A STUDY OF ACOUSTIC FEATURES FOR EMOTIONAL SPEAKER RECOGNITION IN I-VECTOR REPRESENTATION

Lenka MACKOVÁ, Anton ČIŽMÁR, Jozef JUHÁR
Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, tel. +421 55 602 4220,
e-mail: lenka.mackova@tuke.sk, anton.cizmar@tuke.sk, jozef.juhar@tuke.sk

ABSTRACT

Recently recognition of emotions became very important in the field of speech and/or speaker recognition. This paper is dedicated to experimental investigation of best acoustic features obtained for purpose of gender-dependent speaker recognition from emotional speech. Four feature sets - LPC (Linear Prediction Coefficients), LPCC (Linear Prediction Cepstral Coefficients), MFCC (Mel-frequency Cepstral Coefficients) and PLP (Perceptual linear prediction) coefficients - were compared in an experimental setup of speaker recognition system, based on i-vector representation. For evaluation of the system emotional speech recordings from newly created Slovak emotional database and Mahalanobis distance metric as scoring method were used. The results of the experiment showed the MFCC representation as the best fitted for speaker verification from Slovak emotional speech with recognition rate higher than 80%.

Keywords: *emotions, i-vectors, total variability, speaker verification, Mahalanobis*

1. INTRODUCTION

Emotions play very important part in everyday human communication. Since human interaction is multimodal it covers not only speech but gestures, facial expression and "body language" as well. To this kind of communication emotions fill very important background information according which we – people – are able to identify the meaning of the explicit spoken message [1].

The same principle of recognition of the emotions plays essential role in inter-cognitive human-computer interaction (HCI) [2] [3]. One of many of the HCI applications using emotion recognition can be automatic tutoring, where a tutor – agent – may lead the study process according to the emotional expressions of the students. Another application of emotional recognition may be in systems which can alert a user to signs of emotion that call for attention. In forensic the emotion recognition can be very advisable supplement of polygraph or speaker verification and in medicine it can help to the early diagnosis of neurological disorders and diseases.

In this article the focus is on speaker recognition from emotional speech. Speaker recognition is relevant to applications where the access is controlled through the speaker`s voice (e.g. security control applications, telephone based access, etc.) and since speech in everyday life cannot be expressed only in neutral emotional state emotion recognition became very important even in this field of science. In the literature there were experiments which focused on this objective [4] [5]. To the speaker recognition purposes emotional databases in foreign languages were used and the results showed the contribution of emotional speech in process of speaker recognition.

To the disposition of this work several emotional databases in foreign languages were available, namely EMA (Electromagnetic Articulography) database [6], EMO DB (Berlin emotional database) [7], EESC (Estonian emotional speech corpus) [8] and BAUM-2 [9]. None of those corpuses satisfied the requirements of emotional speaker data volume for purposes of proper training of the speaker model. Therefore we decided to create our own native emotional dataset which is presented in this work.

In automatic speaker verification many different techniques may be used [10]. Methods such as the Gaussian Mixture Models based on Universal Background Model (GMM-UBM) [11], Eigenvoices [12] and Eigenchannels [13] belong to the class of mostly applied generative models. To the category of GMM-UBM models also pertains nowadays the most powerful speaker verification technique - Joint Factor Analysis (JFA) [14].

JFA is based on use of the fundamentals of high-dimensional GMM supervectors for the inter-speaker variability modelling and channel/session compensation. Results of the work in [15] showed that in channel space of JFA speaker information is obtained as well. Accordingly Dehak et al. [16] proposed a concept of so called i-vectors. The main idea of i-vectors is based on the use of only one GMM supervector subspace of low dimension. In this space speaker as well as channel variability information is represented thus this space is called total variability space.

To differentiate between speaker and session information in total variability space proper normalization technique has to be used. For suppression of session information mostly LDA (Linear Discriminant Analysis) [17], WCNN (Within-class Covariance Normalization) [18], NAP (Nuisance Attribute Projection) [19] techniques or its extended version known as Eigen Factor Radial (EFR) normalization is employed.

The organization of this article is as follows. In section 2 the recognition system is described, section 3 provides introduction to native Slovak emotional speech database – SUS. Section 4 is dedicated to the actual experiment and section 5 provides discussion of the results and conclusion.

2. SYSTEM DESCRIPTION

In Fig. 1 is depicted system used for emotional speaker recognition. In this system a state of the art technique of i-

vectors is used for representation of audio information in low-dimensional space. To extract i-vectors from emotional recordings Alize/LiaRal [20] toolkit for speaker recognition was used.

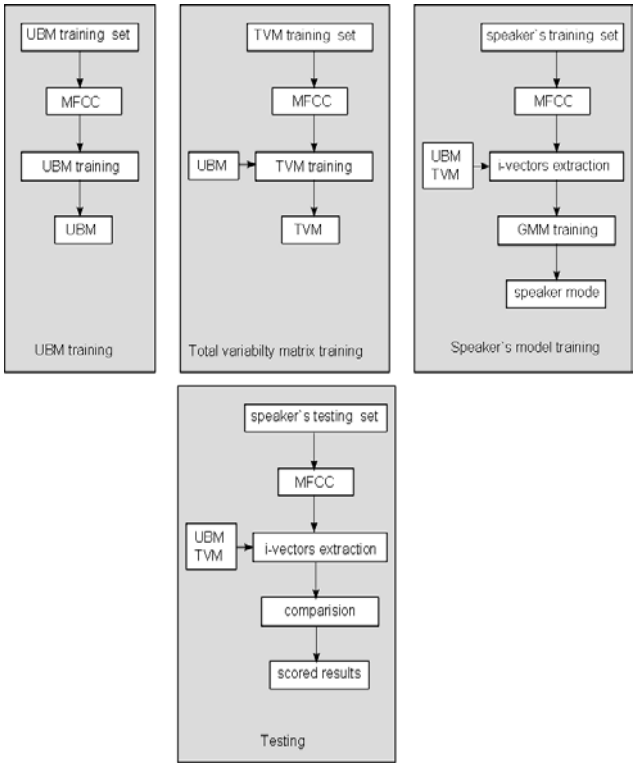

Fig. 1 Scheme of the recognition system

2.1. Universal background model training

Audio features extracted in font-end processing of training data waveforms are used in GMM/UBM training. UBM was trained by means of the EM (expectation-maximization) algorithm.

2.2. Total variability matrix training

In front-end processing of total variability data training set feature vectors are extracted. To obtain UBM supervector for extracted feature vectors the adaptation of UBM using MAP (Maximum a posteriori) algorithm is performed. Then the i-vector technique is applied to the resulted UBM supervectors to obtain vectors without effect of speaker and session variability. With concept of total variability space in an i-vector method an UBM supervector **W** is defined as

$$\mathbf{W} = \mathbf{w} + \mathbf{T}\omega. \tag{1}$$

In equation (1) **T** is a rectangular total variability matrix and **w** stands for speaker and session-dependent supervector. Vector ω is random vector with standard normal distribution N(0,I) which components are called total factors or *i-vectors*.

The training process of **T** matrix is similar to the eigenvoice matrix training. The difference is in considering the set of speaker`s recordings to belong to different subject when training total variability matrix.

To compensate the session variability in total variability space EFR normalization is used. EFR uses the concept of NAP [21] where the channel variability is estimated as partial rank of the within-class covariance matrix. EFR itself then suppress the nuisance dimension of computed i-vectors and continues in their normalization by rotating them to the first principal axe in orthogonal subspace of the speaker where the i-vectors are projected. The rotation of i-vectors in session space is depicted in Fig. 2.

In EFR the reduction of channel variability is defined as

$$\mathbf{w} = \frac{\mathbf{W}_c^{-\frac{1}{2}}(\mathbf{w} - \overline{\mathbf{w}})}{\sqrt{(\mathbf{w} - \overline{\mathbf{w}})\mathbf{V}^{-1}(\mathbf{w} - \overline{\mathbf{w}})}}, \tag{2}$$

where $\overline{\mathbf{w}}$ is the mean of i-vectors, $\mathbf{V}$ is eigenvoice matrix and $\mathbf{W}_c$ is the covariance matrix defined by equation

$$\mathbf{W}_c = \sum_{s=1}^{S} \frac{n_s}{n} \mathbf{W}_s = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \overline{\mathbf{w}_s})(\mathbf{w}_i^s - \overline{\mathbf{w}_s})^t. \tag{3}$$

In equation (3) $\mathbf{W}_s$ is the speaker *s* covariance matrix, *n* is the number of all of the utterances, n_s is the number of speaker *s* utterances with $\overline{\mathbf{w}_s}$ as their mean.

2.3. Speaker model training

To train the speaker model feature extraction from training audio files is performed. Then the total variability matrix **T** and UBM are employed into projection of extracted feature vectors into space of i-vectors. Finally GMM training of speaker model is provided.

2.4. Testing

Based on our previous experiment [22], where Mahalanobis distance metric showed better result than CSS, we decided to use this only metric to evaluate the recognition system.

Mahalanobis distance metric compares training set of entities to the mean of known class distribution. The goal of this method is to allocate an observed entity to the best fitted class. In this work an entity – speaker`s i-vector – is assigned to the class of the speaker *s* as in equation

$$(\mathbf{w} - \overline{\mathbf{w}_s})^t \mathbf{W}_s^{-1} (\mathbf{w} - \overline{\mathbf{w}_s}) = \left\| \mathbf{w} - \overline{\mathbf{w}_s} \right\|^2_{\mathbf{W}_s^{-1}}, \tag{4}$$

where $\mathbf{W}_s$ is the covariance matrix of speaker *s* as in (3) and $\overline{\mathbf{w}_s}$ is the mean of class.
The final Mahalanobis scoring is defined as

$$score(\mathbf{w}_1, \mathbf{w}_2) = -\left\| \mathbf{w} - \overline{\mathbf{w}_s} \right\|^2_{\mathbf{W}_s^{-1}}, \tag{5}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are two i-vectors scored by the log-probability that $\mathbf{w}_1$ and $\mathbf{w}_2$ belong to the same class in accordance to the covariance matrix $\mathbf{W}_s$.

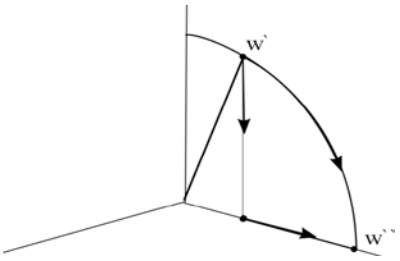

Fig. 2 Eigen Factor Radial normalization

3. EMOTIONAL DATABASE

For purposes of speaker recognition in this experiment we created emotional database in native Slovak language. Emotional audio recordings of different subjects were captured from free FTA DVB-T transmission using PCI digital capture card.

This database can be categorized as induced database [23] since captured sessions of SUS consist fabricated law suits situations in which non-professional actors take parts. Participants are supposed to act according to the storyline in each session, but a slight anticipation of controlled expression of emotional states may be expected.

Each session is oriented to solutions of juridical cases therefore emotional range of recorded utterances covers mostly emotions of neutral state and curiosity as well as negative emotions such as anger, aggressiveness, sadness and disgust. Positive emotions are rare to find in those recordings.

SUS audio sessions were down-sampled to 16 kHz from the original 48 kHz 128kbit mpa2 audio stream, encoded using LIN16 PCM encoding, mono and saved all in WAV format.

All of the emotional utterances were manually labeled. The emotional evaluation was provided on the whole sentences so that the explicit meaning of the utterance was captured in recording. In case where more than one emotion occurred in an utterance the division of such an utterance to the shorter segments following the rule of information relevance was carried out. The whole sentences or segments were then evaluated from emotional point of view and then labeled with capital letters representing the specific emotions (e.g. CU for curiosity, N for neutral, etc.). Using Transcriber software [24], labeled sessions were segmented and then, using proprietary script, were cut into separate emotional utterances of individual speakers with duration from 5 to 6 seconds.

The SUS database consists nowadays from approximately 2000 utterances of 7 speakers (3 male, 4 female) in emotions of neutral, curiosity, anger, sadness and so.

4. EXPERIMENTAL SETUP

In this paper the focus was on performance of speaker verification system when employing recordings of emotional database in Slovak language. For this reason in front-end processing MFCC, LPC, LPCC and PLP features were extracted from emotional utterances of three male subject of the SUS corpus.

- LPC - Linear prediction analysis (LPA) is method in which speech signal is approximated as a linear combination of its p previous samples. The coefficients estimated by this method – Linear Predictive Coefficients (LPC) – describe the formants in speech signal.
- LPCC - Linear system which models the human vocal tract can be described with use of cepstral coefficients as well. Linear Predictive Cepstral Coefficients (LPCC) are obtained by estimation of Power Spectral Density (PSD) of the signal. The advantage of LPCC resides in their smaller correlation in comparison to LPC.
- MFCC - Mel -Frequency Cepstral Coefficients represent the real cepstrum of a windowed short-time signal which is derived from the Fast Fourier Transformation (FFT) of that signal. Since the human auditory system processes a speech signal nonlinearly the MFCC analysis is used to represent such a signal with respect to nonlinear fashion of the frequency. To perform that nonlinear mel-scale bank is used to convert from normal frequency f to mel frequency f_{mel} is given by

$$f_{mel} = 2596.\log_{10}(1+f\ 700). \quad (6)$$

  Additionally the MFCC`s are robust and reliable to variations according to speakers and recording conditions. When calculating the MFCC`s all audio information except those parameters similar to ones that are used by humans for hearing speech, are deemphasized.
- PLP - Perceptual linear prediction (PLP) attempts to approximate to the perception of sound by human auditory organs. By discarding irrelevant information within a speech signal PLP improves the recognition performance. The process of PLP computation is identical to LPC with difference in spectral characteristics which have been transformed to match characteristics of human auditory system.

In the process of feature extraction firstly 19 coefficients of mention acoustic feature sets were computed. For augmentation of spectral parameters obtained in process of LP or mel-filterbank analysis an energy coefficient was appended. This energy term was computed as log of the signal energy. To enhance performance of a speech recognition system to the basic coefficients additional time derivates were appended as well. The first order regression coefficients, referred to as delta coefficients, the second order regression coefficients (acceleration coefficients) and the third order regression coefficients. The dimension of such computed vector was 80 segments per frame.

In the second step the number of extracted basic coefficients was increased to the count of 22. The final size of feature vector obtained in the process of extraction was 92 dimensions. All of the audio features were extracted using 25 ms Hamming window with shift of 10 ms.

In the process follows energetic coefficients of silent frames were normalized with respect to zero mean and variance and then, according to the speech energy, evaluated. Comparing to the specific threshold, frames with higher energy were used in the next i-vector extraction processing. Customized features were then mapped to the fixed-length vector in process of i-vector extraction. After running experiment with different values of dimensional parameter the decision was to set the dimension of extracted i-vectors to the number of 10.

Gender-dependent UBM was trained using 247 recordings of background noise and SUS male speakers not included in training and testing phase. In this experiment 32, 64, 128 and 256 Gaussians were used in UBM training. The comparison of the best number of Gaussians used in UBM training was made in evaluation phase. The total variability matrix was trained with 250 emotional utterances of the total variability training dataset. Several experiments with different number of T iteration were carried out. The best results were obtained with number of 10 iterations.

To created speaker model training and testing sets emotional utterances of three male speakers (spk1, spk2, spk3) from the SUS database in emotions of neutral and curiosity were used. Those emotions were the most common to extract from SUS sessions. In process of training the speaker model two different approaches were chosen to be applied as shown in Fig. 3.

In the first approach to train the speaker model emotions of neutral and curiosity were used. It resulted in generation of one mixed model per speaker trained with all available emotional recordings of the speaker.

The results of the second approach were two different speaker`s models, both trained with only one emotion from emotional training set of required subject.

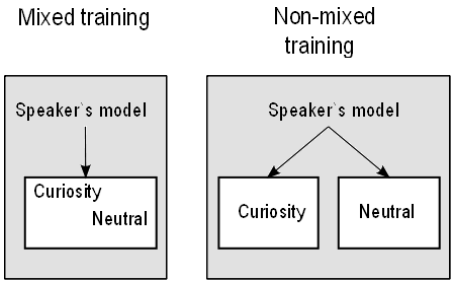

Fig. 3 Mixed and non-mixed model approach

The best resulting score in emotional speaker recognition obtained for different extracted features is shown in form of confusion matrixes in Table 1, Table 2, Table 3 and Table 4.

In a confusion matrix diagonal elements represent correct classification while other elements in each row express misclassification. For example in the first row corresponding to speaker 1 (spk1) of Table 80% of emotional utterances were recognized correctly, 8% of them were misled as utterances of speaker 2 (spk2) and 12% of emotional utterances were presented as utterances of speaker 3 (spk3).

Table 1 Speaker recognition rate (%) for mixed model with 128GMM/UBM training

	22 MFCC			19 LPC			19 PLP		
	spk1	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**
spk1	**80**	8	12	**73**	15	12	**69**	11	20
spk2	0	**82**	18	2	**73**	25	6	**70**	24
spk3	14	7	**79**	8	22	**70**	25	7	**68**

Table 2 Speaker recognition rate (%) for non-mixed model with 128GMM/UBM training (speaker model trained with neutral)

	22 MFCC			19 LPC			19 PLP		
	spk1	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**
spk1	**83**	2	15	**73**	15	12	**69**	11	20
spk2	5	**85**	10	4	**71**	25	6	**70**	24
spk3	12	8	**80**	8	22	**70**	25	7	**68**

Table 3 Speaker recognition rate (%) for non-mixed model with 128GMM/UBM training (speaker model trained with curiosity)

	22 MFCC			19 LPC			19 PLP		
	spk1	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**	**spk1**	**spk2**	**spk3**
spk1	**82**	1	17	**70**	17	13	**70**	20	10
spk2	0	**82**	18	11	**70**	19	9	**70**	21
spk3	10	9	**81**	3	27	**70**	22	9	**69**

Table 4 Speaker recognition rate (%) for 22 LPCC with 64 GMM/UBM training

	Mixed model		
	spk1	**spk2**	**spk3**
spk1	67	11	22
spk2	6	67	27
spk3	32	3	65

	Non-mixed model (neutral)		
	spk1	**spk2**	**spk3**
spk1	67	12	21
spk2	5	67	28
spk3	30	5	65

	Non-mixed model (curiosity)		
	spk1	**spk2**	**spk3**
spk1	68	15	17
spk2	9	66	25
spk3	27	8	65

5. DISCUSSION OF RESULTS AND CONCLUSION

According the results from tables (Table 1, Table 2, Table 3 and Table 4) the best results in both approaches were mostly obtained with 128 GMM/UBM training. The difference was in case of extraction LPCC when the best recognition rate was with 64 GMM/UBM training and 22 coefficients per frame extracted independently on mixed or non-mixed model approach used.

When computing LPC and PLP the best recognition rate (73% and 70% respectively) was obtained with number of 19 coefficients extracted.

The best recognition rate whatsoever was obtained extracting 22 MFCC. The second approach (non-mixed model) showed better results compared to mixed model training. Recognition rate in case of non-mixed training model resulted in 85% while when using approach of mixed model training the best percentual value of recognition rate was 82%.

According investigation of this paper 22 MFCCs provide the best recognition rate in speaker verification on the SUS emotional database. The superior performance of MFCC may be related to the fact that MFCCs are the best features to represent perceptual aspect of short-term speech spectrum.

Since in non-mixed training approach speaker model was trained by utterances of only one emotion the emotional characteristics differ less than in mixed model training and this, we suppose, is the reason of better recognition rate in this case.

In the future we would like to focus on testing the gender-dependent speaker verification system on female emotional dataset of the SUS corpus. We also plan to continue in enlargement of the SUS corpus.

ACKNOWLEDGMENTS

The research presented in this paper was supported partially (50%) by Competence Center for Innovation Knowledge Technology of production systems in industry and services (ITMS project code 26220220155) and partially (50%) by VEGA 1/0075/15.

REFERENCES

[1] VERVERIDIS, D. – KOTROPOULOS, C.: Emotional speech recognition: Resources, features, and methods, *Speech communication* 48.9, pp.1162–1181, 2006.

[2] GALANIS, D. et al.: Classification of emotional speech units in call centre interactions, *4th IEEE Conference on Cognitive Infocommunications*, pp.403-406, 2013.

[3] BARANYI, P. – CSAPO, A.: Definition and synergies of cognitive infocommunications, *Acta Polytechnica Hungarica* 9.1, pp. 67-83, 2012.

[4] KOOLAGUDI, S. G. – SHARMA, K. – RAO, K. S.: Speaker Recognition in Emotional Environment, *Eco-friendly Computing and Communication Systems*, Springer Berlin Heidelberg, pp. 117-124, 2012.

[5] LI, D. – YANG, Y.: Emotional speech clustering based robust speaker recognition system, *Image and Signal Processing*, 2009. CISP'09. 2nd International Congress on. IEEE, pp. 1-5, 2009.

[6] LEE, S. – YILDIRIM, S. – KAZAMZADEH, A. – NARAYANAN, S.: An articulatory study of emotional speech production," *In Interspeech*, pp. 497-500, 2005.

[7] BURKHARDT, F. – PAESCHE, A. – ROLFES, M. – SEDLMEIER, W. – WEISS, B.: A database of German emotional speech, *In Interspeech*, Vol. 5, pp. 1517-1520, 2005.

[8] ALTROV, R. – PAJUPUU, H.: Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers, *In Human Language Technologies – The Baltic Perspective Proceedings of the Fourth*

International Conference Baltic HLT, pp. 25 – 32, 2010.

[9] ERDEM, C. E. – CIDGEM, T. – ZAFER, A.: BAUM-2: a multilingual audio-visual affective face database, *Multimedia Tools and Applications*, Online first, pp. 1-31, 2014.

[10] HRIC, M. – CHMULÍK, M. – JARINA, R.: Comparison of selected classification methods in automatic speaker identification, *Komunikácie (Communications)*, 13 (4), pp. 20-24, 2011.

[11] REYNOLDS, A. D.: A gaussian mixture modeling approach to text-independent speaker identification, Ph.D. thesis, Georgia Institute of Technology, 1992.

[12] KUHN, R. et al.: Eigenvoices for speaker adaptation, *ICSLP*. Vol. 98, pp. 1774-1777, 1998.

[13] MATROUF, M. – SCHEFFER, N. – BONASTRE, J. F.: A straightforward and efficient implementation of the factor analysis model for speaker verification, *In Interspeech*, pp. 1242-1245, 2007.

[14] KENNY, P. – BOULIANNE, G. – OUELLET, P. – DUMOUCHEL, P.: Joint factor analysis versus eigenchannels in speaker recognition, *Audio, Speech, and Language Processing*, IEEE Transactions on, pp. 1435-1447, 2007.

[15] DEHAK, N.: Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification″, Ph.D. thesis, École de Technologie Supérieure, Montreal, QC, Canada, 2009.

[16] DEHAK, N. – DEHAK, R. – KENNY, P. – BRUMMER, N. – OUELLET, P. – DUMOUCHEL, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification″, *In Interspeech*, Vol. 9, pp. 1559-1562, 2009.

[17] VISZLAY, P. – JUHAR, J. – PLEVA, M.: Modified estimation of between-class covariance matrix in linear discriminant analysis of speech, *Systems, Signals and Image Processing*, 20th International Conference on. IEEE, pp.167-170, 2013.

[18] HATCH, A. – KAJAREKAR, S. – STOLCKE, A.: Within-class covariance normalization for Svm-based speaker recognition, *Interspeech 2006 and 9th International conference on Spoken Language Processing* - ICSLP, vol. 3, pp. 1471 – 1474, 2006.

[19] MACHLICA, L. – ZAJÍC, Z.: Factor Analysis and Nuisance Attribute Projection Revisited, *In Interspeech*, pp. 1570-1573, 2012.

[20] LARCHER, A. et al.: ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition, *In Interspeech*, pp. 1-5, 2013.

[21] CAMPBELL, W. M. et al.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, *ICASSP*, Vol. 1, pp. I97-I100, 2006.

[22] MACKOVÁ, L. – ČIŽMÁR, A.: Emotional Speaker Verification Based on I-vectors, *5th IEEE International Conference on Cognitive Infocommunications*, pp. 533-536, 2014.

[23] KOOLAGUDI, S. G. – RAO, K. S.: Emotion recognition from speech: a review, *International Journal of Speech Technology*, Springer, 15, pp. 99-117, 2012.

[24] BARRAS, C. – GEOFFROIS, E. – WU, Z. – LIBERMAN, M.: Transcriber: Development and use of a tool for assisting speech corpora production, *Speech Communication*, 33 (1-2), pp. 5-22, 2001.

Received April 8, 2015 , accepted ,May 4, 2015

BIOGRAPHIES

Lenka Macková was born in Košice, Slovakia in 1982. In 2007 she graduated (MSc) with distinction at the department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics at Technical University in Košice. She is currently PhD student at the Department of electronics and multimedia communications at the Technical university of Košice. Her research interests include emotion recognition from speech signal.

Anton Čižmár was born in Michalovce, Slovakia in 1956. He graduated from the Slovak technical university in Bratislava in 1980, at the Department of telecommunications. He received his Ph.D. degree in Radioelectronics from the Technical university of Kosice in 1986, where he works as a full professor at the Department of electronics and multimedia communications and a rector of the Technical university of Košice. He is author and co-author of more than 170 scientific papers. His scientific research areas are broadband information and telecommunication technologies, multimedia systems, telecommunication networks and services, NGN mobile communication systems and localization algorithms.

Jozef Juhár was born in Poproč, Slovakia in 1956. He graduated from the Technical university of Košice in 1980. He received Ph.D. degree in Radioelectronics from technical university of Košice in 1991, where he works as a full professor at the Department of electronics and multimedia communications. He is author and co-author of more than 200 scientific papers. His research interest includes digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.