Audio Event Detection in Noisy Environments

Htet Htet Aung
University of Computer Studies, Yangon
htethtetaung.ucsy@gmail.com

Abstract

This paper describes an approach for an audio event detection system in noisy environments. The system specifically focuses on classification of an audio event as gunshot, scream or ambient noise. For discriminating gunshot from noise and scream from noise, two parallel Gaussian Mixture Model (GMM) classifiers are applied. Acoustic features such as zero-crossing rate, mel frequency cepstral coefficients, spectral flatness measures are firstly extracted to train GMM classifier. Each GMM classifier is trained using different set of audio features. To reduce the false detection rate, the decision that an event (gunshot or scream or noise) is taken by computing logical OR of the two classifiers. The efficiency of this scheme is investigated over audio recordings taken from internet repositories. The experimental results show that the overall accuracy of the system is as high as 97%.

1. Introduction

Audio is useful especially in situations when other sensor such as video fails to reliability detect the events. Automated audio analysis is a challenging problem. Audio events detection/ classification are receiving a growing interest by the scientific community. A large amount of work has been carried out in speech recognition, in audio segmentation and classification and more recently, in audio source separation and localization. Audio based surveillance has been studied earlier for detecting various types of acoustic events such as human's coughing in the office environment, impulsive sound like glass break, explosions or door alarms. Audio events classification is considered in this work. In particular, the events such as gunshot, scream and noise are considered in this paper.

Recently, works on audio classification/audio event detection have proved that a hierarchical classification scheme consisting of different levels of binary classifiers generally achieves higher performance than classification of multiple audio classes using a single level classifier. A set of cascaded GMM is used to classify 5 different sound classes in [1]. A method used for audio based event detection for multimedia surveillance is proposed in [2]. The hierarchical approach has also been employed to design a specific system able to detect the scream in public transport systems. This system is tested using both GMMs and SVMs as classifier. It is shown that in general GMMs provide higher precision. Event detection for an audio-based surveillance system is proposed in [3]. The use of audio sensors in surveillance and monitoring applications has proved to be useful for the detection of events like gunshots. For feature extraction, a large set of audio features for sound description (similarity and classification) is presented in [4]. Unsupervised learning of finite mixture model is applied for audio classification [5]. Audio based surveillance stems from the field of automatic audio classification and matching. Traditional tasks in this area are speech/music segmentation and classification [6, 7] and audio retrieval [8].

In this work, there are two separate steps in gunshot and scream detection from noise. In the first step, the audio signal is split into segments by detecting abrupt changes in the signal statistics (Feature Extraction). Second step, the extracted segments are classified as gunshot, scream and noise by using GMM classifiers (Classification).

2. Background

2.1. Audio Features

A number of audio features have been used for the tasks of audio analysis and content-based audio retrieval. There are many types of audio features: temporal features, e.g. Zero Crossing Rate (ZCR): energy features, e.g. Short Time Energy (STE): spectral features, e.g. spectral moments, spectral flatness: perceptual features, e.g. loudness, sharpness or Mel Frequency Cepstral Coefficients (MFCCs). In this work, different types of features are employed for classification. (1) Zero crossing

rate (ZCR) (2) spectral flatness measure (SFM) (3) spectral decrease (4) spectral centriod (5) mel frequency cepstral coefficients (MFCC). For impulsive noises, like gunshots, much of the energy is concentrated in the first time lags, while for harmonic sounds, like screams the energy is spread over a wide range of time lags.

2.1.1. Zero Crossing Rate (ZCR)

It is a measure of the number of time the signal value cross the zero axis. This feature helps in distinguishing the excited events from the normal events.

$$Z_n = \frac{1}{2}\sum_m |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| w(n-m) \quad (1)$$

where $\mathrm{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$

2.1.2. Spectral Flatness Measure

It is a measure of the noisiness (flat, decorrelation)/sinusoidality of a spectrum (or a part of it). It is computed by the ratio of the geometric mean to the arithmetic mean of the energy spectrum value.

$$SFM(num\_band) = \frac{\left[\prod_{k \in num\_band} a(k)\right]^{\frac{1}{k}}}{\frac{1}{k}\sum_{k \in num\_band} a(k)} \quad (2)$$

$a(k)$ is the amplitude in frequency band number k

2.1.3. Spectral Decrease

It represents the amount of decreasing the spectral amplitude.

$$decrease = \frac{1}{\sum_{k=2:K} a} \sum_{k=2:K} \frac{a(k) - a(1)}{k-1} \quad (3)$$

2.1.4. Mel Frequency Cepstral Coefficients (MFCC)

It represents the shape of the spectrum with very few coefficients. It is the coefficients of the Mel cepstrum. The cepstrum, is the Fourier Transform (or Discrete Cosine Transform DCT) of the Mel cepstrum. The Mel cepstrum is the cepstrum computed on the Mel bands instead of the Fourier spectrum. The use of mel scale allows better to take into the mid-frequencies part of the signal. Figure 1 shows the block diagram for computing MFCC coefficients.

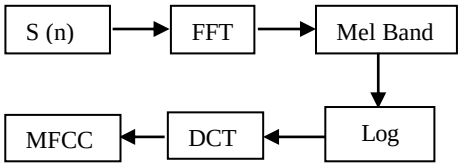

Figure 1. The block diagram of MFCC calculation

2.1.5. Spectral Centriod

It is computed considering the spectrum as a distribution which values are the frequency and the probabilities to observe these are the normalized amplitude.

$$\mu = \int x.p(x)\delta x \quad (4)$$

x are the observed data: $x = freq\_v(x)$
$p(x)$ is the probability to observe x:

$$p(x) = \frac{ampl - v(x)}{\sum_x amp - v(x)} \quad (5)$$

2.2. GMM classifier

Classifier is an algorithm with features as input and concludes what it means based on information that is encoded into the classifier algorithm and its parameters. The output is usually a label, but it can contain also confidence values. The classifiers usually employed in most classification frameworks are Gaussian Mixture Models (GMM), Maximum a Posterior (MAP), Hidden Markov Models (HMM), Support Vector Machine (SVM), neural networks and etc. GMM is a parametric method. It belongs to the class of pattern recognition systems. The parameters in GMM are tuned using a complex iterative procedure called expectation-maximization (EM) algorithm, which aims at maximizing the likelihood of the training set generated by the estimated probability density function (PDF). The major applications of GMM are clustering and classification. The well-known applications of GMM are image segmentation, edge detection, pattern recognition, motion tracking, watermarking, speaker verification, voice recognition, infrared object modeling and etc.

3. Method

The purpose of this paper is to segment the input audio stream into 3 label classes such as gun shot segment, noise segment or scream segment. The

framework of this audio event detection is illustrated in Figure 2.

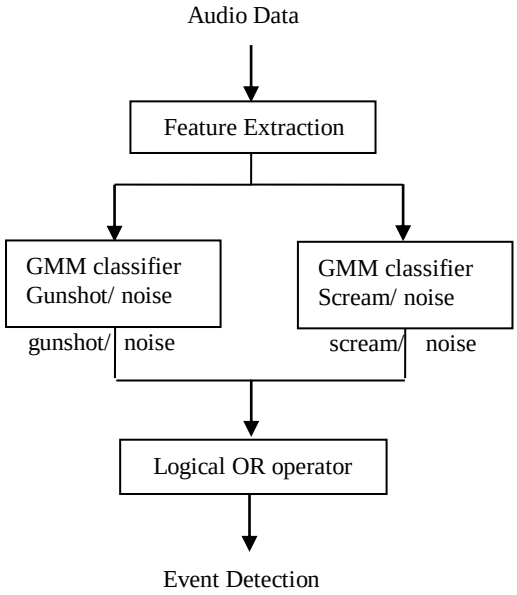

Figure 2: Framework of the method

3.1. Feature Extraction

This system considers 35-dimensional features comprising of energy, zero-crossing rate, spectral centroid, spectral flatness measure, spectral decrease and 30 MFCC coefficients. Input audio stream is assumed to have sampling frequency of 22k Hz. To cope with the dynamic nature of complex audio, all features are extracted from 20ms frames with 50% overlap. Hamming window is then applied to these divided frames. After each frame is windowed, time domain features such as energy and zero crossing rates are computed. Windowed signal is again converted into frequency domain using fast Fourier transform. Take the absolute value to compute frequency domain features such as spectral decrease, spectral centriod, spectral flatness measure and Mel frequency cepstral coefficients. Figures 3-7 show the example plots of extracted features (zero crossing rate and spectral flatness measure) for gunshot, scream and noise. Here, each feature is extracted from a fragment of audio with 4 seconds in length.

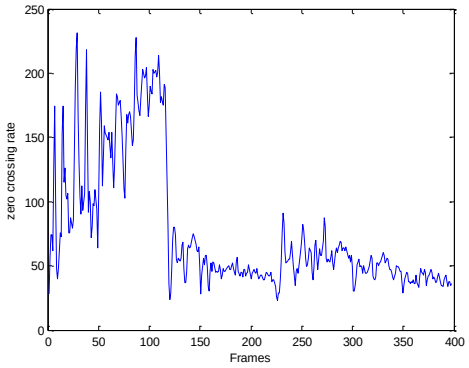

Figure 3: Zero crossing rate of a gunshot signal of 4s long

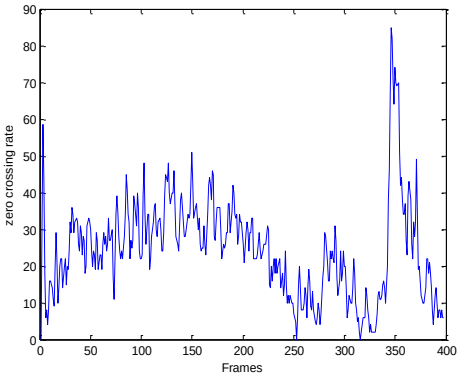

Figure 4: Zero crossing rate of a noise signal of 4s longs

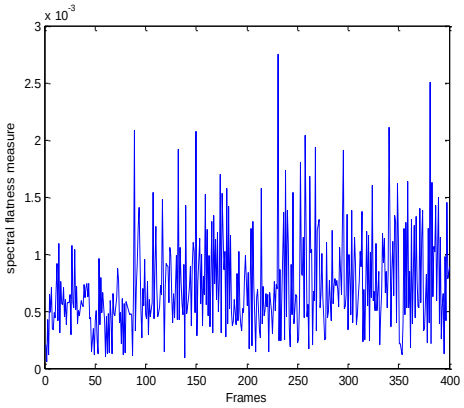

Figure 5: Spectral flatness measure of a gunshot signal of 4s long

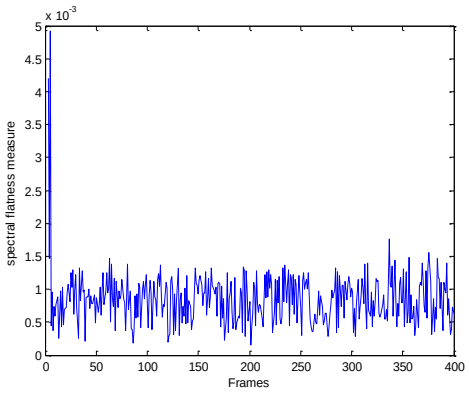

Figure 6: Spectral flatness measure of a scream signal of 4s long

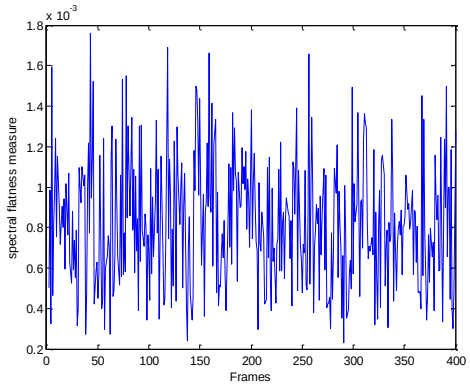

Figure 7: Spectral flatness measure of a noise signal of 4s long

3.2. Classification

This audio event detection approach is designed with two parallel GMMs for discriminating gunshot or noise and scream or noise. For each class, a GMM is built upon different feature sets. Feature vectors used for each GMM are listed in Table 1. Each binary classifier is trained with the samples of its respective classes. The parameters of the models are then estimated using the traditional Expectation-Maximization algorithm.

At the testing phase of this method, each frame from the input audio stream is firstly classified independently. The decision that an event is gun-shot, scream or noise is finally taken from computing the logical operation (OR) of the two classifiers.

Table 1: Feature vectors employed in this system

No	Gunshot/Noise classifier	Scream/Noise classifier
1	Spectral centriod	SFM
2	MFCC 1	Spectral Decrease
3	MFCC 2	MFCC 2
4	MFCC 3	MFCC 3
5	MFCC 11	MFCC 9
6	MFCC 28	MFCC 12
7	MFCC 29	
8	MFCC 30	
9	ZCR	
10	SFM	

4. Experimental Study

4.1. Data

The classification of gunshot, scream and noise were tested on audio from www.findsounds.com, www.waveplanet.com, www.pacdv.com and other internet repositories. File format includes (.wav) as well as (.au). Gunshot/noise classifier is trained with 25 gunshot files and 1 noise file. The gunshot sounds are composed of pistol, rifle, ak47, ak1, auto rifle and machine gun with varying length. This GMM classifier uses only one file of noise sound with 13s long. To train scream/noise GMM classifier, 1 scream file of 6s long and 1 noise file of 6s are used. The length of gunshot, scream and noise signals for training ranges from 1s to 13s. The system is also tested with the signal files with sampling frequency of 22k Hz. 13 experiments were conducted for testing stage. Some tested audio files include one type of audio such as gunshot only, scream only and noise only. The other contained mixed audio with different combinations of scream, gunshot and noise sounds. Scream audio files used in both training and testing include scream sounds from man, woman and children. The noise sound files collected for this system were reported to be recorded at the public places.

4.2. Performance Criteria

The efficiency of this system is measured with the performance criteria given in Table 2. The ratio of correctly identified gunshot segments over the total number of gunshot segments is defined as true gunshot whereas the gunshot segments which are wrongly labeled as scream or noise is defined as false gunshot. The ratio of correctly identified scream segments over the total number of scream segments is defined as true scream whereas the scream segments which are wrongly labeled as gunshot or noise is named as false scream. The ratio of correctly identified noise segments over the total number of noise segments is defined as true noise whereas the noise segments which are wrongly labeled as gunshot and scream are defined as false scream. In this experimental study, the accuracy of an event is measured over a segment of 1s.

Table 2: Performance Metric

Type	Gunshot	Scream	Noise
Gunshot	True Gunshot (TG)	False Gunshot (FG)	False Gunshot (FG)
Scream	False Scream (FS)	True Scream (TS)	False Scream (FS)
Noise	False Noise(FN)	False Noise(FN)	True Noise(TN)

TG = gunshot segments classified as gunshot
TS = scream segments classified as scream
TN = noise segments classified as noise
FG = gunshot segments classified as noise or scream
FS = gunshot segments classified as noise or scream
FN = noise segments classified as scream or gunshot

4.3. Experimental Result

Firstly, the detection accuracy of this audio event classification approach is tested with 11 audio files. The classification accuracy on the 11 tests is shown in Table 3. Test number 1, 2 and 3 are gunshot only audio file with the length of 2s, 11s and 5s respectively. Types of gunshot in these tests are machine gun 16, machine gun 60 and pistol. Test number 4, 5 and 6 contain scream sounds of children, woman, and man with the length of 2s, 3s and 13s respectively. Test number 7, 8 and 9 are public noises with 13s, 12s and 7s long noise signals. Test number 10 is the mixture of audio with gunshot and noise of 23s consisting of m 16 gunshots and a people talk as noise. Experiment 11 is also tested with mixed types of audio with the total length of 31s. In this test, public noise of 13s, woman scream of 13s and shot gun signal of 5s are contained.

Table 3: Classification Accuracy

Test No	TG	TS	TN	FG	FS	FN	Accuracy
1	100%	-	-	-	-	-	100%
2	100%	-	-	-	-	-	100%
3	100%	-	-	-	-	-	100%
4	-	100%	-	-	-	-	100%
5	-	100%	-	-	-	-	100%
6	-	100%	-	-	-	-	100%
7	-	-	100%	-	-	-	100%
8	-	-	100%	-	-	-	100%
9	-	-	72%	-	-	28%	72%
10	100%	-	100%	-	-	-	100%
11	100%	100%	100%	-	-	-	100%
Avg	100%	100%	94%	-	-	28%	97%

The effects of signal-to-noise ratio (SNR) on the performance of the proposed method are tested with two noisy audio files at 10dB and 5dB. This experimental study is studied without adding noise to the existing training data. Audio signals of gunshot and scream were mixed with the noise signal at specified SNR. The length of test signal in this experiment is 22s. The detection accuracy for this test is 80% for gunshot detection, 100% for noise and scream detection at 10dB SNR. However, the classification accuracy decreases to 20% for true gun, 100% for true scream and 100% for true noise at 5dB SNR. This test illustrates that the method can sufficient detect for noisy situation. However, to be robust for very noisy situations (e.g. 5dB SNR) two GMM classifiers may need to train with noisy data.

5. Conclusion

An approach is presented for a classification system which is able to detect events such as gunshots, screams and noises. Audio features are extracted from both time domain and frequency domain. Two simultaneous GMM classifiers classify the audio event into gunshot/noise and scream/noise. Different features vectors have been applied for two GMM classifiers for better performance of the system. The effectiveness of the system is tested on audio clips containing varying durations of scream, gunshot and noise sounds. Experimental results reports that the overall accuracy of the method is about 97% with false rejection rate around 28%.

6. References

[1] P. K. Atrey, N. C. Maddage and M. S. Kankanhalli, "Audio Based Event Detection for Multimedia Surveillance", Acoustics, Speech, and Signal Processing, 2006. ICASSP, 2006. *IEEE International Conference on, 2006.*

[2] J. L. Rouas, J. Louradour, and S. Ambellouis, "Audio Events Detection in Public Transport Vehicle", Proc. of the 9th International *IEEE Conference on Intelligent Transportation Systems: 733-738, 2006*.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events Detection for an Audio-Based Surveillance System", Multimedia and Expo, 2005, ICME 2005, 1306-1309, *IEEE International Conference.*

[4] G. Peeters, "A large set of audio features for sound description" (similarity and classification) in the CUIDADO Project Report, 2004.

[5] M. A. F. Figueiredo and A. K. Jain. "Unsupervised Learning of Finite Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (3):381-396, 2002.*

[6] L. Lu, H.J. Zhaung, and H. Jiang, "Content analysis for audio classification and segmentation" Speech and Audio Processing, *IEEE Transactions on, 10(7):504-516, 2002.*

[7] J. Pinquier, J.L. Rouas, and R. Andre-Obrecht " Robust Speech/Music Classification in Audio Documents." International Conference on Spoken Language Processing, ICSLP, 5: 10-15, 2002.

[8] T. Zhang and C.C.J. Kuo. " Hierarchical system for content based audio classification and retrieval."

Conference on Multimedia Storage and Archiving Systems
III, SPIE, 3527:398-409, 1998.