, 23-61.

[13] REYNOLDS, D. A., ROSE, R.: *Robust Text-independent Speaker Identification Using Gaussian Mixture Speakers Models,* IEEE Trans. On Speech and Audio Processing 3, 1995, pp. 72-83.

[14] BISHOP, C.: *Pattern Recognition and Machine Learning, Springer Science+Business Media*, LLC : New York, 2006.

[15] LINDE, Y., BUZO, A., GRAY, R.: *An Algorithm for Vector Quantizer Design.* IEEE Transactions on Communications 28, 1980, pp. 84-95.

[16] KINNUNEM, T., LI, H. *An Overview of Text-Independent Speaker Recognition: From Features to Supervectors*, Speech Communication, 2009

[17] GAROFOLO, J., LAMEL, J. et al.: *DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. National Institute of Standards and Technology, 1990.

[18] RAJESHWARA, R. R., PRASAD, A., KEDARI RAO, CH.: Robust Features for Automatic Text-Independent Speaker Recognition Using Gaussian Mixture Model, *Intern. J. of Soft Computing and Engineering (IJSCE)*, vol. 1, No. 5, November 2011.

[19] ITU: *Pulse Code Modulation (PCM) of Voice Frequencies*, Intern. Telecommunication Union : Geneva, ITU-T Rec. G.711, 1988.

[20] ITU: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear prediction (CS-ACELP), Intern. Telecommunication Union : Geneva, ITU-T Rec. G.729, 2007.

[21] 3GPP: Mandatory Speech CODEC Speech Processing Functions; AMR speech Codec; General description, Third Generation Partnership Project, 3GPP TS 26.071, 2012.

[22] REYNOLDS, D. A.: *An Overview of Automatic Speaker Recognition Technology,* IEEE, 2002.

[23] POLACKY, J., GUOTH, I.: *Comparative Evaluation of GMM and GMM/UBM Speaker Identification Systems,* Proc. of intern. conference TRANSCOM 2015, University of Zilina, June 2015.