DUAL SHOTS DETECTION

Eva VOZARIKOVA[1], *Jozef JUHAR*[1], *Anton CIZMAR*[1]

[1]Department of Electronics and Multimedia Communications, Technical University of Kosice, Park Komenskeho 13, 042 00 Kosice, Slovak Republic

eva.vozarikova@tuke.sk, jozef.juhar@tuke.sk, anton.cizmar@tuke.sk

Abstract. *The identification of a special kind of acoustic events such as dual gunshots and single gunshots in the traffic background is described in this work. The recognition of dangerous sounds may help to prevent the abnormal or criminal activities that happened near to the public transport stations. Therefore in this paper the methodology of dual shots detection in a noisy background was developed and evaluated. For this purpose, we investigated various feature extraction methods and combinations of different feature sets. These approaches were evaluated by the widely used classification technique based on the Hidden Markov Models.*

Keywords

Detection, dual shots, feature supervector.

1. Introduction

In last few years a huge interest was concentrated to the detection of selected sound, where one category of sounds can include potential dangerous ones, for example gunshots, sounds of a car crash accidents, calling for help, etc. The detection of the mentioned sound category is relevant for surveillance or security systems. They are created by using the knowledge of different scientific areas, e.g. signal processing, pattern recognition, machine learning, etc. As it was indicated, surveillance or security systems can help to protect lives and properties and in some cases these systems can be very helpful for police because they operate according to the given instructions and they are not affected by any irrational behavior (bystander effect), [1], [2].

Principally, each surveillance system [3] can be divided into the two blocks, i.e. a feature extraction block and a classification block. In the feature extraction block, the input audio signal is transformed into the feature vectors according to the chosen feature extraction algorithm. Many approaches can be used for representing input signal [13], [14], e.g. speech based approach such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction coefficients (PLP), [7], etc., which were primary developed for speech or speaker recognition tasks. MPEG-7 standard [4], [5], [16], or other extraction algorithm such as statistical moments (skewness and kurtosis), Zero Crossing Rate (ZCR), etc. can be also used. In the block of classifier the input patterns are classified [9] into predefined classes according to the knowledge obtained in the training process of a classifier. There are several methods that can perform it, e.g. Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), Bayesian Networks (BN), Support Vector Machines (SVM) and Neural Networks (NN), etc.

The research described in this paper is focused on the dual shots detection in the noise traffic background. For this task, feature extractions in the time, spectral and cepstral domain were performed. The cepstral domain is represented with Mel-Frequency Cepstral coefficients (MFCC), spectral domain features represent MEL−SPECtral coefficients (MELSPEC) and the time domain is represented with the statistical moments (skewness, kurtosis) and Zero Crossing Rate (ZCR). We have supposed that the information from MFCC and MELSPEC (with their first and second time derivations of coefficients) and also with the time expressing features (skewness, kurtosis and ZCR) could distinguish the difference between single shots and dual shots. This approach is applied and evaluated by popular statistical classification method like Hidden Markov Models [15]. Dual shot detection is a partial task in the complex surveillance system [12].

The rest of the paper has the following structure: section 2 presents used feature extraction methods. Section 3 gives information about the sound database. Section 4 describes the proposed system and section 5 presents and summarizes obtained results.

2. Feature Extraction

The effective feature extraction is a very important phase of the overall process, because the recognition

performance directly depends on the quality of extracted feature vectors. In this work, feature extraction algorithms such as MFCC, MELSPEC [7], [8], skewness, kurtosis [11] and ZCR [5] were used.

2.1. Mel-Frequency Cepstral Coefficients

Ear's perception of the frequency components in the audio signal does not follow the linear scale, but rather the Mel-frequency scale [7], [8], which should be understood as a linear frequency spacing below 1 kHz and logarithmic spacing above 1 kHz. So filters spaced linearly at a low frequency and logarithmic at high frequencies can be used to capture the phonetically important characteristics. The relation between the Mel-frequency and the frequency is given by the formula:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \quad (1)$$

where f is frequency in Hertz [10]. MFCC coefficients are computed according to the Fig. 1. Normally, signal is filtered using preemphasis filter then the 25 ms Hamming window method was applied on the frames. Then, they are transformed to the frequency domain via Discrete Fast Fourier Transform (FFT), and then the magnitude spectrum is passed through a bank of triangular shaped filters. The energy output from each filter is then log-compressed and transformed to the cepstral domain via the Discrete Cosine Transform (DCT).

2.2. MELSPEC Coefficients

MELSPEC coefficients represent linear Mel-filter bank channel outputs. They are computed during extraction of MFCC coefficients. This process is illustrated in the Fig. 1, [10].

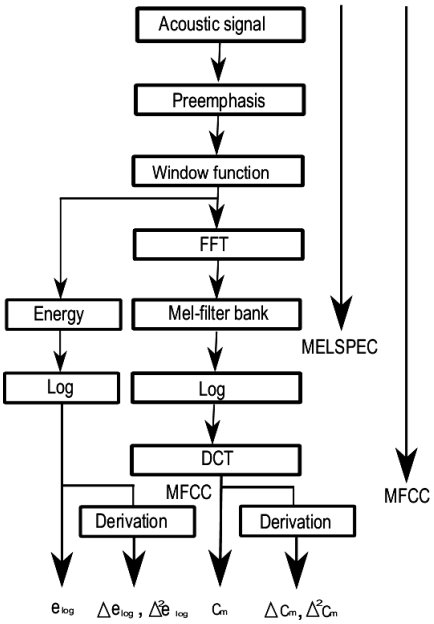

Fig. 1: Block scheme of MFCC and MELSPEC feature extraction.

2.3. Skewness and Kurtosis

Skewness and kurtosis [11] are statistical measures that describe the shape of data set. Skewness is a measure of symmetry, or more precisely, the lack of symmetry.

A distribution, or data set, is symmetric, if it looks the same to the left and right of the centre point. It is described by the formula:

$$Skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)\sigma^3}, \quad (2)$$

where $Y_1, Y_2, ..., Y_N$ are sample of data set, $\bar{Y}$ is the mean, σ is the standard deviation, and N is the number of data points. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Kurtosis is a measure of the "peakedness" of the probability distribution and it is described by the formula:

$$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)\sigma^4}. \quad (3)$$

The kurtosis for a standard normal distribution is equal to three.

2.4. Zero Crossing Rate

Zero Crossing Rate (ZCR), [5], is a very useful audio feature. It is defined as the number of times that the audio waveform crosses the zero axis. It is defined by the formula:

$$ZCR = \frac{1}{2}\left(\sum_{n=1}^{N-1}\left|sign(s(n)) - sign(s(n-1))\right|\right)\frac{Fs}{N}, \quad (4)$$

where N is total samples of the signal $s(n)$, Fs is a sampling frequency and $sign(x)$ function is defined follows:

$$sign(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases}. \quad (5)$$

3. Sound Database

Extended acoustic event database [6] used in the training phase contained 387 single gun shots, 123 dual shots and approximately 53 minutes of the traffic background. In the testing process, 38 single gun shots and 20 dual shots were recognized. The recordings (48 kHz, 16 bits per sample) were cut and manually labeled using Transcriber [17]. Some dual shots were recorded and the rest (approximately 90 %) were mixed using Audacity software [18]. In the Fig. 2 is the example of the one testing recording, which contained several dual shots and

single shots.

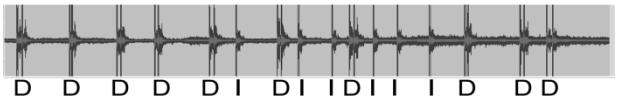

Fig. 2: The example of the testing recording (D − dual shot, I – single shot) with the traffic background.

4. System Description

As was mentioned before, each detection system consists of two different function blocks i.e. feature extraction block and classification block. The input acoustic signal was transformed in to the feature vectors. These vectors were used for training acoustic models for single shot, dual shots and background.

Our system used Hidden Markov Model (HMM) based learning technique. For classes of single shots, dual shots and background, HMMs from one to four states and from 1 to 1024 Probability Density Functions (PDFs) were trained and then evaluated in offline tests. In our experiments, different features sets with different dimensions were created, see the Tab. 1.

Tab.1: The list of feature sets and dimensions of feature vectors.

SKEWNESS, KURTOSIS, ZCR	3
MFCC_E	13
MELSPEC_E	13
MFCC_E + SKEWNESS, KURTOSIS, ZCR	16
MELSPEC_E + SKEWNESS, KURTOSIS, ZCR	16
MFCC_E_D + SKEWNESS, KURTOSIS, ZCR	29
MELSPEC_E_D + SKEWNESS, KURTOSIS, ZCR	29
MFCC_E_D_A	39
MELSPEC_E_D_A	39
MFCC_E_D_A + SKEWNESS, KURTOSIS, ZCR	42
MELSPEC_E_D_A + SKEWNESS, KURTOSIS, ZCR	42

The input sound was represented by the features from the cepstral domain (MFCC coefficients), the spectral domain (MELSPEC coefficients) and with common used time statistical parameters such as skewness, kurtosis and ZCR. Mentioned kinds of features could be able to distinguish single shots, dual shots and background. We also extracted first (_D) and second time derivations (_A), [8], which represent time expressing features.

MFCC and MELSPEC coefficients were extracted by HCopy tool [8]. Various settings of MFCC and MELSPEC were performed (e.g. with energy coefficient (_E), energy and delta (_E_D), energy, delta and acceleration coefficients (_E_D_A). Extractions of MFCC and MELSPEC coefficients were done without the pre-emphasis filtration, then the 25 ms Hamming window was applied. The signal was filtered by the Mel-filter bank with 29 triangular filters and finally 12 coefficients were computed per one signal frame.

Extractions of ZCR, skewness and kurtosis were done in Matlab environment. Speech based feature vectors were jointed with mentioned parameters (moments and ZCR) to the one supervector with higher dimension. It was done in Matlab too.

5. Results of Experiments

The detection and classification of events and background sounds were realized from one to four states HMM classifiers. In the testing phase, different feature extraction approaches were evaluated and discussion of obtained results is presented in this section.

First experiments were focused on the evaluation of the basic feature sets (skewness, kurtosis and ZCR with 3 dimensions (dim); MFCC_E and MELSPEC_E feature vectors had dim 13). These results are described in the following lines.

In our internal experiments, HMM models trained with statistical moments and ZCR achieved the maximum 71,54 % Accuracy – ACC [8] for 1 state HMM with 1PDF (1 HMM – 1 PDF). Then followed 1 HMM – 2 PDFs with ACC equal to 56,16 %. Accuracy values decreased to the value 16,92 % in case of 1 HMM – 1024 PDFs. The rest of the 2, 3 and 4 states HMMs achieved worse results than 1 HMMs – 1 PDF.

The efficiency of MFCC_E and MELSPEC_E was evaluated in offline tests and results of experiments are depicted in the Fig. 3. Generally for MFCC_E feature set, the 2HMM reached recognition results (about 60 % ACC), but more suitable results (e.g. 1 HMM – 8, 16, 32 PDFs) belonged to the MELSPEC_E parameterization, where the best recognition result 74,62 % was occurred for 1 HMM – 8 PDFs. These results are presented in the Fig. 3.

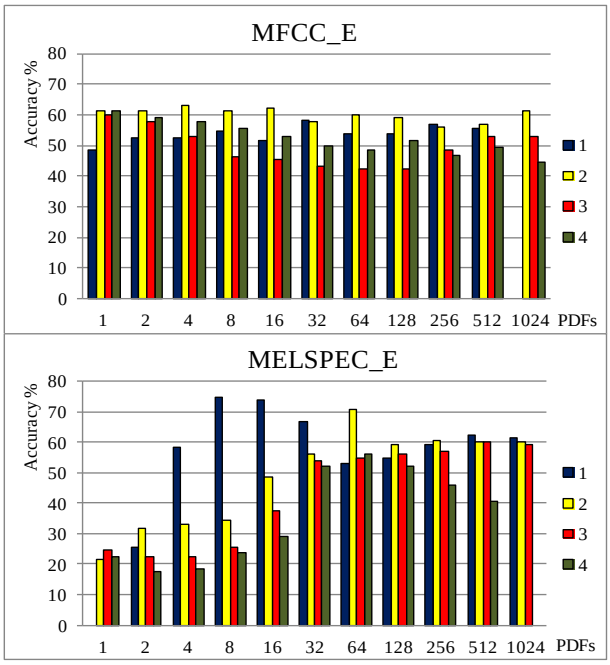

Fig. 3: Recognition results for MFCC_E and MELSPEC_E (the number of HMM states is illustrated by colors).

We also investigated the influence of ZCR, skewness, kurtosis, which were added to the MFCC_E and MELSPEC_E feature vectors, see the Fig. 4. This step brought better recognition performances in both cases. Positive impacts were observed mainly for MFCC_E + skewness, kurtosis, ZCR models up to 32 PDFs and also for 1HMM with more than 16 PDFs in the case of MELSPEC_E + skewness, kurtosis, ZCR.

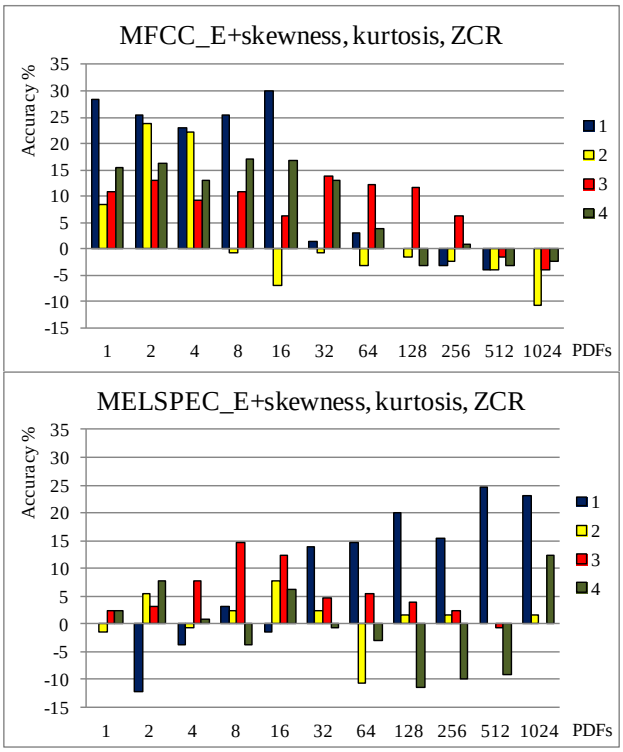

Fig. 4: Differences between baseline MFCC_E and extended MFCC_E + skewness, kurtosis, ZCR and also between second baseline approach MELSPEC_E and its extension MELSPEC_E + skewness, kurtosis, ZCR.

Generally, the recognition performance was positively influenced by delta and acceleration coefficients. Baseline models MFCC_EDA and MELSPEC_EDA with dimension 39 were evaluated and results are depicted in the Fig. 5. Then various combinations of features were also performed, evaluated and differences against baseline methods (MFCC_EDA and MELSPEC_EDA) were depicted in the Fig. 6 and Fig. 7.

The results of MFCC_EDA + skewness, kurtosis, ZCR with dim 42 and MFCC_ED + skewness, kurtosis, ZCR with dim 29 are presented in the Fig. 6. This figure expresses the differences between baseline approach and a new approach. The significantly improvements were occurred mainly in the case of 1HMM – 16, 32, 64 PDFs with MFCC_EDA + skewness, kurtosis, ZCR features set. Also the higher recognition results were occurred in the case of MFCC_ED + skewness, kurtosis, ZCR up to 32 PDFs. This approach did not bring the improvement for models with the higher number of PDFs (more than 64), Fig. 6. The mentioned effect was observed in MFCC_EDA + skewness, kurtosis, ZCR features set too.

The same experiments were performed also with the MELSPEC feature set. MELSPEC_EDA + skewness, kurtosis, ZCR with dim 42 and MELSPEC_ED + skewness, kurtosis, ZCR with dim 29 were evaluated and obtained results are depicted in the Fig. 7.

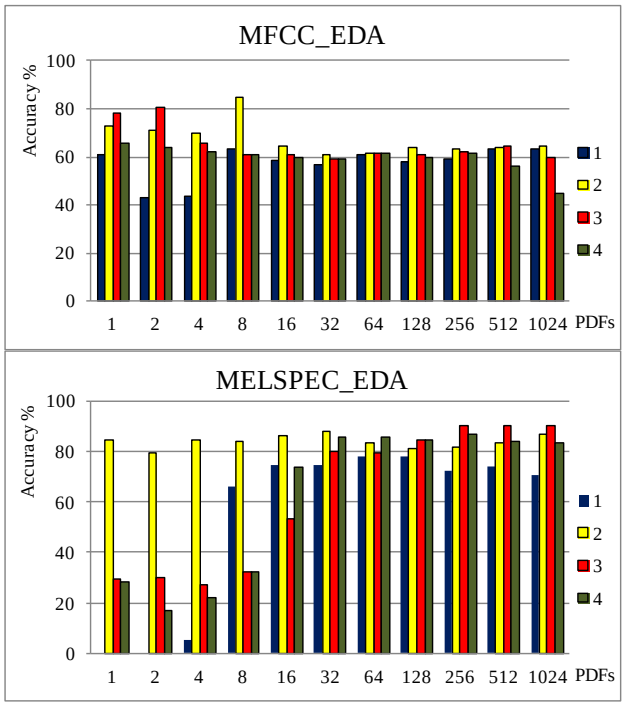

Fig. 5: Recognition results for baseline models MFCC_EDA and MELSPEC_EDA (the number of HMM states is illustrated by colors).

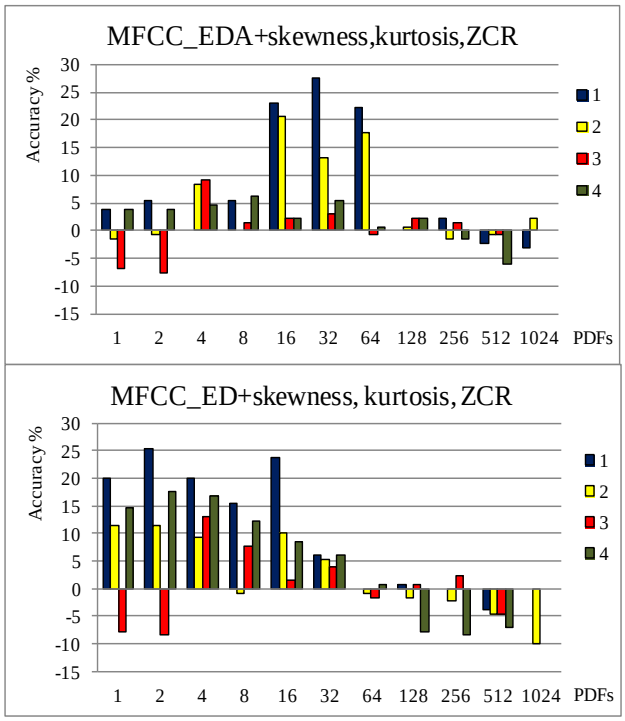

Fig. 6: Differences between baseline method (MFCC_EDA) and extended methods (MFCC_EDA + skewness, kurtosis, ZCR and MFCC_ED + skewness, kurtosis, ZCR).

The presence of skewness, kurtosis and ZCR with

a combination of MELSPEC_ED and MELSPEC_EDA did not bring an unambiguous improvement of recognition results, see the Fig. 7. In many cases, they caused worst results, especially for MELSPEC_ED + skewness, kurtosis and ZCR set, as you can see in the Fig. 7. Results of this feature set appointed to the lower recognition performance, which was occurred with 4HMM especially for 16, 128, 256, 1024 PDFs. Some positive impact of MELSPEC_ED + skewness, kurtosis and ZCR feature set was occurred for 1HMM from 8 to 1024 PDFs.

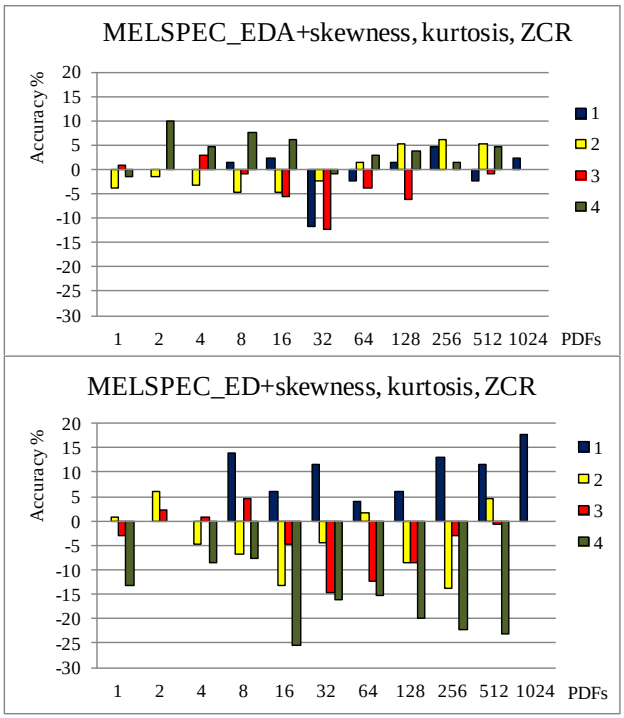

Fig. 7: Differences between baseline method (MELSPEC_EDA) and extended methods (MELSPEC_EDA + skewness, kurtosis, ZCR and MELSPEC_ED + skewness, kurtosis, ZCR).

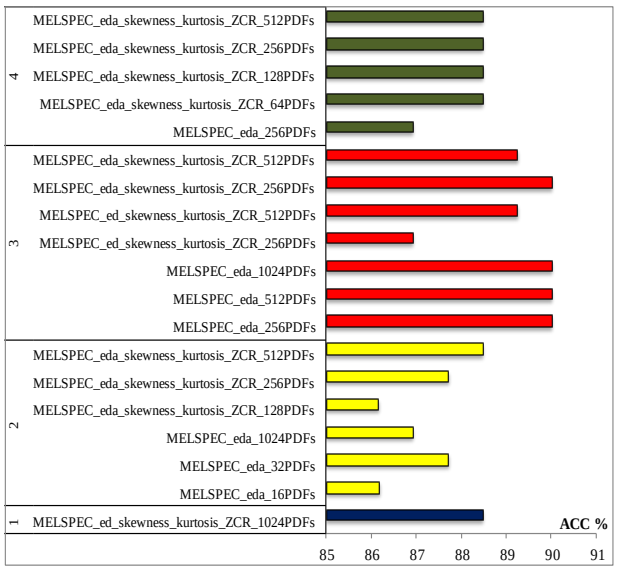

Fig. 8: Comparison of selected models (with ACC higher than 86 %).

Recognition results for the best HMM models, which were evaluated for this work are depicted in the Fig. 8. These models reached ACC higher than 86 %.

6. Conclusion

The dual shot detection with different features sets was performed and evaluated in this work. The addition of skewness, kurtosis and ZCR to the baseline MFCC often brought better results against baseline approach (particularly for the lower number of PDFs (64)). The achieved improvement did not overcome recognition results of the MELSPEC based approach.

The using of MELSPEC coefficients together with skewness, kurtosis and ZCR was unbalanced. Worst results are occurred in many cases. The correlated MELSPEC features with other time correlated parameters could lead to the misclassification.

For the dual shots detection task, the highest values of ACC (90 %) were reached by the MLESPEC_eda_3HMM – 256, 512, 1024 PDFs and MELSPEC_eda_skewness_ kurtosis_ZCR_3HMM – 256 PDFs.

Acknowledgements

The research presented in this paper was supported by the EU ICT Project INDECT (FP7 - 218086) (35 %) and Research and Development Operational Program funded by the ERDF under the projects ITMS-26220120030 (35 %) and by the Ministry of Education of Slovak Republic under research project VEGA 1/0386/12 (30 %).

References

[1] LATANE, B. and J. M. DARLEY. Bystander apathy. *American Scientist*. 1969, vol. 57, no. 2, pp. 244–268. ISSN 0003-0996.

[2] LATANE, B. and J. M. DARLEY. Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*. 1968, vol. 10, no. 3, pp. 215–221. ISSN 0022-3514. DOI: 10.1037/h0026570.

[3] NTALAMPIRAS, S., I. POTAMITIS and N. FAKOTAKIS. On acoustic surveillance of hazardous situations. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 - ICASSP 2009*. Taiwan: IEEE, 2009, pp. 165-168. ISSN 1520-6149. ISBN 978-1-4244-2353-8. DOI: 10.1109/ICASSP.2009.4959546.

[4] MITROVIC, D., M. ZEPPELYAUER and H. EIDENBERGER. Analysis of the data quality of audio descriptions of environmental sounds. *Journal of Digital Information Management*. 2007, vol. 5, no. 2, pp. 48–55. ISSN 0972-7272.

[5] KIM, H.G., N. MOREAU and T. SIKORA. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Chichester: John Wiley & Sons, 2005. ISBN 978-0-470-09334-4.

[6] PLEVA, M., E. VOZARIKOVA, L. DOBOS and A. CIZMAR. The joint database of audio events and backgrounds for monitoring of urban areas. In: *Journal of Electrical and Electronics Engineering*. 2011, vol. 4, no. 1. p. 185-188. ISSN 1844-6035.

[7] PSUTKA, J., L. MULLER and J.V. PSUTKA. Comparison of MFCC and PLP parametrizations in the speaker independent continuous speech recognition task. In: *Proceedings of Eurospeech 2001*. Aalborg: ISCA, 2001, pp. 1813-1816. ISBN 87-90834-09-7.

[8] YOUNG, Steve, Gunnar EVERMANN, Mark GALES, Thomas HAIN, Dan KERSHAW, Xunying (Andrew) LIU, Gareth MOORE, Julian ODELL, Dave OLLASON, Dan POVEY, Valtcho VALTCHEV, Phil WOODLAND. The THK Book: for HTK Version 3.4. In: *Cambridge University* [online]. 2009. Available at: http://www.ee.ucla.edu/~weichu/htkbook/.

[9] JAIN, A.K., R.P. W. DUIN and J. MAO. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*n. 2000, vol. 22, iss. 1, pp. 4-37. ISSN 0162-8828. DOI: 10.1109/34.824819.

[10] PSUTKA, J., L. MULLER, J. MATOUSEK and V. RADOVA. *Mluvime s pocitacem cesky.* Praha: Academia, 2006. ISBN 80-200-1309-1.

[11] E-Handbook of statistical methods. *NIST Information Technology Laboratory* [online]. 2012. Available at: http://www.itl.nist.gov/div898/-handbook/eda/section3/eda35b.htm.

[12] Indect: For the Security of Citizens. *Indect* [online]. 2000. Available at: http://www.indect-project.eu/

[13] VOZARIKOVA, E., M. LOJKA, J. JUHAR and A. CIZMAR. Performance of Basic Spectral Descriptors and MRMR Algorithm to the Detection of Acoustic Events. In: *Communications in Computer and Information Science: Multimedia Communications, Services and Security 2012.* Krakow: Springer, 2012, no. 287, pp. 350-359. ISBN 978-3-642-30720-1. DOI: 10.1007/978-3-642-30721-8_34.

[14] VOZARIKOVA, E., P. VISZLAY, J. JUHAR and A. CIZMAR. Acoustic Events Detection via PCA-based Feature Extraction. *Journal of Computer Science and Control Systems*. 2010, vol. 3, no. 2, pp. 99-102. ISSN 1844-6043.

[15] ZIEGER, Christian. An HMM based system for acoustic event detection. In: *Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science.* Berlin: Springer, 2008, vol. 4625. pp. 338-344. ISBN 978-3-540-68584-5. DOI: 10.1007/978-3-540-68585-2_32.

[16] VOZARIKOVA, E., M. PLEVA, J. JUHAR and A. CIZMAR. Surveillance system based on the acoustic events detection. *Journal of Electrical and Electronics Engineering*. 2011, vol. 4, no. 1, pp. 255-258. ISSN 1844-6035.

[17] BARRAS, C., E. GEOFFROIS, Z. WU and M. LIBERMAN. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication.* 2000, vol. 33. no. 1-2, pp. 5-22. ISSN 0167-6393

[18] Audacity software. *Audacity* [online]. 2012. Available at: http://audacity.sourceforge.net.

About Authors

Eva VOZARIKOVA was born in Liptovsky Mikulas**,** Slovakia in 1984. In 2009 she graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. She is a Ph.D. student at the same department in the field of Telecommunications. Her research is oriented on the field of the acoustic event detection and classification, speaker recognition and digital speech and audio processing.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as an Associate Professor at the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.

Anton CIZMAR was born in Michalovce, Slovakia in 1956. He is a graduate of the Slovak Technical University in Bratislava in 1980, at the Department of Telecommunications. He holds a 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. Now he works as a rector of the Technical University of Kosice. He is author and co-author of more than 170 scientific papers. His scientific research areas are broadband information and telecommunication technologies, multimedia systems, telecommunications networks and services, 4. generation mobile communications systems and localization algorithms.