Acoustic Events Classification Using Support Vector Machines (SVMs)

Htat Htat Aung, Daw Hlaing Thida Oo
University of Computer Studies, Yangon
htathtataung92@gmail.com

Abstract

In this paper, an approach is built to automatically detect acoustic events that are produced in a meeting or lecture room environment. Six audio classes are to be classified through this approach. The classes considered are music, speech, clapping, door slam, cough, and laughter. Several events samples are collected from the Internet. Support Vector Machines (SVMs) perform training and testing the events classification on perceptual and MFCC features set. A hierarchical clustering scheme is used therefore the required number of binary SVM classifier is also reduced. The system is tested on different data sets and its effectiveness is determined with classification accuracy on audio event frames.

1. Introduction

Acoustic event detection (AED) is a part of computational auditory scene analysis may be used to detect and identify acoustic events. AED task is closely related to the more general task of noise classification and recognition. Acoustic event detection is providing a lot of advantages in surveillance and security applications. Video sensors used to get the information but due to the occurrence of any situation when information is unreliable, unavailable in the darkness and expensive. However audio sensors are simple and cheap so that audio information becomes important cues in many diverse areas.

Meeting recording is reflected in a rich variety of acoustic events, either produced by the human or objects. Speech is the most informative acoustic event, but other kind of sounds may provide helpful information in environments. In this implementation, a system which is able to detect six types of audio events; clapping, door slam, laughter, speech and music in the meeting room environment is developed for the Acoustic Event Classification task.

Acoustic event detection and classification (AED/AEC) may help to detect and describe the human and social activity that takes place in the meeting room for example;

- clapping or laughter inside a speech discourse
- a cough in the middle of meeting
- a clapping at the end of the meeting
- door noise when the meeting has just started.

This information is useful in applications such as event monitoring systems, multimedia information retrieval and intelligent meeting or lecture rooms.

2. Related Work

In [1], the author compared two different approaches to alarm sound detection and classification, namely: ANN and a technique specifically designed to exploit the structure of alarm sounds and minimize the influence of background noise. The authors compared to the task of non-speech environmental sound recognition in [2]. The Learning Vector Quantization (LVQ) and ANN have been used. Bird species sound recognition has been performed in [3]. The acoustic event recognition for four different environments - kitchen, workshop, office and outdoors has been applied in [4]. The paper discusses a prototype of a sound recognition system focused on an ultra low power hardware implementation in a button-like miniature form. In [5], the authors have considered the detection of “laughter” in meetings with SVM. In their experiments, MFCC features outperform the proposed spatial features and modulation spectrum features. The authors have considered human activity detection in public places mainly by concentrating on coffee shop activity detection in [6]. A wide range of features and two distinct classifiers (k-nearest neighbors and GMM) have been compared. A system of non-speech environmental sound classification for autonomous surveillance has been discussed in [7]. Features based on a wavelet transformation and MFCC features performed the best. In [8], the authors have applied two classification techniques (SVM and GMM) to audio indexing. They have performed a discrimination of “speech” and “music” in radio programs. The system analyzes the acoustic activity at the recording site, in [9] and using a set of low-level acoustic features the system is able to separate all interesting events in an unsupervised manner in office environment. Recognition of sounds related to the bathroom environment has been done in [10] and an HMM classifier and MFCC features have been used. Preliminary results showed high average accuracy. The work in [11] presents a hierarchical approach of audio based event detection

for surveillance. A given audio frame is firstly classified as vocal or non-vocal, and then further classified as normal and excited. The approach is also based on a GMM classifier and LPC features.

3. Audio Feature Representation

The first task of audio feature extraction is down sampling the input signal from various sampling rates to 8 kHz, mono channel .wav format with 16 bit resolution. Frames are of size 128 samples (16 ms) with 50% (64 samples or 8 ms) overlap in each of the two adjacent frames. Next, a frame is hamming-windowed. Two types of features are computed from each frame: (i) perceptual features composed of zero crossing rate (ZCR), short-time energy (STE), spectral centroid (CE), spectral roll-off (RF), spectral bandwidth (BW) and (ii) mel-frequency cepstral coefficients (MFCCs). These statistics are considered as feature sets for the audio sound to be determined.

3.1. Perceptual Features

3.1.1. Zero Crossing Rate (ZCR)

It measures the number of zero crossings of the waveform within a frame and is calculated as:

$$ZCR = \sum_{n=0}^{N-1} |s(n) - s(n-1)| \qquad (1)$$

where s*(n)* is sign of the signal value at the time index *n* and *N* is frame length.

3.1.2. Short Time Energy (STE)

Total signal energy in a frame calculated as:

$$STE = \sum_{n=0}^{N-1} s(n)^2 \qquad (2)$$

where *s(n)* is signal value at the time index *n* and *N* is frame length.

3.1.3. Spectral Centroid (CE)

The centroid is a measure of the spectral "brightness" of the spectral frame and is defined as the linear average frequency weighted by Discrete Fourier Transform (DFT) amplitudes, divided by the sum of the amplitudes:

$$CE = \frac{\sum_{\forall i} f(i)\, a(i)}{\sum_{\forall i} a(i)} \qquad (3)$$

where *f(i)* is the frequency value at the frequency *i* and *a(i)* is DFT amplitude.

3.1.4. Spectral Roll-off (RF)

It is a measure of the skewness of the spectral shape and is defined as a frequency *f* below which the c percentage of the spectral amplitudes is concentrated (here *c*=95):

$$RF = \frac{c}{100} \sum_{\forall i} a(i) \qquad (4)$$

where *a(i)* is amplitude of DFT.

3.1.4. Spectral Bandwidth (BW)

A measure of spreading of the spectrum around the spectral centroid:

$$BW = \sqrt{\frac{\sum_{\forall i} (f(i) - CE)^2 a(i)^2}{\sum_{\forall i} a(i)^2}} \qquad (5)$$

where *f(i)* is the frequency value at the frequency *i* and *a(i)* is DFT amplitude.

3.2 Mel-frequency Cepstral Coefficients(MFCC)

MFCC are short-terms spectral features and have been used very successfully in the field of speech recognition as classification features for speech audio signals. The processing sequence for finding the MFCCs of an audio signal is following:

- Window the data with a Hamming Window
- Find the amplitude values of the DFT of the data
- Convert the amplitude values of filter bank outputs
- Calculate the log base 10
- Find the discrete cosine transform

4. Support Vector Machines (SVMs)

SVM is a supervised learning machine and outperforms many popular methods for text classification. It is widely used in pattern recognition areas such as auditory context recognition, face detection, isolated handwriting digit recognition, and pattern recognition and gene classification. SVM classifier that discriminated the data by creating boundaries between classes rather than estimating class conditional densities, it may need considerably less data to perform accurate classification. SVM are fundamentally binary classifiers, but any number of classes can be accommodated by first considering linearly separable classes, i.e., two classes which can be perfectly separated using a linear hyper plane as a decision boundary. SVM training is based on the idea of maximizing the margin between any decision boundary and the closest observation at each side of the hyper plane, i.e., the goal is to maximize the

distance from the closest class representative points to the decision boundary. In nonlinearly separable case, the SVM replaces the inner product $x \,.y$ by a kernel function $K(x \,.\, y)$ implicitly maps the input vectors into a high dimensional feature space. The most often used kernel functions in SVM applications are the following:

i. Radial Basis Function (RBF)

$$K(x, y) = \exp(-\| x - y \|^2 / 2\sigma^2) \quad (6)$$

ii. Polynomial

$$K(x, y) = (x \,.\, y + 1)^d \quad (7)$$

iii. Multi-layer perception

$$K(x, y) = \tanh(\kappa (x \,.\, y) - \mu) \quad (8)$$

Where σ, d, κ and μ are kernel parameters. This method uses Radial Basis Function (RBF) kernel, because it was empirically observed to perform better than other two.

In this paper, we have implemented multi-class classification task is done by a hierarchical classification structure so that number of SVMs to be used can be reduced to N-1 instead of $N(N-1)/2$ classifiers which is usually needed for classifying N classes.

5. Proposed Method

Based on the cepstral features and perceptual features, the proposed system is designed as depicted in Figure 1. Exploiting both perceptual and cepstral features described in Section 4, the overall system is built.

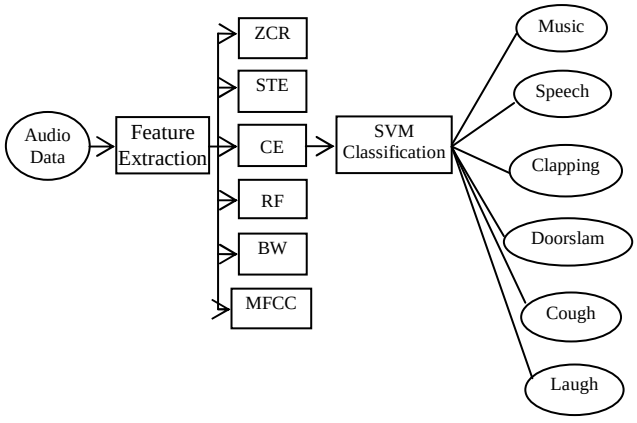

Figure 1. Structure of Acoustic Event Detection

5.1 Classification Architecture

A hierarchically clustering scheme with a SVM at each node of this acoustic event detection system is employed as the classification scheme. In order to detect the above nominated six classes, the classification is built by five SVMs as illustrated in Figure 2. In this classification hierarchy, three levels classification is adopted. First, SVM 1 classifies music/speech or clapping /door slam/cough/laughter. In the second level, there are two classification nodes: SVM 2 and SVM 3. At this level music or speech classification is performed by SVM 2 while differentiation between clapping/door slam or cough/ laughter is accomplished by SVM 3. Finally, SVM 4 and SVM 5 operate for discrimination between clapping or door slam and cough or laughter.

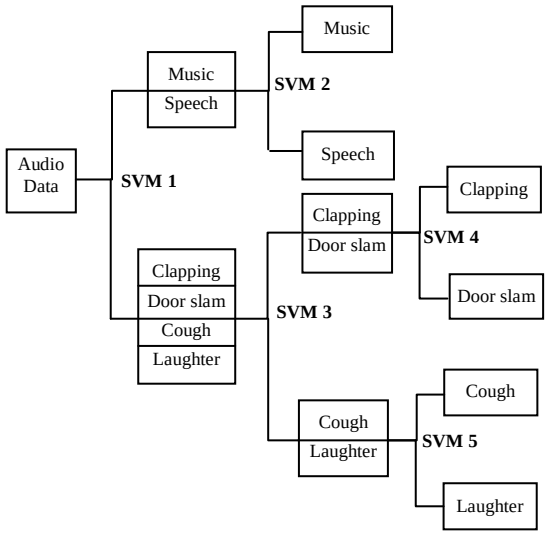

Figure 2. Architecture of SVM classification

In Table 1, the analysis of feature representation suitable for all nodes of SVM classification is shown. In this table, average discriminating accuracy using several features set for each SVM is listed. It could be noticed that for most of the classification nodes, high accuracy can be obtained with combination of MFCC and perceptual features rather than using one type of feature.

Table 1. Experiment result on several features set

Classifiers	MFCC	Perceptual	MFCC+Per:
SVM1	96.35±0.50	90.00±1.00	93.37
SVM2	99.35±0.05	79.35±0.05	99.40
SVM3	87.20±0.20	82.35±0.35	90.73±0.32
SVM4	98.50	95.75±0.15	98.45±0.35
SVM5	88.00±1.00	89.50±0.50	91.70±0.10

6. Experimental Study

The audio files used in experiment are collected from various repositories including following:

1. http://www.dcs.shef.ac.uk/spandh/projects/shatr
2. http://www.partnershnrthyme.com/soundfx/human.shtml
3. http://www.archive.org/search.php
4. http://www.mit.edu

Table 2. Duration and Accuracy of the Training and Testing Data

Type	Training		Testing1		Testing2		Testing3		Testing4		Testing5	
	Duration	Accuracy	Duration	Accuracy	Duration	Accuracy	Duration	Accuracy	Duration	Accuracy	Duration	Accuracy
Music	12s	86.50%	60s	71.71%	104s	63.91%	60s	70.11%	119s	86.79%	120s	86.73%
Speech	20s	96.00%	173s	53.40%	360s	69.28%	638s	89.82%	813s	81.79%	1055s	96.11%
Clapping	4s	100%	7s	100%	14s	99.37%	18s	99.96%	21s	100%	21s	100%
Door slam	4s	80.60%	_	_	_	_	_	_	1s	77.60%	2s	62.4%
Cough	4s	73.00%	_	_	2s	69.6%	_	_	2s	100%	3s	69.6%
Laughter	4s	95.40%	_	_	_	_	4s	99.20%	4s	99.20%	_	_

The speech audio files are selected from lecture room recording where language of media is English. Most of the speakers are males. The database used in our experiments is composed of 3648 seconds in total length including training and testing in the meeting room or lecture room. Events are classified in 1s clip and each clip is labeled as one of the pre-defined six audio classes. It is partitioned into training set of about 48 seconds and five test sets of about 3600 seconds. In Table 1, data compositions of train and test sets are listed along with their duration. These test data sets are created by concatenating the specific events with varying durations where cough and laughter events are drawn from training samples.

In experiments, all SVMs are learned with RBF kernel. In creating training data, it is designed with segment of 12 s music, 20 s long speech segment, and clapping, door slam, cough, laughter clips each of which are 4 s in length. However, in all testing data sets, number audio events and their duration contained are at random and only the total length in each class are described Table 2. According to the experimental results, it is found that the proposed approach can obtain highest accuracy at clapping event in all test sets.

Table 3 also shows the overall accuracy with corresponding total duration used in training and testing. The average accuracy for training set is 88.58% and that of all test sets is 83.79%.

Table 3. Accuracy of the training and testing

Acoustic Events	Training set		Testing set	
	Duration	Accuracy	Duration	Accuracy
Music	12s	86.50%	463s	75.85%
Speech	20s	96.00%	3039s	78.08%
Clapping	4s	100%	81s	99.87%
Doorslam	4s	80.60%	3s	70%
Cough	4s	73.00%	7s	79.73%
Laughter	4s	95.40%	8s	99.20%

In Figure 3, the resulted accuracy for training and testing are illustrated. Overall error rates for both training and testing are also shown in Figure 4. Through the tables and figures, it is obvious that the proposed method achieve reasonable accuracy rates at all events. It is expected that higher accuracy could be obtained with the expense of more training data.

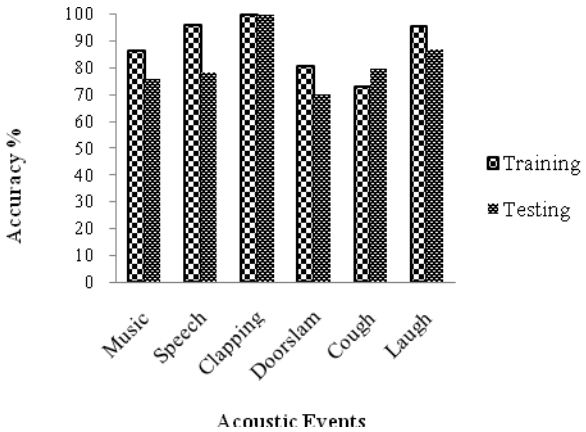

Figure 3. Overall accuracy of training and testing

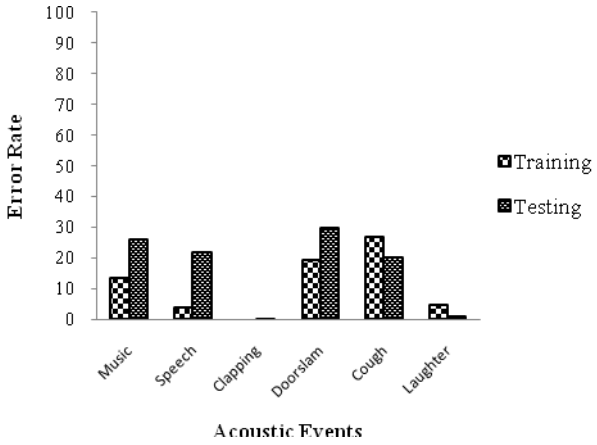

Figure 4. Overall error rate of training and testing

7. Discussion

When SVM based method is used for acoustic event detection and classification, it is shown that it has good performance in audio classification. SVM provide efficiency, illustrating it gives generalized ability to classify events unseen in the training set. SVM takes long time to train and needs to select kernel function and labels them which are practiced by trial and error. Combination of perceptual features and cepstral features can enhance the performance in detecting the events rather using only one of them. Thus, in all classifiers, both perceptual and cespstral features are applied. In the experiments, data are collected from many sources. Using training and testing samples drawn from a unique environment can enhance the overall accuracy of the system.

8. Conclusion

This implementation attempts to deal with the problem of classifying acoustic events in a meeting/lecture room environments. To obtain the best performance, the system analyzed several feature sets and SVM classification on data sets. The best results could be achieved with features that is the combination of MFCC and Perceptual Features. As the system uses a hierarchical clustering scheme, the required number of classifier is reduced. Experimental study conducted on 5 data sets indicates that this approach is suitable for classifying audio events. Classification of more acoustic events could be encountered in the domain environment with more balanced classification architecture is one possible direction of future work.

9. References

[1] D. Ellis, "Detecting Alarm Sounds", in *Proc. Workshop on Consistent and Reliable Acoustic Cues*, 2001.

[2] M. Cowling, R. Sitte, "Analysis of speech Recognition Techniques for use in a Non-Speech Sound Recognition System", in *Proc. International Symposium on Digital Signal Processing for Communication Systems*, 2002.

[3] A.Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] M. Stäger, P. Lukowicz, N. Perera, T. von Büren, G. Tröster, T. Starner, "Sound Button: Design of a Low Power Wearable Audio Classification System", in *Proc. IEEE International Symposium on Wearable Computers*, pp. 12-17, 2003.

[5]L. Kennedy, D. Ellis, "Laughter Detection in Meetings", NIST Meeting Recognition Workshop, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[6] B. Lukic, *Activity Detection in Public Spaces*, M.Sc. Thesis, Royal Institute of Technology,
Stockholm, Sweden, 2004.

[7] M. Cowling, *Non-Speech Environmental Sound Classification System for Autonomous Surveillance*, PhD Thesis, Griffith University, Australia, 2004.

[8] J. Arias, J. Pinquier, R. André-Obrecht, "Evaluation of classification techniques for audio indexing", in *Proc. European Signal Processing Conference*, 2005.

[9] A. Härmä, M. McKinney, J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment", in *Proc. International Conference on Multimedia and Expo*, 2005.

[10] C. Jianfeng, Z. Jianmin, A. Kam, L. Shue, "An automatic acoustic bathroom monitoring system", in *Proc. IEEE International Symposium on Circuits and Systems*, 2005.

[11] P. Atrey, N. Maddage, M. Kankanhalli, "Audio Based Event Detection for Multimedia Surveillance", in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.