Acoustic Events Detection Using MFCC and MPEG-7 Descriptors

Eva Vozáriková, Jozef Juhár, and Anton Čižmár

Technical university of Košice, Slovak Republic
Dept. of Electronics and Multimedia Communications, FEI TU Košice
Park Komenského 13, 041 20 Košice, Slovak Republic
{eva.vozarikova,jozef.juhar,anton.cizmar}@tuke.sk
http://kemt.fei.tuke.sk/

Abstract. This paper is focused on the acoustic events detection. Particularly two types of acoustic events (gun shot, breaking glass) were investigated. For any detection task the feature extraction methods play very important role. The feature extraction influences the recognition rate, therefore it is most important in any pattern recognition task. In this paper the impact of Mel-Frequency Cepstral Coefficients - MFCC and selected set of MPEG-7 low-level descriptors were examined. The best feature set contained MFCC and selected descriptors such as ASC, ASS, ASF. They were used to represent the sounds of acoustic events and background. We obtained the improvement of the detection rate using the mentioned set of features. In this task GMM classifiers are used to model the sound classes. This paper describes a basic aspect of our work.

Keywords: Acoustic events, feature extraction, MFCC, MPEG-7 low-level descriptors.

1 Introduction

Audio surveillance systems are highly requested systems nowadays. Over the past decades a great deal of work has been done and published in the area of detection and classification input pattern [1], [2]. The complex surveillance system [3] can be created by the fusion of sound and video information. The effectiveness of the surveillance system depends on environmental conditions. The visual part of detection system will probably fail in terms of the bad light condition or if the problematic situation is not in the visual field of surveillance camera. On the other hand, the audio part of system is very sensitive on the sound similarity. Generally, extreme weather conditions limit the performance of any surveillance system.

Surveillance systems are usually used at monitor public places, stadiums, vehicles and stations of public transport, etc. Some dangerous situations are more easily detectable via sound information than the video information for example calling for help, sound of gun shots, etc. Of course, there are many applications

A. Dziech and A. Czyżewski (Eds.): MCSS 2011, CCIS 149, pp. 191–197, 2011.

[4], [5], where the detection of specific sounds can be very helpful, e.g. in smart rooms, health care or industry applications, etc.

The goal of each surveillance system is help to protect life and property. Detection systems should generate the alert only if dangerous event is detected. It is very important to reduce false alarm as a result of incorrectly classified pattern. The proposed detection system is created to recognize potentially dangerous situations via sound information. Especially, the effort is to detect two types of acoustic events such as gun shot and breaking glass. These sounds represent abnormal behavior and they point to existence of some dangerous situation.

Acoustic signals have information redundancy and for this reason is necessary to specify effective feature extraction methods. The effective feature extraction should highlight the relevant information and reduce the number of input data by removing irrelevant information using various kinds of decorrelation methods.

One of the most popular approaches is based on the feature extraction methods that primary for speech signals were developed. The well known feature extraction methods such as MFCC (Mel-Frequency Cepstral Coefficients) [6], [7], or PLP (Perceptual Linear Prediction) coefficients [7] are used to extract the relevant information from the input speech signal.

Other effective approach exploits the low-level audio descriptors [8], [9], defined in MPEG-7 standard. MPEG-7 is focused on the describing of the multimedia content. It is oriented on the indexing, searching and retrieval of audio using the 17 low-level descriptors [10]. These descriptors can capture the nature of input acoustic signal.

This paper is focused on the feature extraction method that include the selected set of MPEG-7 low-level descriptors such as Audio Spectrum Spread (ASS), Audio Spectrum Centroid (ASC) and Audio Spectrum Flatness (ASF) and the advantages of speech parametrization like MFCC. In the classification stage Gaussian Mixture Models (GMMs) are used.

The rest of this paper has following structure. Section 2 describes the feature extraction methodology and Section 3 gives information about the proposed detection system. Section 4 includes experiments and results, finally the conclusion and future work proposal follows in Section 5.

2 Feature Extraction Methodology

As was mentioned in previous section the feature extraction method is very important part of the detection system. In general, a typical pattern recognition task can be divided into the feature extraction and classification task. The efficient feature extraction is a very important phase of overall process, because the recognition performance directly depends on the quality of extracted feature vectors.

In this paper we investigate the discriminative power of selected MPEG-7 descriptors in comparison of conventional speech parametrization MFCC.

2.1 MPEG-7 Low-Level Descriptors

MPEG-7 is an ISO/IEC standard developed in by the Moving Picture Experts Group (MPEG). It became an international standard in September 2001 [10] and includes the part dealing with audio information. This part of standard is MPEG-7 Audio. There are defined 17 low-level descriptors.

In our experiments basic spectral descriptors namely Audio Spectrum Centroid, Audio Spectrum Spread and Audio Spectrum Flatness were chosen according to the good results that were presented in the works [2], [8], [11].

Audio Spectrum Centroid - ASC

The Audio Spectrum Centroid (ASC) [10] gives the centre of gravity of a log-frequency power spectrum. All power coefficients below 62.5 Hz are summed and represented by a single coefficient, in order to prevent a non-zero DC component and /or very low-frequency components which can have a disproportionate weight. For a given frame of signal, ASC descriptor is computed from the modified power coefficients and their frequencies. In the Eq. (1) $P'(k')$ represents the power spectrum and $f'(k')$ represent corresponding frequencies.

$$ASC = \frac{\sum_{k'=0}^{(N_{FT}/2)-K_{low}} log_2\left(\frac{f'(k')}{1000}\right) P'(k')}{\sum_{k'=0}^{(N_{FT}/2)-K_{low}} P'(k')}. \quad (1)$$

ASC descriptor gives information on the shape of the power spectrum. It indicates whether in a power spectrum are dominated by low or high frequencies and can be regarded as an approximation of the perceptual sharpness of the signal.

Audio Spectrum Spread - ASS

The Audio Spectrum Spread (ASS) [10] is also called instantaneous bandwidth. It is measure of the spectral shape. In MPEG-7, it is defined as the second central moment of the log-frequency spectrum. For a given signal frame ASS is computed following way:

$$ASS = \frac{\sum_{k'=0}^{(N_{FT}/2)-K_{low}} \left[log_2\left(\frac{f'(k')}{1000}\right) - ASC\right]^2 P'(k')}{\sum_{k'=0}^{(N_{FT}/2)-K_{low}} P'(k')}. \quad (2)$$

ASS descriptor is extracted by taking the root-mean-square (RMS) deviation of the spectrum from its centroid ASC. The ASS gives indications about how

the spectrum is distributed around its centroid. A low ASS value means that the spectrum may be concentrated around the centroid, whereas a high value reflects a distribution of power across a wider range of frequencies.

Audio Spectrum Flatness - ASF

The Audio Spectrum Flatness (ASF) [10] reflects the flatness properties of the power spectrum. More precisely, for a given signal frame, it consists of a series of values, each one expressing the deviation of the signal's power spectrum from a flat shape inside a predefined frequency band. In MPEG-7, the power coefficients are computed from non-overlapping frames where the spectrum B is divided into $1/4$ octave resolution logarithmically spaced overlapping frequency bands. For each band b, the spectral flatness descriptor is estimated as the ratio between the geometric mean and the arithmetic mean of the spectral power coefficients within this band:

$$ASF(b) = \frac{\sqrt[hiK'_b - loK'_b + 1]{\Pi_{k'=loK'_b}^{hiK'_b} P_g(k')}}{\frac{1}{hiK'_b - loK'_b + 1} \sum_{k'=loK'_b}^{hiK'_b} P_g(k')}, \qquad (1 \leq b \leq B). \tag{3}$$

For all bands under the edge of 1 kHz, the power coefficients are averaged in the normal way. For all bands above 1 kHz, power coefficients are grouped $P_g(k')$. The terms $hiK'b$ and $loK'b$ represent the high and low limit for band b. High values of ASF coefficients reflect noisiness, on the other hand, low values indicate a harmonic structure of the spectrum.

2.2 Mel-Frequency Cepstral Coefficients - MFCC

Ear's perception of the 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 Eq.(4):

$$Mel(f) = 2595 \times log_{10}\left(1 + \frac{f}{700}\right), \tag{4}$$

where f is frequency in Hertz. MFCC coefficients are computed following way: a segment of signal is divided into the short frames, where the parameters of the signal are constant. The Hamming window method was applied on the frames. Then, they are transformed to the frequency domain via the Discrete Fast Fourier Transform (DFFT), 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) [8].

3 System Overview

Our system was developed to the purpose of the gun shots and breaking glass detection. Gaussian Mixture Models (GMMs) were used as a learning algorithm. For each acoustic event GMMs were trained up to 64 Gaussian Probability Density Functions (PDFs). Three types of feature sets (MFCC E, ASS ASC ASF and MFCC E+ASS ASC ASF) were compared.

Three types of low-level audio descriptors were used in our work. ASS, ASC and ASF descriptors that give promising results were computed from acoustic signal. ASF generated the vector with 24 coefficients for each signal frame and a scalar value was generated by ASS and ASC. The final MPEG-7 feature vector composed of 26 (24+1+1) coefficients per frame. The extraction of these descriptors was done in Matlab.

MFCC features were computed from the signal divided into the 30 ms frames with 50% overlapping between the neighboring frames. 29 triangular band filters were used. MFCC feature extraction algorithm generated 13 coefficients (12 static coefficients and log-energy coefficient E). Extraction of MFCC was done using HCopy from HTK toolkit.

The coefficients of mentioned feature extraction algorithms were finally jointed to the one supervector with dimension 39 (MFCC E+ASS+ASC+ASF = 13+1+1+24). Each feature extraction approach was performed and evaluated. The audio data of acoustic events used throughout the paper were recorded in relatively quiet environment. The sound data were recorded with sampling frequency 48 kHz with resolution of 16 bits per sample. Then, they were split into the training and the testing set. Recordings were cut and manually labeled using Transcriber. It is important to note that the definition of acoustic event and sound of background was specific for this application. Each sound that was not any acoustic event was considered as a background sound.

The GMM models were trained by the 40 recordings of shot, with the 40 recordings of breaking glass and 11 minutes of background sounds. Testing recordings were used to evaluate the detection system. The total duration of testing recordings was 40 seconds. They contained non-overlapping sounds of acoustic events and background sounds.

4 Experiments and Results

The performed experiments were focused on the detection of breaking glass and gun shot. HTK toolkit was used in this task. In the first step MFCC coefficients and log energy coefficients E were computed from acoustic signal. Then ASS, ASC, ASF descriptors were computed and added in to the one MPEG-7 vector. Finally MFCC E + MPEG-7 vector was used to describe the input acoustic signal. For evaluating the proposed detection system the measure accuracy [%] was used. Accuracy [%] is defined as:

$$ACC\ [\%] = \frac{N - D - S - I}{N} \times 100, \tag{5}$$

Table 1. Breaking glass detection - *ACC* [%]

Num. of PDFs	MFCC E	MPEG-7	MFCC E+MPEG-7
1	77.78	44.44	100
2	33.33	44.44	88.89
4	22.22	44.44	100
8	55.56	66.67	100
16	33.33	66.67	100
32	11.11	55.56	100
64	11.11	77.78	33.33

Table 2. Gun shot detection - *ACC* [%]

Num. of PDFs	MFCC E	MPEG-7	MFCC E+MPEG-7
1	87.50	37.50	87.50
2	62.50	50.00	87.50
4	87.50	50.00	87.50
8	87.50	50.00	87.50
16	62.50	50.00	87.50
32	37.50	62.50	62.50
64	12.50	62.50	37.50

where D is the number of deletion errors, S the number of substitution errors, I the number of insertion errors, and N is the total number of labels in the reference transcription files [12]. The results are depicted on the tables.

Table 1 presents the results of the breaking glass detection. The best recognition results were obtained with MFCC E and MPEG-7 feature extraction. Every sound of testing recording was recognized correctly in case of GMMs (1, 4, 8, 16, 32 PDFs). Table 2 shows the results of the gun shot detection. In this case, the joint set of MFCC E and MPEG-7 enabled to achieve comparable or better results against first two parametrization approaches. For the breaking glass and gun shot detection, the higher number of PDFs brought higher values of accuracies for sets of MPEG-7 descriptors in compare of MFCC E, where better results were occurred with the lower number of PDFs.

5 Conclusion and Future Work Proposal

Presented results give us some base information about gun shot and breaking glass detection using of selected MPEG-7 descriptors and MFCC. We can notice that the results of the breaking glass detection was better against the results of the shot detection. It was probably caused by the location of acoustic events (gun shots) in the testing recordings. They were placed immediately behind one to another. In the future, we would like to evaluate the discrimination capability of each descriptor to the particular types of acoustic events. Also, we suppose that the combination of speech feature extraction algorithm, low-level descriptors and

feature reduction method can be very effective in this detection task, because the use of several descriptors and speech based parametrizations e.g. MFCC leading to the multidimensional feature vectors (supervectors). For this reason, Principal Component Analysis (PCA) can be apply to reduce the input feature supervectors. Extending of the sound corpus is also in progress.

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 1/0065/10.

References

1. Huang, W., Lau, S., Tan, T., Li, L., Wyse, L.: Audio events classification using hierarchical structure. In: ICICS-PCM, pp. 1299–1303 (2003)
2. Ghulam, M., Yousef, A.A., Mansour, A., Mohammad, N.H.: Environment recognition using selected MPEG-7 audio features and Mel-Frequency Cepstral Coefficients. In: International Conference on Digital Telecommunications, pp. 11–16 (2010)
3. Cristiani, M., Bicego, M., Murino, V.: Audio-visual event recognition in surveillance video sequences. IEEE Transactions on Multimedia 9/2, 257–266 (2007)
4. Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M.: Acoustic Event Detection and Classification in Smart-Room Environment: Evaluation of CHIL Project Systems. In: The IV Biennial Workshop on Speech Technology (2006)
5. Rougui, J.E., Istrate, D., Souidene, W.: Audio sound event identification for distress situations and context awareness. In: International Conference of Engineering in Medicine and Biology Society, Minneapolis, September 3-6, pp. 3501–3504 (2009)
6. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. Journal of Computer Science and Technology 16/6, 582–589 (2001)
7. Psutka, J., Müller, L., Psutka, J.V.: Comparison of MFCC and PLP parametrizations in the speaker independent continuous speech recognition task. In: Eurospeech, Aalborg, September 3-7, pp. 1813–1816 (2001)
8. Mitrovic, D., Zeppelzauer, M., Eidenberger, H.: Analysis of the data quality of audio descriptions of environmental sounds. Journal of Digital Information Management 5/2, 48–55 (2007)
9. Casey, M.: General sound classification and similarity in MPEG-7, pp. 153–164. Cambridge University Press, Cambridge (2001)
10. Kim, H.G., Moreau, N., Sikora, T.: MPEG-7 audio and beyond: Audio content indexing and retrieval, p. 304. Wiley, Chichester (2005); ISBN: 978-0-470-09334-4
11. Ntalampiras, S., Potamitis, I., Fakotakis, N.: Automatic recognition of urban environmental sounds events. In: CIP, Santorini, June 9-10, pp. 110–113 (2008)
12. Young, S., et al.: The HTK Book. Cambridge University, Cambridge (2009)