Comparison of Different Feature Types for Acoustic Event Detection System

Eva Kiktova, Martin Lojka, Matus Pleva,
Jozef Juhar, and Anton Cizmar

Technical University of Kosice
Dept. of Electronics and Multimedia Communications, FEI TU Kosice
Park Komenskeho 13, 041 20 Kosice, Slovak Republic
{eva.kiktova,martin.lojka,matus.pleva,jozef.juhar,anton.cizmar}@tuke.sk
http://kemt.fei.tuke.sk

Abstract. With the increasing use of audio sensors in surveillance or monitoring applications, the detection of acoustic event performed in a real condition has emerged as a very important research problem. This paper is focused on the comparison of different feature extraction algorithms which were used for the parametric representation of the foreground and background sounds in a noisy environment. Our aim was to automatically detect shots and sounds of breaking glass in different SNR conditions. The well known feature extraction method like Mel-frequency cepstral coefficients (MFCC) and other effective spectral features such as logarithmic Mel-filter bank coefficients (FBANK) and Mel-filter bank coefficients (MELSPEC) were extracted from an input sound. Hidden Markov model (HMM) based learning technique performs the classification of mentioned sound categories.

Keywords: Feature extraction, acoustic event detection, SNR, HMM.

1 Introduction

The acoustic event detection is currently a very attractive research domain. It used knowledge from different scientific areas such as pattern recognition, machine learning, signal processing, artificial intelligence, etc. The term of acoustic event denotes the specific sound category, which is relevant for the particular task. It usually has a rare occurrence and it is hard to predict when and whether it occurs. This paper is focused on two sound classes, i.e. gun shot and breaking glass. These sounds are relevant for security applications and they belong to the foreground sounds, which appointed to an abnormal situation. In normal conditions, the foreground sounds do not occur, but when they occur, it probably determines some criminal activity. The intelligent security (or surveillance) system works according to the given instructions and it is not influenced or limited by various factors, not as a human. It should work autonomously and generate alert only when some dangerous situation is detected. It should also provide a constantly support for a police operator. The detection of acoustic events is a partial task in the complex security system [1].

A. Dziech and A. Czyżewski (Eds.): MCSS 2013, CCIS 368, pp. 288–297, 2013.

As was indicated in previous paragraph, the detection of acoustic event in real condition is a challenging task [2], [3], [4]. An unstable background sound, different weather conditions, changing values of SNR and many similar noisy non-event sounds (like trucks, trams, etc.) limit the performance of each security system. In this paper we investigated the impact of very important limiting factor- the noise level measured as Signal-to-Noise-Ratio (SNR).

Different types of audio features (MELSPEC, FBANK and MFCC) [5] and HMM prototypes (different number of states) were tested and finally we identified the suitable parametric representation. Events and background sounds were modeled using well known Hidden Markov Models (HMMs). Their number of states and topology (mainly ergodic) were based on the experiments.

The rest of the paper has the following structure: Section 2. presents the motivation and related works. Section 3. gives information about applied feature extraction methods and Section 4. gives information about the used part of sound database. Section 5. describes performed experiments and Section 6. summarizes obtained results, then follows the conclusion is Section 7.

2 Motivation and Related Works

Many algorithms can be used for the extraction of relevant features [6], [9]. They describe the nature of input sounds in the time, spectral, cepstral or some other transformation domain. MPEG-7 descriptors [8], speech inspired features [11], [12] are used very often. Some works contain the description of the acoustic event detection for surveillance applications, e.g. [2], [3].

Our previous works are focused on the feature extraction, which combines different approaches with the respect to the on-line applicable post-processing of features [6], [7] or another work which describes the long term monitoring performed by our own detector, which is based on the modified approach to detection of decision point, when it is located inside decoder (without any dependency on the front-end) [4].

Our proposed solution of the intelligent audio surveillance system has a promising results for a long-term audio-events monitoring application.

3 Feature Extraction

The feature extraction is a crucial aspect for each detection system, because the recognition performance depends on the quality of extracted feature vectors. In this work Mel-frequency cepstral coefficients (MFCC), Log Mel-filter bank coefficients (FBANK) and Mel-filter bank coefficients (MELSPEC) were used [11], [12]. Their computation process is described below.

The ear's perception of frequency components in the audio signal does not follow the linear scale, but rather the Mel-frequency scale 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:

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

where f is frequency in Hertz. MFCC, FBANK and MELSPEC coefficients are computed according to the Fig. 1.

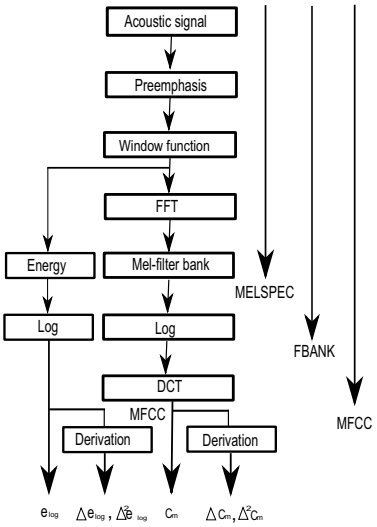


Fig. 1. Principial block scheme of MELPSEC, FBANK and MFCC coefficients

Normally, signal is filtered using preemphasis filter then the 25ms Hamming window method was applied on the frames. Then, they are transformed to the frequency domain via the discrete Fast Fourier Transform (FFT), and the magnitude spectrum is passed through a bank of triangular shaped filters. After this step we have MELSPEC features. The energy output from each filter is then log-compressed and represent FBANK features. Finally MFCC coefficients were obtained after the transformation to the cepstral domain by the Discrete Cosine Transform (DCT).

As it was described earlier the MELSPEC coefficients represent linear Mel-filter bank channel outputs and FBANK coefficients are logarithmic Mel-filter bank channel outputs. Mentioned features are computed during the extraction of popular MFCC coefficients.

4 Acoustic Event Database

The extended acoustic events database JDAE TUKE [10] used in this work involved the gun shot recordings with 463 realisations of shots from commonly

used weapons, breaking glass recordings with 150 realizations of broken glass and background recording (traffic sounds) with the duration of 53 minutes.

SNR influence was investigated on the recordings with different levels of SNR (Signal Noise Ratio). Nine new recordings (approximately 53 min) with different SNR i.e. -3dB, 0dB, 3dB, 6dB, 8dB, 11dB, 14dB, 17dB, 20dB were created and used in the training process. In the testing phase 33 min recordings with different SNR were recognised. They contained two series of continuous gun shots and two series of continuous breaking glass events. Shots and breaking glass sounds formed one class i.e. the acoustic events class, so they were evaluated together.

Background recordings include traffic sound, car honks, singing birds, and other loud sounds. All recordings have wav format with 48 kHz sampling frequency, resolution 16 bits per sample. They were cut and manually labeled using Transcriber[1].

5 Description of Experiments

Many internal experiments with speech based features such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC) were done. Relatively good results were obtained by MFCC coefficients, which achieved better recognition performance than the rest of the mentioned parametrization approaches. Features were extracted with 25 ms of Hamming window and 10 ms of frame step. 22 Mel-filter were used and the final number of coefficients was set to 12 (or 13 when using also energy or zero cepstral coefficient). Each acoustic HMM model was evaluated by the classifier based on the Viterbi decoder. Experiments presented in this paper were performed in HTK (Hidden markov model ToolKit) environment [5].

For this reason we decided to investigate MFCC approach more precisely. We evaluated different settings of MFCC with, zero cepstral coefficient (0), energy (E), delta (D), acceleration (A) coefficients and with or without cepstral mean normalisation (Z).

Generally the presence of energy E or 0th coefficient brought significant improvement in the relatively quiet environment without any louder sounds and with stable value of SNR. But, it is very difficult to restrict louder sounds and to have appropriate level of SNR in a real environment.

First and second time derivations of the baseline coefficients expanded feature vectors from 12 to 36 coefficients (or 13 to 39). The derivated coefficients had positive impact for each tested scenario. Generally speaking, also HMMs with higher number of PDFs achieved significant improvements.

Acoustic models trained with MFCC_EDA or MFCC_0DA features have in a testing process low detection performance because tested input sound had SNR ratio between -3dB and 20dB. These models were successfully applied only for particular SNR value of analysed audio signals.

[1] `http://trans.sourceforge.net/`

Cepstral mean normalisation (CMN) [5], [9] is very effective for the recognition performed in a noisy environment and it is de facto standard operation for most large vocabulary speech recognition systems. The CMN algorithm computes a long-term mean value of the feature vectors and subtracts the mean value from the all feature vectors. CMN reduces the variability of the data and allows simple but effective feature normalization. CMN in HTK is realized by the optional value "Z" [5].

On the basis of promising results with MFCC_DAZ, we decide to investigate also features, which were obtained before final feature set. Therefore MELSPEC_DAZ and FBANK_DAZ were extracted too.

In this work, two different system architectures were analysed. First is based on the HMM models for shots, glass and several backgrounds, see Fig. 2. We used a separate background model for each SNR value. In these experiments ergodic HMMs from one to four states and from 1 to 1024 Probability Density Functions (PDFs) were trained and evaluated in offline tests.

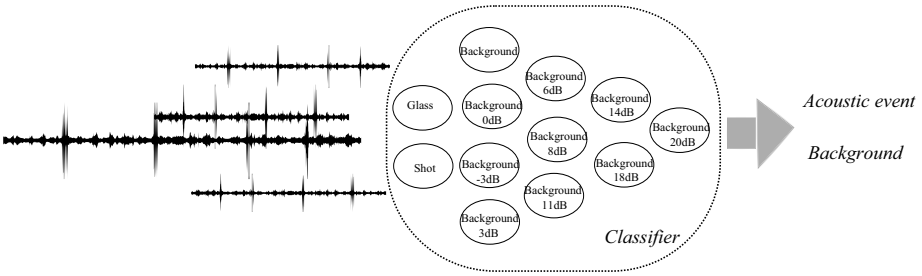


Fig. 2. Description of first detection system scheme

We considered also the second system architecture, which is based only on the one HMM for background. The rest of background models were not applied in the testing process. Models for shots and glass were the same. The principal scheme of this approach is depicted in the Fig. 3.

6 Results of Experiments

Three types of features MELSPEC_DAZ, FBANK_DAZ and MFCC_DAZ were used for the evaluation of the proposed robust system. We supposed that several background models will by more suitable for recognition of acoustic events in different SNR conditions. The testing set consisted from 9 recordings, more details are available in the Section 4. Obtained results of experiments is depicted in the Fig.4, where the results were averaged for each model type using widely used Accuracy measurement technique [5], [6].

The approach based on the MFCC_DAZ parametrization achieved results higher than 75% only in the two cases, i.e. ACC = 76,67% was achieved by four states HMM with 1 PDF (HMM_4_1) and ACC = 75,56% with HMM_3_1.

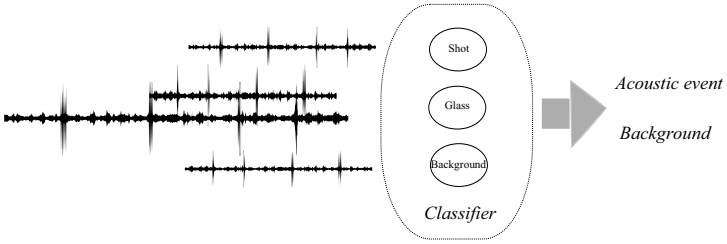


Fig. 3. Description of second detection system scheme

FBANK_DAZ features obtained promising results. The accuracy higher then 75% were achieved nine times. The highest value of ACC=87,78% was reached by HMM_1_32, second best results ACC=86,67% belonged to HMM_3_1.

Like a previous case, MELSPEC_DAZ features were evaluated by the same way. Four times ACC higher then 75% were yielded. The best MELSPEC models HMM_1_64 achieved 80% ACC.

Generally suitable results were yielded by FBANK features, otherwise MFCC features seem to be the less appropriate. As it was mentioned, the results from Fig. 4 are averaged.

More detailed information about SNR impact for selected best models that yielded three highest ACC [%] (based form Fig. 4) will be presented in details for first system architecture in Fig. 5 and for second architecture in Fig.6.

The Fig.5 depicted the obtained results for system with several background models and Fig.6 refers to the system with one background model.

As you can see in the Fig.5 and Fig.6 many models had the same performance. From this point of wiev the presence of several models was not as important as we supposed. The CMN operation as a partial operation in the feature extraction process fixed the SNR problem very effective way.

- **The first system architecture with several SNR background models (Fig.5)**
  Perfect recognition results only for MELSPEC_DAZ were reached. Recordings with SNR 20dB, 17dB, 14dB were recognised overall seven times with ACC = 100% with using HMM_1_32, HMM_1_64 PDFs and HMM_1_128 PDFs. In the cases of very noisy conditions (SNR = -3 dB, 0dB) one state MELSPEC_DAZ models failed. These models achieved the lowest values of ACC = 40%. FBANK_DAZ models detected events in range of 80% to 90% of ACC. MFCC_DAZ models achieved lower ACC values.

- **The second system architecture with one background model (Fig.6)**
  The proposed system worked similarly as in the case of several background models. The perfect recognition result of ACC=100% was achieved nine times only for MELSPEC_DAZ approach. HMM_1_64 correctly recognised testing recordings with SNR 11dB, 14dB, 17dB and 20dB. Good performances were achieved also by MELSPEC_DAZ HMM_1_32 for recordings with SNR 14dB, 17dB, 20dB and HMM_1_128 for SNR 17dB and 20dB.

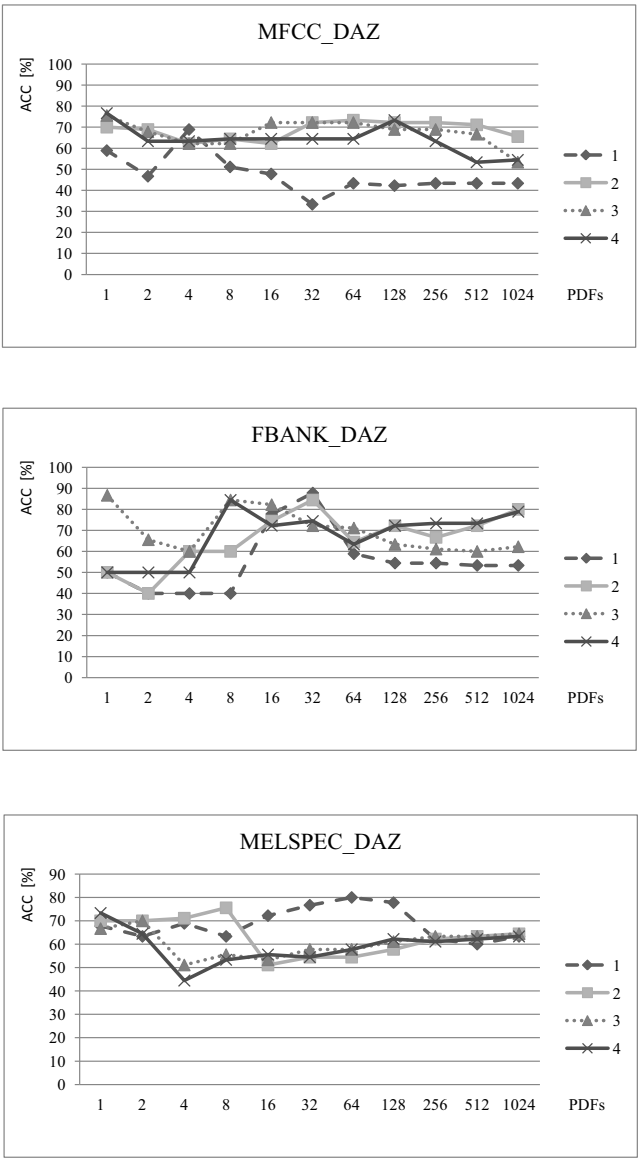
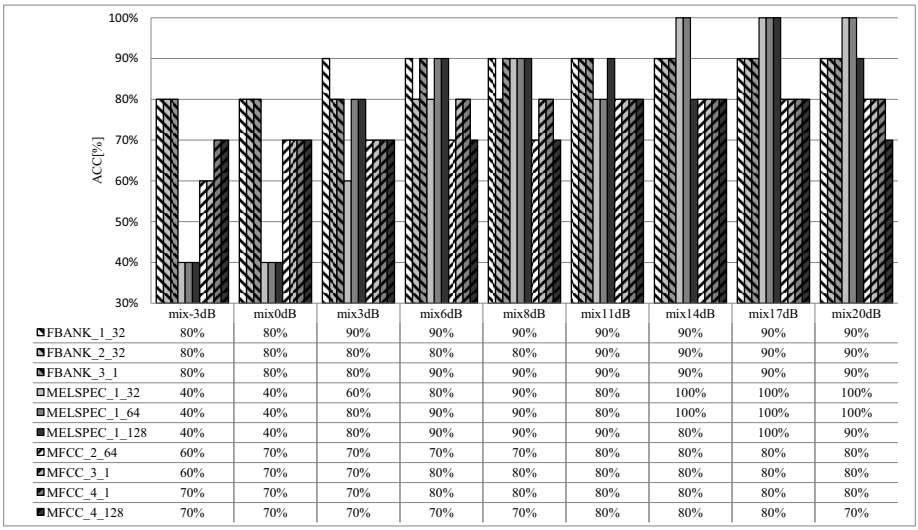
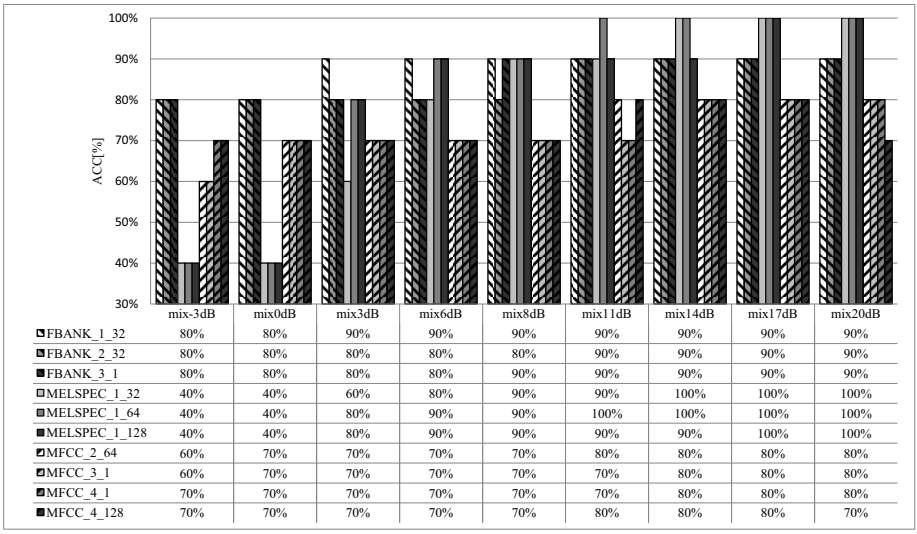


Fig. 4. The recognition results of average ACC [%] for each parametrization, (number of HMM states is depicted in legend)



	mix-3dB	mix0dB	mix3dB	mix6dB	mix8dB	mix11dB	mix14dB	mix17dB	mix20dB
FBANK_1_32	80%	80%	90%	90%	90%	90%	90%	90%	90%
FBANK_2_32	80%	80%	80%	80%	80%	90%	90%	90%	90%
FBANK_3_1	80%	80%	80%	90%	90%	90%	90%	90%	90%
MELSPEC_1_32	40%	40%	60%	80%	90%	80%	100%	100%	100%
MELSPEC_1_64	40%	40%	80%	90%	90%	80%	100%	100%	100%
MELSPEC_1_128	40%	40%	80%	90%	90%	90%	80%	100%	90%
MFCC_2_64	60%	70%	70%	70%	70%	80%	80%	80%	80%
MFCC_3_1	60%	70%	70%	80%	80%	80%	80%	80%	80%
MFCC_4_1	70%	70%	70%	80%	80%	80%	80%	80%	80%
MFCC_4_128	70%	70%	70%	70%	70%	80%	80%	80%	70%

Fig. 5. The recognition results of ACC [%] for selected HMMs (several background models)

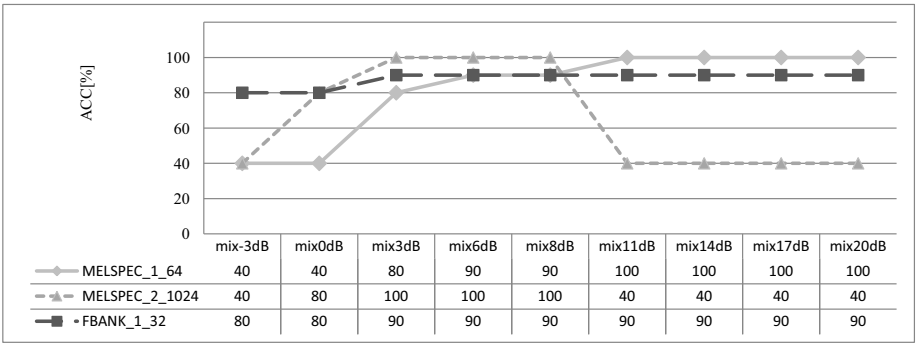


	mix-3dB	mix0dB	mix3dB	mix6dB	mix8dB	mix11dB	mix14dB	mix17dB	mix20dB
FBANK_1_32	80%	80%	90%	90%	90%	90%	90%	90%	90%
FBANK_2_32	80%	80%	80%	80%	80%	90%	90%	90%	90%
FBANK_3_1	80%	80%	80%	80%	90%	90%	90%	90%	90%
MELSPEC_1_32	40%	40%	60%	80%	90%	90%	100%	100%	100%
MELSPEC_1_64	40%	40%	80%	90%	90%	100%	100%	100%	100%
MELSPEC_1_128	40%	40%	80%	90%	90%	90%	90%	100%	100%
MFCC_2_64	60%	70%	70%	70%	70%	80%	80%	80%	80%
MFCC_3_1	60%	70%	70%	70%	70%	70%	80%	80%	80%
MFCC_4_1	70%	70%	70%	70%	70%	70%	80%	80%	80%
MFCC_4_128	70%	70%	70%	70%	70%	80%	80%	80%	70%

Fig. 6. The recognition results of ACC [%] for selected HMMs (one background model)

The results presented in the Fig.5 and the Fig.6 appointed better recognition performance for systems with one background model. For more noisy recordings (SNR= 6dB, 8dB) the system with SNR depended HMM models seems to be slightly better. MELSPEC_DAZ especially for HMM_1_64 yielded very good results when SNR ratio was higher. The detailed analysis of MELSPEC_DAZ results showed that other perfect recognition results were yielded by HMM_2_512, HMM_2_1024 and HMM_3_512 for SNR= 3dB, 6dB and 8dB.

Other MELSPEC_DAZ models reached usually ACC=40% for SNR -3dB and 0dB, therefore we analysed other approaches (MFCC_DAZ and FBANK_DAZ) for finding the most suitable models regarding to the SNR (-3dB and 0dB). Balanced results were achieved by the FBANK_DAZ where the recognition results for low and high SNR were in the range from 80% to 90% of ACC. The selection of the overall best recognition results is depicted in the Fig. 7.



	mix-3dB	mix0dB	mix3dB	mix6dB	mix8dB	mix11dB	mix14dB	mix17dB	mix20dB
MELSPEC_1_64	40	40	80	90	90	100	100	100	100
MELSPEC_2_1024	40	80	100	100	100	40	40	40	40
FBANK_1_32	80	80	90	90	90	90	90	90	90

Fig. 7. The best achieved recognition results for different SNR ratio

7 Conclusion

This paper evaluated the acoustic event detection in the urban environment in consideration of the very important limiting factor - the noise level measured by SNR. We analysed the detection performance for different SNR conditions with using MFCC, FBANK and MELSPEC features. Enhancements such as delta, acceleration coefficients and cepstral mean normalisation were applied as robust features for acoustic events recognition.

Performed experiments showed that MELSPEC_DAZ and FBANK_DAZ features are able to distinguish the presence of acoustic events in different SNR conditions more accurately than widely used MFCC_DAZ. Promising results especially for MELSPEC_DAZ models were achieved. Delta and acceleration coefficients were used for incorporate temporal information in acoustic event features and cepstral mean normalisation contributed apparently to the robustness of created acoustic models.

Acknowledgments. This work has been performed partially in the framework of the EU ICT Project INDECT (FP7 - 218086) and by the Ministry of Education of Slovak Republic under research VEGA 1/0386/12 and under research project ITMS-26220220155 supported by the Research & Development Operational Programme funded by the ERDF.

References

1. INDECT project homepage, http://www.indect-project.eu/
2. Clavel, C., Ehrette, T., Richard, G.: Events Detection for an Audio-Based Surveillance System. In: IEEE International Conference on Multimedia and Expo. 2005, pp. 1306–1309 (2005)
3. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio Based Event Detection for Multimedia Surveillance. In: IEEE International Conference on Acoustics, Speech and Signal Processing 2006, vol. 5, pp. 813–816 (2006)
4. Pleva, M., Lojka, M., Juhar, J., Vozarikova, E.: Evaluating the Modified Viterbi Decoder for Long-Term Audio Events Monitoring Task. In: 54th International Symposium Croatian Society Electronics in Marine - Elmar, pp. 179–182 (2012)
5. Young, S., et al.: The HTK Book, p. 368. Cambridge University (2006)
6. Vozarikova, E., Lojka, M., Juhar, J., Cizmar, A.: Performance of Basic Spectral Descriptors and MRMR Algorithm to the Detection of Acoustic Events. In: Dziech, A., Czyżewski, A. (eds.) MCSS 2012. CCIS, vol. 287, pp. 350–359. Springer, Heidelberg (2012)
7. Vozáriková, E., Juhár, J., Čižmár, A.: Acoustic Events Detection Using MFCC and MPEG-7 Descriptors. In: Dziech, A., Czyżewski, A. (eds.) MCSS 2011. CCIS, vol. 149, pp. 191–197. Springer, Heidelberg (2011)
8. Kim, H.G., Moreau, N., Sikora, T.: MPEG-7 audio and beyond: Audio content indexing and retrieval, p. 304. Wiley (2005)
9. Toh, A.M., Togneri, R., Nordhoolm, S.: Investigation of Robust Features for Speech Recognition in Hostile Environments. In: Asia-Pacific Conference on Communications 2005, pp. 956–960 (2005)
10. Pleva, M., Vozarikova, E., Dobos, L., Cizmar, A.: The joint database of audio events and backgrounds for monitoring of urban areas. Journal of Electrical and Electronics Engineering 4(1), 185–188 (2011)
11. Psutka, J., Müller, L., Psutka, J.V.: Comparison of MFCC and PLP parametrizations in the speaker independent continuous speech recognition task. In: Eurospeech 2001, pp. 1813–1816 (2001)
12. Wong, E., Sridharan, S.: Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In: International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 95–98 (2001)