SEMANTIC CLASSIFICATION AND RETRIEVAL SYSTEM FOR ENVIRONMENTAL SOUNDS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇİĞDEM OKUYUCU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2012

Approval of the thesis:

SEMANTIC CLASSIFICATION AND RETRIEVAL SYSTEM FOR ENVIRONMENTAL SOUNDS

submitted by **ÇİĞDEM OKUYUCU** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** _____

Prof. Dr. Adnan Yazıcı
Supervisor, **Computer Engineering Dept., METU** _____

Assist. Prof. Dr. Mustafa Sert
Co-supervisor, **Computer Engineering Dept., Başkent University** _____

Examining Committee Members:

Assist. Prof. Dr. Murat Koyuncu
Information Systems Engineering Dept., Atılım University _____

Prof. Dr. Adnan Yazıcı
Computer Engineering Dept., METU _____

Assist. Prof. Dr. Mustafa Sert
Computer Engineering Dept., Başkent University _____

Assist. Prof. Dr. Sinan Kalkan
Computer Engineering Dept., METU _____

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Dept., METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ÇİĞDEM OKUYUCU

Signature :

ABSTRACT

SEMANTIC CLASSIFICATION AND RETRIEVAL SYSTEM FOR ENVIRONMENTAL SOUNDS

Okuyucu, Çiğdem

M.Sc., Department of Computer Engineering

Supervisor : Prof. Dr. Adnan Yazıcı

Co-Supervisor : Assist. Prof. Dr. Mustafa Sert

September 2012, 99 pages

The growth of multimedia content in recent years motivated the research on audio classification and content retrieval area. In this thesis, a general environmental audio classification and retrieval approach is proposed in which higher level *semantic classes* (outdoor, nature, meeting and violence) are obtained from lower level *acoustic classes* (emergency alarm, car horn, gun-shot, explosion, automobile, motorcycle, helicopter, wind, water, rain, applause, crowd and laughter). In order to classify an audio sample into *acoustic classes*, MPEG-7 audio features, Mel Frequency Cepstral Coefficients (MFCC) feature and Zero Crossing Rate (ZCR) feature are used with Hidden Markov Model (HMM) and Support Vector Machine (SVM) classifiers. Additionally, a new classification method is proposed using Genetic Algorithm (GA) for classification of *semantic classes*. Query by Example (QBE) and keyword-based query capabilities are implemented for content retrieval.

Keywords: Hidden Markov Model (HMM), Support Vector Machines (SVM), MPEG-7 Audio Features, Zero Crossing Rate (ZCR), Mel Frequency Cepstral Coefficients (MFCC), Genetic Algorithm (GA), Query by Example (QBE)

ÖZ

ÇEVRESEL SESLER İÇİN ANLAMSAL SINIFLANDIRMA VE GERİ ERİŞİM SİSTEMİ

Okuyucu, Çiğdem
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Adnan Yazıcı
Ortak Tez Yöneticisi : Y. Doç. Dr. Mustafa Sert
Eylül 2012, 99 sayfa

Çoklu ortam içeriğindeki artış son yıllarda ses sınıflandırma ve geri erişim sistemleri üzerine çalışmaları motive etmektedir. Bu çalışmada, düşük seviyedeki akustik sınıflar (acil durum alarmları, araba kornası, silah, patlama, otomobil, motosiklet, helikopter, rüzgar, su, yağmur, alkış, kalabalık ve gülme sesi) kullanılarak yüksek seviyedeki anlamsal sınıfların (dış ortam, doğa, şiddet ve toplantı) elde edildiği genel bir çevresel ses sınıflandırma ve geri erişim yaklaşımı önerilmişir. Bir ses örneğini sınıflandırmak için MPEG-7 ses öznitelikleri, Mel Frekansı Kepstrum Katsayıları (MFCC) özniteliği ve Sıfırı Geçme Oranı (ZCR) özniteliği; Saklı Markov Modelleri (HMM) ve Destek Vektör Makineleri (SVM) sınıflandırıcıları üzerinde kullanılmıştır. Bunlara ek olarak, anlamsal sınıfların elde edilmesi için Genetik Algoritma (GA) kullanılarak yeni bir sınıflandırma metodu önerilmişir. İçerik geri erişimi için anahtar kelime ve ses örneğiyle sorgulama (QBE) yetenekleri geliştirilmişir.

Anahtar Kelimeler: Saklı Markov Modeli (HMM), Destek Vektör Makineleri (SVM), MPEG-7 Ses öznitelikleri, Sıfır Geçme Oranı (ZCR), Mel Frekansı Kepstrum Katsayıları (MFCC), Genetik Algoritma (GA)

To Life

ACKNOWLEDGMENTS

I express my sincere appreciation to my supervisor, Prof. Dr. Adnan Yazıcı for his guidance, criticism, wisdom, encouragement and insight throughout the research. I am very grateful to my co-supervisor, Assist. Prof. Dr. Mustafa Sert for his advice, motivation, guidance and support through his previous works and experience.

I am grateful to my parents and my sister for their love and support. I thank to, my dear, Çağrı İlçe for his help, support and nice surprises. I hope there will be a big change in our lives after this thesis work.

I thank to my dear friends İpek Tatlı, Hande Dağlı, N. İlker Erçin, Kerem Hadımlı, Utku Erdoğdu and Merve Aydınlılar for their unlimited emotional and technical support.

Thanks to Melike Bozkaya for energy to keep me motivated as much as possible. I am also very thankful to Hakan Yıldırım, Marc Bartels and Fatih Arslan for their creative ideas, advices and comments.

I am also thankful for my colleagues from Philips Healthcare, Danny Havenith, Pim Otto and Frank Verburg for their understanding and support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Due to the continuous growth in multimedia content in digital archives, the problem of categorization and retrieval of these archives is an emerging research area. For instance, records of broadcast news, sports videos, television shows, radio programs, films and medical surveys are increasing rapidly. Auditory and visual components of these records are employed to categorize, using video, image and audio classification techniques. In order to gather information from the categorized records, content retrieval techniques are employed.

The initial step for audio classification and content retrieval research is audio classification which is basically labelling an audio sample as belonging to a single category from a set of predefined categories. For instance, an audio sample from a football game record can be classified into speech and non-speech categories. The first problem of audio classification is deciding on the intended categories. In the early studies of audio classification, an audio sample is classified into general categories such as speech, music and environmental sound [4, 5]. Speech signals and music signals have repetitive pattern characteristics and tonal characteristics, respectively. However, not following an obvious pattern might be counted as the characteristics of environmental sound signals. Therefore, mentioned characteristics of speech, music and environmental sounds allow such a general categorization approach.

This general categorization can be seen as a key approach in order to obtain mixed-type or more detailed hierarchical categorizations. For instance, an audio sample can be first classified into silence, speech, music and environmental categories and then speech parts can be classified into emotional categories or music parts can be classified into genre categories. For environmental sounds, the situation becomes more complicated to decide on the categories because they do not have an obvious pattern. For example, applause and rain sounds should

be treated as different categories logically, but they are similar for human hearing and difficult to distinguish.

Several researches have differentiating approaches for this categorization problem. In Beritelli's study [6] an audio sample is classified into categories such as bus, car, construction, dump, factory, office, pool, station, stadium and train. Muhammad [7] also proposed a categorization like restaurant, crowded street, quiet street, shopping mall, car with open window, car with closed window, corridor of university campus, office room, desert and park. Feki [8] categorized environmental sounds into speech, music, ring tones, train, motorcycle, explosion, helicopter, slamming door, dog barking, bird, breeze glasses, applause, horse, cat, care, slot machine, wind, plane, laugh and police alarm categories.

After the categorization decision problem, a second audio classification problem emerges. This problem is exploring the best method to classify an audio sample into the intended categories. In order to handle this problem, several machine learning techniques are applied with several audio features. Support Vector Machines (SVM), Hidden Markov Models (HMM), Neural Networks (NN) and Gaussian Mixture Models (GMM) are commonly applied machine learning techniques with MPEG-7 audio features, Mel Frequency Cepstral Coefficents (MFCC) feature and Zero Crossing Rate (ZCR) feature. Existing studies propose solutions to this problem focusing on analysing the best representative audio features [7, 9–11], discovering the best machine learning methods [6, 12, 13] and discovering the best machine learning and feature combination [8, 14].

In this thesis study, different combinations of MPEG-7 audio features, MFCC feature and ZCR feature are employed on SVM and HMM classifiers in order to obtain the best representative feature set and machine learning technique combination. An environmental audio clip is classified into thirteen categories, namely *acoustic classes* such as emergency alarm, car horn, gun-shot, explosion, automobile, motorcycle, helicopter, wind, water, rain, applause, crowd and laughter. These categories containing considerably similar sounds are intentionally selected in order to experiment the performance of the selected features and classifiers.

In the proposed system, an audio clip is first divided into one-second segments. Silence segments are detected and non-silence segments are classified into selected *acoustic classes*. This sample audio clip containing the classified segments, is then passed through a smoothing process to discard the classification errors. After silence detection, classification and smoothing

processes; labelled audio segments are used as input for the proposed semantic classification in order to be classified into outdoor, nature, violence and meeting classes. A table containing the impact values of each *acoustic class* on each *semantic class* is optimized using Genetic Algorithm (GA) for this classification.

Content retrieval capabilities are also implemented. Temporal and keyword-based queries are supported and related audio segments are retrieved. Queries by example audio are provided to search for similar audio segments.

A new environmental audio classification and retrieval tool is implemented to the present the proposed work. This tool provides a user interface to classify an audio clip and retrieve content information from the results.

The work presented in this thesis contributes to the previously conducted studies in the following aspects:

- A general solution for classification and segmentation of environmental sounds is proposed. This solution utilizes a bottom up approach in which underlying basic categories are processed in order to obtain more general and complex categories. Higher level *Semantic classes* are obtained utilizing lower level *acoustic classes*.
- Considerably similar categories (gun-shot, explosion; motorcycle, helicopter and automobile; rain, wind, water categories) which human perception is even insufficient to distinguish are intentionally selected and experimented in order to explore the performances of the selected features with SVM and HMM classifiers.
- A comparative evaluation and analysis is presented for SVM and HMM classification performance. Additionally, a comprehensive study is presented for MPEG-7, MFCC and ZCR audio features to discover the best representative feature combination.
- A new environmental audio classification and retrieval tool is implemented. This tool provides an efficient environment for the user in order to classify audio samples and retrieve the related content.

The rest of the thesis is organized a follows: In Chapter 2, related studies are mentioned and summarized. Necessary background information about MPEG-7, MFCC and ZCR features, brief explanations of HMM and SVM classifiers and GA are explained in Chapter 3. In

Chapter 4, the proposed system is explained in detail. Experimental results and interpretations are given in Chapter 5. Chapter 6 is about the implementation and utilization of the proposed classification tool. Finally the conclusions and future work are provided in Chapter 7.

CHAPTER 2

LITERATURE SURVEY

In this chapter, previous studies about audio classification are deeply explained in related sections. The studies concentrate in audio feature and classifier selection. There are also studies introducing new techniques for both feature extraction and classification subjects.

2.1 General Sound Classification

Xiong et al. proposed a comparison-based study for sports audio classification [15]. Audio is classified into applause, ball-hit, cheering, music, speech and speech with music classes. They compared two different HMMs which are Maximum Likelihood HMM (ML-HMM) and Entropic Prior HMM (EP-HMM) with and without trimming of the model parameters. MFCC and MPEG-7 audio features are used in the experiments. Regarding their experiment results, combination of MPEG-7 and EP-HMM with trimming achieves a classification accuracy of 94.7%. Second best result is the combination of MFCC and ML-HMM with a classification accuracy of 94.6%. It is also stated that all combinations provides an average classification accuracy around of 90%.

Kim et al. introduces an MPEG-7 based audio classification technique for analysis of film material [16]. Two recognition systems are offered in their study which are speaker recognition and sound effect recognition. They experimented HMMs as classifiers using MPEG-7 ASP feature based on Audio Spectrum Basis (ASB). They conducted various tests and come up with the result that usage of Independent Component Analysis (ICA) is providing better results than Normalized Audio Spectrum Envelope (NASE) and Principal Component Analysis (PCA) in the speaker recognition system. An accuracy of 96% is achieved in the sound

effect recognition experiments in real time.

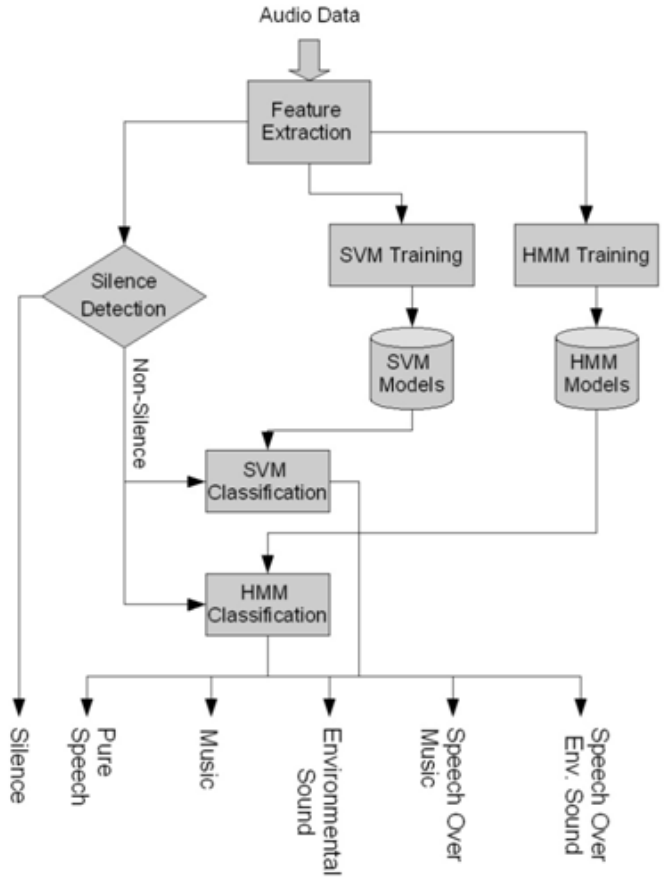

Figure 2.1: Block diagram of proposed classification system of Doğan [1].

Doğan et al. presents a complete content-based audio management and retrieval system for news broadcasts [1]. This system considers classification, segmentation, analysis and retrieval of an audio stream. An audio stream is segmented into six different classes: silence, pure speech, music, environmental sound, speech over music and speech over environmental sound (Figure 2.1). In addition, various audio classification experiments are presented to exploit the ability of MPEG-7 features and the selected classification methods (SVM and HMM). The proposed system is composed of silence detection, classification of non-silent frames and smoothing steps. Classification of mixed type audio data (for instance, speech over music and speech over environmental sound) using the combination of MPEG-7 ASC, ASS and ASF features achieves considerably high accuracy rates in news domain. The combined feature has accuracy of 91.9% with HMM and 90.5% with SVM in classification of non-speech and

with-speech classes. For pure speech and mixed speech classification, accuracies of 89.4% and 86.0% are achieved with SVM and HMM classifiers, respectively.

Song et al. proposed a feature extraction and audio classification method in which audio is analysed to short-time energy ratio, ZCR, bandwidth, low short-time energy ratio, high zero crossing rate ratio and noise rate features [17]. In their study, they introduce a new audio classification technique for news audio based on decision tree method. Proposed classification is applied on four classes: silence, pure speech, music and non-pure speech. Their proposed classification method and selected features result into reasonable accuracy levels around 90%.

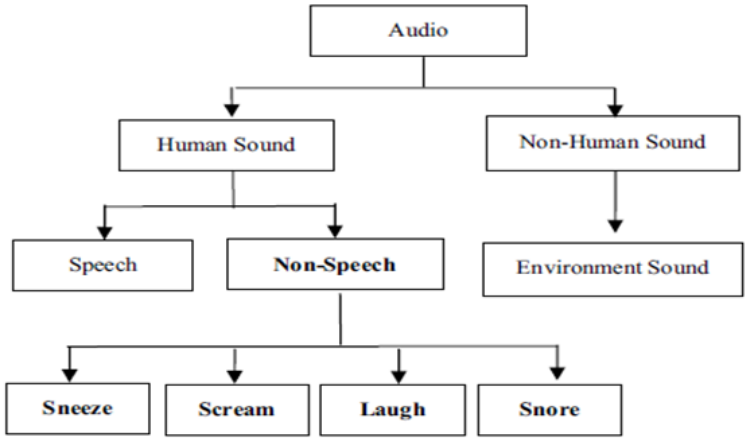

Figure 2.2: Diagram of proposed classification tree of Liao [2].

Liao et al. studied on a classification system in which audio is classified into human and non-human sounds as a first step [2]. In the second step human sounds are classified into speech and non-speech and then non-speech human sounds are classified into sneeze, scream, laugh and snore sounds (Figure 2.2). For their experiments, they used two sets of acoustic features. First set is composed of commonly used audio descriptors which are fundamental frequency, spectral centroid (SC), spectral spread spectral flatness, entropy and two format frequencies. Second set contains 20 MFCC features after applying Discrete Cosine Transform (DCT) on 20 log-energy values. They used Multivariate Adaptive Regression Spline (MARS) to build the optimal model and select the most representative feature. They used SVM classifier for feature selection and noise sensitivity experiments. The experimental results show that spectral centroid, fundamental frequency, spectral spread and spectral flatness play important roles in the classification task. In addition, these four feature also achieve better noise insensitivity

when the audio signals were in noisy environments. In order to compare the results, MFCC based experiments are conducted. The features selected by MARS indicates that low frequency components play significant role for this particular classification problem.

Lu et al. proposed an audio classification method which detects cheering events in audio extracted from videos of live sports games [18]. Four audio classes are studied: speech, music, cheering and other. Experiments are conducted in audio streams of beach volleyball, badminton, ping-pong, volleyball and hockey games. Fixed-length sliding window technique is used for pre-segmentation from start to end. Short-Time Energy (STE), Sub-Band Energy Distribution (SBED), Spectrum Flux (SF), Brightness and Bandwidth features are extracted in order to used with the Gaussian Mixture Model (GMM) classifier. A new smoothing algorithm called boundary-seek algorithm is proposed to overcome the shortcomings of sliding window technique. They prepared an HMM based event detection framework with the same feature set to compare with the sliding window based framework. Their system achieved an average F value of 82.99% considering five kinds of sports after integration of all approaches. In the experiments, sliding window based framework is more successful than HMM based event detection framework.

2.2 Environmental Sound Classification

Dufaux et al. proposed an automatic sound detection and recognition system for noisy environments [12]. HMM and GMM are used for classification of impulsive sounds like door slam, glass break, human scream, explosions, gun shots and other noises. Their system is composed of an impulsive sound detection module and a sound recognition module. They used a non-linear median filter to analyse the energy variations which is used in impulsive sound detection. Experimental results indicates that HMM (recognition rate is 98.54%) performs better than GMM (recognition rate is 97.32%) for the proposed recognition system.

Nishura et al. studied environmental source identification based on HMM for robust speech recognition [13]. They categorized sounds into three categories. The first category contains collision sounds of wood, plastic and ceramics. Second category contains sounds which occurs from human activities and the third category contains sounds such as coins, telephones and pipes. They proposed a new HMM composing speech HMMs and an HMM of cat-

egorized environmental sounds for robust environmental sound-added speech recognition. In experimental results it is stated that their new HMM is more successful (95.8%) than conventional-HMM (85.2%) and speech-HMM (41.2%).

Wang et al. proposed an environmental sound recognition technique using MPEG-7 audio Low-Level Descriptors (LDD) [9]. They categorized home environmental sound into seven categories including male speech, female speech, dog barks, cat mews, doorbell rings, knock and laughing. Regarding the experiments, their recognition rate is 82% if they only adopt spectrogram as the parameter. Later they improve their recognition rate about 95% by using three MPEG-7 audio Low Level Descriptors (LDD) which are ASC, ASS and ASF descriptors.

Cai et al. [10] studied the problem of highlight sound effects detection focusing on laughter, applause and cheer sound effects, which are highlight events in sports, meeting and entertainment videos. They used HMMs with MFCC, STE, ZCR, sub-band energies, brightness and bandwidth features. They combined all mentioned features in one feature vector to introduce satisfying results during the experiments. According to their experiments, the system reaches approximately 90% of recall and precision values.

Dong et al. proposed a sound environment classifier for hearing aid applications which is implemented on a low-power DSP chip [19]. The system uses an HMM-based classifier using MFCC and delta-MFCC coefficients to highlight five sound sources which are speech, music, car noise and babble. According to their experiments they had evaluation results higher than 95% accuracy.

Beritelli et al. proposes a pattern recognition system for background sounds like bus, car, construction, dump, factory, office, pool, station, stadium and train sounds [6]. They used NN as a classifier and MFCC parameters as features. The average accuracy is in a range of 75% and 95% depending on the sound.

Shin et al. studied on a system for cough sounds to detect abnormal heath symptoms using acoustical information [14]. They proposed a hybrid model consisting of ANN and HMM to select cough sounds from other sounds in the environment. The input of their ANN model is human-auditory-characteristic-based filter banks on which Energy Cepstral Coefficients (ECC) are employed. Ergodic HMM is trained with the output of the ANN module and a

filtered envelope of the audio signal is used to handle temporal variation of the sound signal. Their proposed hybrid model introduced better results comparing to conventional HMM and MFCC usage in low SNR values.

Muhammad et al. studied on an environment recognition system using selected MPEG-7 audio low level descriptors and MFCC features [7]. MPEG-7 descriptors are ranked using Fisher's Discriminant Ratio (FDR) and top ranked descriptors are passed through PCA to obtain 13 features. These features are combined with MFCC features to complete the feature set. Environmental sounds like restaurant, crowded street, quiet street, shopping mall, car with open window, car with closed window, corridor of university campus, office room, desert and park are studied in this research. Experiments showed that they have an important improvement in classification performance in both systems using only MFCC and MPEG-7 low level features with GMM classifier.

Güvensan et al. [20] proposed an environmental sound recognition system for house appliances to create intelligent home environments. They used sounds of house appliances like refrigerator, blender, exhaust fan, dish washer, washing machine, blow dryer and ventilation units. ZCR, STE, Band-level energy (BLE), SC, Spectral Roll-off (SRO), SF and MFCC features are used to feed SVM and K-Nearest Neighbor (k-NN) classifiers. Their experiments showed that usage of SVM and MFCC features provides 98% accuracy success which is slightly higher than k-NN with MFCC usage.

Choi et al. proposed a real-time acoustic and visual context awareness system for mobile applications [21]. In audio part they introduced categories such as babble, car, bag, music, noisy, office, one-talk, public, subway and water. Long window length-MFCC and GMM are used for classification. An overall average accuracy of 98% is achieved.

Feki et al. proposes a framework for audio classification based on audio stream analysis [8]. They classified audio into classes like speech, music, ring tones, train, motorcycle, explosion, helicopter, slamming door, dog barking, bird, breeze glasses, applause, horse, cat, care, slot machine, wind, plane, laugh and police alarm. Their proposed system consists of three steps. The first step is the pre-processing part where audio stream is segmented and silence segments are detected. They used STE, low short-time energy ratio (LSTER), SF, band periodicity (BP) and MFCC feature to determine characteristics of the audio. In the second step speech, music and environmental sounds are automatically classified into detailed classes using NN,

HMM and SVM. In last step, they implemented a novel framework that encapsulates binary classifiers. The experiments show that, their proposed system has an accuracy higher than 90% for audio concept identification.

Chu et al. [11]. proposed an environmental sound classification system to understand a scene or context surrounding an audio sensor. They classified environmental sounds into classes like nature-daytime, vehicle, restaurant, casino, nature-nighttime, police, playground, traffic, thundering, train, rain, stream, waves and ambulance. They focused on feature selection using Matching Pursuit (MP) algorithm to obtain effective time-frequency features. The MP-based feature is adopted to supplement the MFCC features to yield higher recognition accuracy for environmental sounds. They conducted extensive experiments using GMM to demonstrate the advantages of MP features as well as joint MFCC and MP features in environmental sound classification.

Roma et al. present a method to search for environmental sounds in large unstructured databases of user-submitted audio, using a general sound events taxonomy from ecological acoustics [22]. In their study, frame level descriptors like MFCC and MPEG-7 are selected and only mean and variance of each frame-level descriptor are used. Selected frame-level descriptors chosen by feature selection process are: High frequency content, Instantaneous confidence of pitch detector (yinFFT), Spectral Contrast (SC) Coefficients, Silence Rate, Spectral Centroid, Spectral Complexity, Spectral crest, Spectral spread, Shape-based SC, Ratio of energy per band, ZCR, Inharmonicity and Tristimulus of harmonic peaks. In order to describe the temporal evolution of the frame level features, they computed several measures of the time series of each feature, such as the log attack time and a measure of decay and several descriptors derived from the statistical moments. Music, speech and voice sounds are classified with SVM with an accuracy of 96.19% in the first experiment. In the second experiment, environmental sounds are classified into rolling, scraping, deformation, impact, drip, pour, ripple, splash, explosion and whoosh sounds. Several sets of features are generated by progressively adding derivatives, attack and decay and temporal descriptors. Proposed feature set performs better than MFCC. In their third experiment they compare hierarchical and direct classification methods and results of direct method outperforms the hierarchical one.

CHAPTER 3

BACKGROUND

In this chapter, commonly used audio features and general classification methods are explained in detail in order to make the ideas in this thesis about audio classification concept more comprehensible for the reader. In the following sections, required information about MPEG-7 Audio Descriptors, MFCC and ZCR features; HMM and SVM classifiers and Genetic Algorithm is provided.

3.1 Audio Features

Audio features are basically some values containing meaningful information extracted from audio signals in order to compare and classify audio data. After the extraction of such information, it is stored in a content description in a compact way. A data descriptor is generally called a feature vector and the process for extracting such feature vectors from audio is called feature extraction. Audio feature extraction is generally based on audio analysis of spectral energy distribution, harmonic ratio or fundamental frequency of the audio signal.

3.1.1 MPEG-7 Audio Features

MPEG-7 standard is a widely used standard in audio classification area. It provides a large set of audio tools to create descriptions. MPEG-7 standard provides the following main elements [3]:

- Descriptors (D) define semantics and syntax of audio feature vectors.

- Description Schemes (DSc) define the semantics and syntax of the relationships between the components of descriptor.
- Description Definition Language (DLL) defines the syntax of description tools.

The interest of this study is the Descriptors in which semantic and syntax of feature vectors are defined. They are low-level audio descriptors containing temporal and spectral descriptors. These descriptors are classified into basic, basic spectral, single parameter, timbral temporal, timbral spectral and spectral basis descriptors (see Figure 3.1).

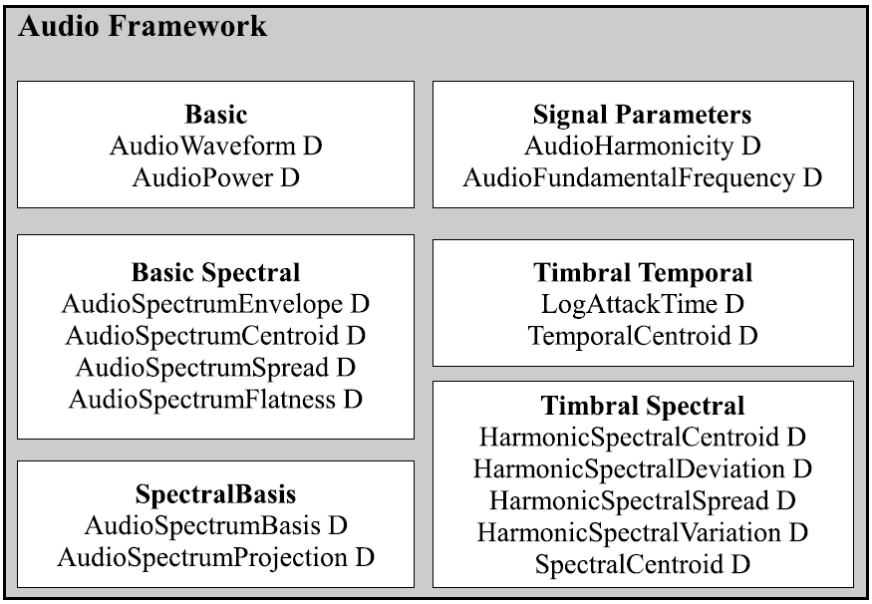

Figure 3.1: MPEG-7 audio framework overview.

3.1.1.1 Basic Descriptors

There exists two basic descriptors, namely Audio Waveform (AWF) and Audio Power (AP) descriptor. These are time domain descriptors of the audio content. The characteristic of the original audio signal can be observed in AWF and AP desciptors shown in Figure 3.2).

AWF descriptor is an estimate of the signal envelope in time domain storing the minimal and maximum samples. It is a compact and straightforward storage, display or comparison technique of waveforms.

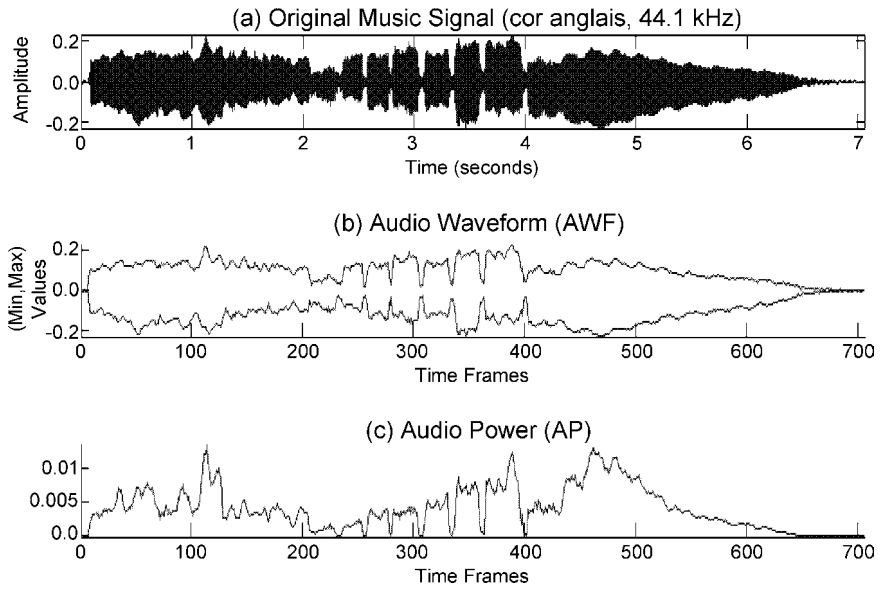

Figure 3.2: MPEG-7 basic descriptors extracted from a music signal [3].

AP describes the temporary smoothed instantaneous power of the audio signal in time domain. This descriptor provides a very simple information about signal amplitude and useful to detect silence parts of an audio stream.

3.1.1.2 Basic Spectral Descriptors

There are four basic spectral low level descriptors based on the estimation of short-term power spectra within overlapping time frames. The importance of these type of descriptors is the similarity to the sensitivity of human ear.

Audio Spectrum Envelope (ASE) Descriptor is a log-frequency power spectrum that is used to generate a reduced spectrogram of the original signal. It is the sum of the energies of power spectrum through series of frequency bands providing a compact representation of the spectrogram of the input signal. See Figure 3.3 (b) for the illustration of this descriptor.

Audio Spectrum Centroid (ASC) gives information about the shape of the signal, in other words, the the center of gravity of a log-frequency power spectrum of an audio signal. It contains perceptual sharpness information which indicates whether the power spectrum is dominated by high or low frequencies (see Figure 3.3 (c)). The log-frequency scaling approx-

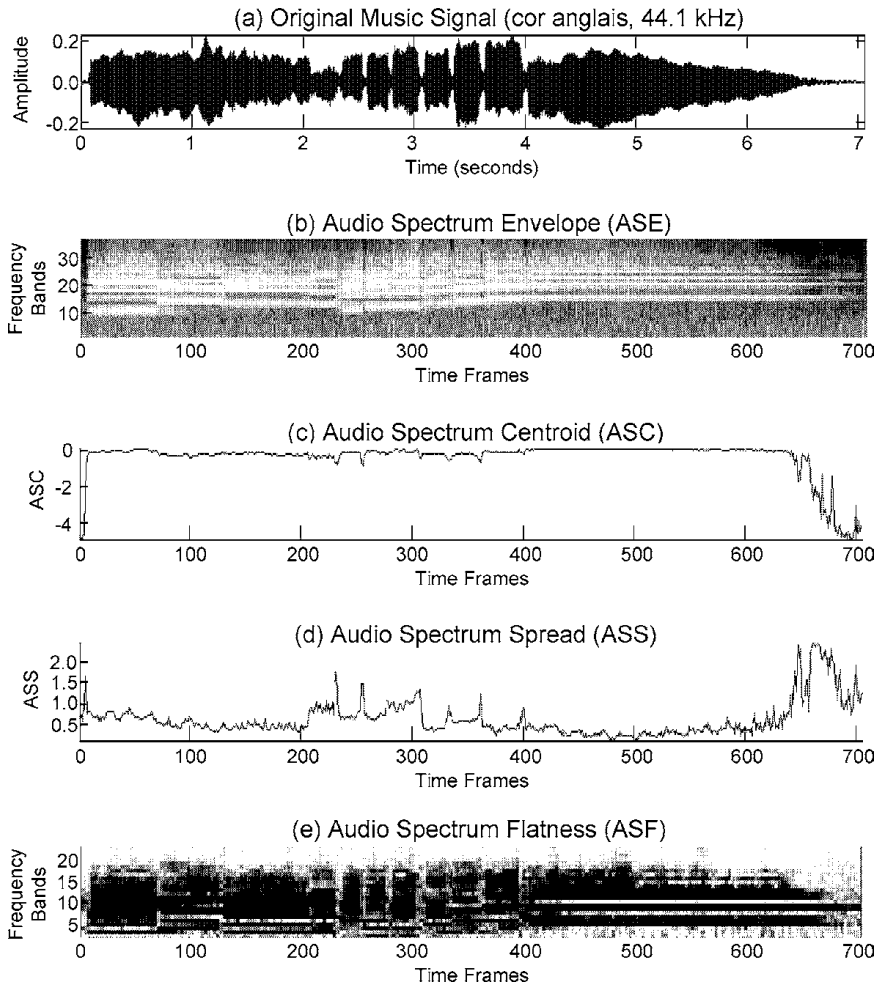

Figure 3.3: MPEG-7 basic spectral descriptors extracted from a music signal [3].

imates the perception of frequencies in the human hearing system.

Audio Spectrum Spread (ASS) is another measure of the spectral shape which is obtained by taking root mean square deviation of ASC (see Figure 3.3 (d)). Spectral spread is also called instantaneous bandwidth of an audio signal. ASS gives ideas about spectrum distribution around centroid. It is useful in discrimination of noise-like and tonal sounds.

Audio Spectrum Flatness (ASF) reflects the flatness properties of the power spectrum of a signal. It consists of a series of values expressing the deviation of the power spectrum of a signal from a flat shape inside a predefined frequency band (see Figure 3.3 (e)). A large deviation from a flat shape generally depicts tonal components. It helps discriminating white noise and impulse signals.

3.1.1.3 Signal Parameters Descriptors

Basic signal parameters reflect the harmonic structure of periodic sounds using frequency resolution (see Figure3.4). There are two of these descriptors: Audio Harmonicity (AH) and Audio Fundamental Frequency (AFF).

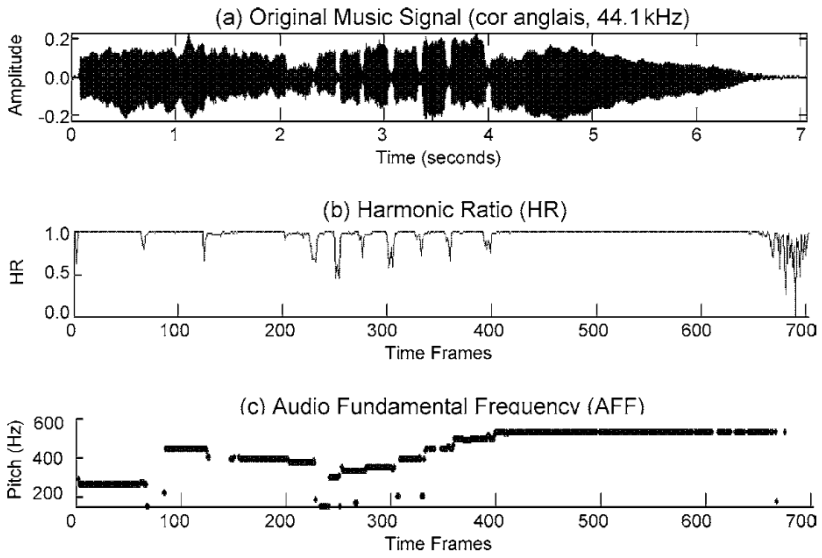

Figure 3.4: MPEG-7 basic signal parameters extracted from a music signal [3].

AH descriptor is a measure of the proportion of harmonic components in the power spectrum. It consists two measurements, namely Harmonic Ratio (HR) and Upper Limit of Harmonicity (ULH). HR is the ratio of harmonic power to the total power and ULH is the frequency beyond which the spectrum cannot be considered harmonic. Both HR and ULH are capable of to distinguishing between music and noisy sounds.

AFF descriptor provides estimations of the fundamental frequency in segments where the signal is assumed to be periodic. AFF is mainly used as an estimate for the pitch of music and voiced speech sounds.

3.1.1.4 Spectral Basis Descriptors

In this section Audio Spectrum Basis (ASB) and Audio Spectrum Projection (ASP) descriptors are explained.

ASB and ASP descriptors are obtained from ASE and Normalized Audio Spectrum Envelope (NASE) values of an audio signal. ASP and ASB contains rich information about the audio content besides the trade-off between dimension and wealth of information. Therefore, ASP and ASB are dimension-reduced versions of NASE using ICA and Single Value Decomposition (SVD).

3.1.1.5 Timbral Temporal Descriptors

The two timbral temporal descriptors Log Attack Time (LAT) and Temporal Centroid (TC) describe temporal characteristics of sounds. They are useful for the description of musical timbre (characteristic tone quality independent of pitch and loudness).

3.1.1.6 Timbral Spectral Descriptors

The five timbral spectral descriptors aim at describing the structure of harmonic spectra in a linear-frequency space. They are generally used for musical sounds.

3.1.2 Mel Frequency Cepstral Coefficients

The mel-frequency cepstrum (MFC) represents the short-term power spectrum of an audio signal, based on a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. MFCC is derived from a type of cepstral representation of the audio signal. The difference between the cepstrum and the mel-frequency cepstrum is equally spaced frequency bands on the mel scale in the MFC. It approximates the response of human auditory system more closely than the linearly-spaced frequency bands used in the normal cepstrum. In order to obtain the MFCC vectors, Fourier Transform (FT) of the signal is taken after sampling and windowing operations. Obtained powers of the spectrum is mapped onto the mel scale using triangular overlapping windows. Discrete Cosine Transform (DCT) of the list of mel

log powers are produced which provides a result spectrum where the amplitudes are MFCC vectors. This method is illustrated in Figure 3.5.

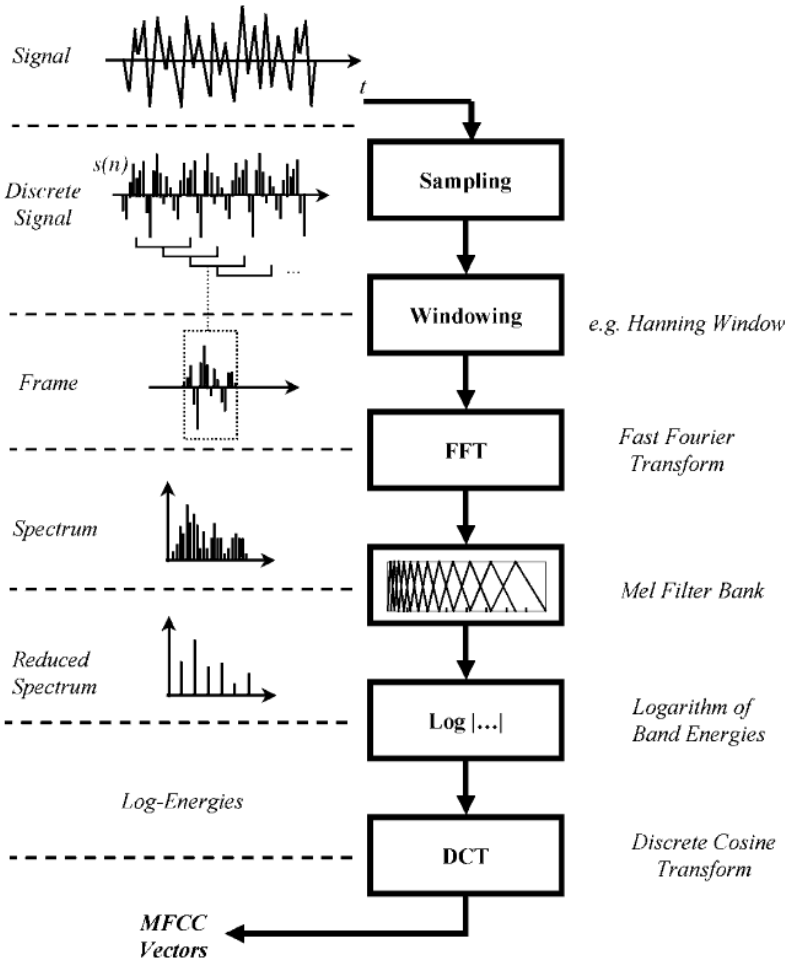

Figure 3.5: Extraction of MFCC vectors.

MFCC feature is powerful feature for speech recognition and music genre recognition with the powerful approximation to the human auditory system's response.

3.1.3 Zero Crossing Rate

Zero-crossing is a commonly used term in electronics, mathematics, and image processing. In mathematical terms, a "zero-crossing" is a point where the sign of a function changes (e.g. from positive to negative) which is then represented by a crossing of the axis (zero value) in

the function graph.

ZCR is the rate of sign-changes along a signal, more precisely, the rate at which the signal changes from positive to negative or vice verse. The ZCR is computed by counting the number of times that the audio waveform crosses the zero axis. ZCR is normalized by the length of the input signal $s(t)$ with the following formula in Wang's study [23]:

$$ZCR = \frac{1}{2}\Big(\sum_{t-1}^{T-1} |sign(s(t)) - sign(s(t-1)|\Big)\frac{F_s}{T} \tag{3.1}$$

where T is the total number f samples in $s(t)$ and F_s is the sampling frequency. $sign(x)$ function can be defined as:

$$sign(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{3.2}$$

This feature has been used heavily in both speech recognition, music genre classification and multimedia content analysis.

3.2 Classification Methods

3.2.1 Hidden Markov Model

Hidden Markov Model is an statistical method introduced by L.E. Baum and co-workers. HMM is a Markov process with hidden states. In Markov model states are visible by the observer and state transition probabilities are the parameters while states are hidden in HMM.

A complete HMM model can be defined as following [24]:

$$\lambda = (A, B, \pi) \tag{3.3}$$

- N, number of states in the model,
- M, number of distinct observation symbols per state, individual symbols are denoted as $V = \{v_1, v_2, ..., v_M\}$
- $A = \{a_{ij}\}$ state transition distribution where

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i];\ 1 \leq i, j \leq N. \tag{3.4}$$

- $B = \{b_j(k)\}$, observation symbol probability distribution in state j, where

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j];\ 1 \leq j \leq N, 1 \leq k \leq M. \tag{3.5}$$

- $\pi = \{\pi_i\}$, initial state distribution where

$$\pi_i = P[q_1 = S_i];\ 1 \leq i, j \leq N. \tag{3.6}$$

Given appropriate values of N, M, A, B and π, HMM can be used as a generator to give an observation sequence

$$O = O_1 O_2 \cdots O_T \tag{3.7}$$

HMM has three major problems:

- Problem 1 (Evaluation Problem): Computing probability of observation sequence $P(O|\lambda)$, given observation sequence $O = O_1 O_2 \cdots O_T$ and model $\lambda = (A, B, \pi)$. This problem is solved using Forward algorithm (see Section 3.2.1.1).
- Problem 2 (Decoding Problem): Choosing a corresponding state sequence $Q = q_1 q_2 \cdots q_T$, given observation sequence $O = O_1 O_2 \cdots O_T$ and model λ. This problem is solved using Viterbi algorithm (see Section 3.2.1.2).
- Problem 3 (Learning Problem): Adjusting model parameters $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$. This problem is efficiently solved using Baum-Welch algorithm(see Section 3.2.1.3).

3.2.1.1 Forward-Backward Algorithm

Consider a forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i | \lambda) \tag{3.8}$$

probability of the partial observation sequence , $O_1 O_2 \cdots O_t$ and state S_j at time t, given the model λ. Problem can be solved for $\alpha_t(i)$ inductively as follows:

1) Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1);\ 1 \leq i \leq N. \tag{3.9}$$

2) Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_(t+1)); \ 1 \leq t \leq T-1, \ 1 \leq j \leq N. \tag{3.10}$$

3) Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{3.11}$$

In the similar manner we can assume a backward variable $\beta_t(i)$ define as:

$$\beta_t(i) = P(O(t+1)O(t+2)) \cdots O_T | q_t = S_i, \lambda) \tag{3.12}$$

The observation sequence is thought to start from $t+1$, given state S_i at time t and the model λ. If we solve the equation for $\beta_t(i)$ inductively:

1) Initialization

$$\beta_T(i) = 1; \ 1 \leq i \leq N. \tag{3.13}$$

2) Induction

$$\beta_t(j) = \sum_{j=1}^{N} a_{ij} b_j(O_{t1} \beta_{t+1}(j)); \ t = T-1, T-2, \cdots, 1, \ 1 \leq i \leq N. \tag{3.14}$$

3.2.1.2 Viterbi Algorithm

Viterbi Algorithm finds the optimum state sequence given the observation sequence. To find the single best state sequence, $Q = \{q_1 q_2 \cdots q_t\}$, for given the observation sequence $O = \{O_1 O_2 \cdots O_T\}$, we define a quantity δ which is the highest probability along a single path, at time t:

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P[q_1 q_2 \cdots q_t = i, O_1 O_2 \cdots O_t | \lambda] \tag{3.15}$$

If we induct through Equation 3.15 we have the variable δ_{t+1} :

$$\delta_{t+1}(i) = [\max_i \delta_t(i) a_{ij}] \cdot b_j(O_{t+1}) \tag{3.16}$$

We keep track of argument maximized in (3.16), for each t and j via array $\psi_t i$. The procedure is shown below:

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1); \ 1 \leq i \leq N \tag{3.17a}$$

$$\psi_1(i) = 0. \tag{3.17b}$$

2) Recursion:

$$\delta_t(j) = \max_{i \le i \le N} [\delta_{t-1} a_{ij}] b_j(O_t);\ 2 \le t \le T,\ 1 \le j \le N \tag{3.18a}$$

$$\psi_(j) = \operatorname*{argmax}_{i \le i \le N} [\delta_{t-1}(i) a_{ij}];\ 2 \le t \le T,\ 1 \le j \le N \tag{3.18b}$$

3) Termination:

$$p* = \max_{i \le i \le N} [\delta_{t-1}] \tag{3.19a}$$

$$q_T^* = \operatorname*{argmax}_{i \le i \le N} [\delta_{t-1}] \tag{3.19b}$$

4) Backtracking:

$$q_T^* = \psi_{t+1}(q_{t+1}^*);\ t = T-1, T-2, \cdots, 1. \tag{3.20}$$

The major difference between forward calculation is the maximization over previous states in Equation 3.17a.

3.2.1.3 Baum-Welch Algorithm

The Baum-Welch algorithm is a particular case of a generalized expectation-maximization algorithm. It can compute maximum likelihood estimates and posterior mode estimates for the parameters (transition and emission probabilities) of an HMM, when given only emissions as training data.

For the procedure of re-estimation of HMM parameters, $\xi_t(i, j)$ is defined as the probability of being in a state S_i at time t and state S_j at time $t + 1$, given the model and the observation sequence ξ_t can be formalized:

$$\xi_t(i, j) = P(q_t = S\,i, q_{t+1} = S\,j | O, \lambda) \tag{3.21}$$

$\xi_t(i, j)$ can be written in the form below using the forward backward variables:

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1} \beta_{t+1}(j))}{PO|\lambda} \tag{3.22}$$

Probability of being in state S_i at time t, given the observation sequence and the model can be related with $\gamma_t(i)$ to $\xi_t(i, j)$ by summing over j:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j). \tag{3.23}$$

When $\gamma_t(i)$ is summed over time index t, the expected number of transitions made from state S_i is obtained. Similarly sum of $\xi_t(i, j)$ over t can be called as the expected number of transitions from state S_i to S_j, which is:

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{expected number of transitions from } S_i \tag{3.24a}$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{expected number of transitions from } S_i \text{ to } S_j \tag{3.24b}$$

Using the the formulas above, a method for re-estimation of parameters of HMM is proposed. Formulas for π, A and B are:

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i, \text{ at time } (t = 1) = \gamma_t(i) \tag{3.25a}$$

$$\bar{a_{ij}} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \tag{3.25b}$$

$$\bar{b_i}(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} \tag{3.25c}$$

Since the current model is defined as $\lambda = (A, B, \pi)$ and use this model to compute right-hand sides of 3.25a, 3.25b and 3.25c, the re-estimated model is defined as $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$. If $\bar{\lambda}$ is used in place of λ iteratively and re-estimation calculation is repeated, the probability of O being observed from the model can be improved until some limiting point is reached. Final result of this re-estimation is called maximum likelihood estimate of the HMM. The formulas 3.25a, 3.25b and 3.25c can be derived maximizing Baum's function over $\bar{\lambda}$. Maximization of $Q(\lambda, \bar{\lambda})$ leads to increased likelihood:

$$\max_{\bar{\lambda}}[Q(\lambda, \bar{\lambda})] = \max_{\bar{\lambda}}[\sum_{Q} P(Q|O, \lambda) log[P(O, Q|\bar{\lambda})]] \Rightarrow P(O|\bar{\lambda}) \geq P(O|\lambda). \tag{3.26}$$

3.2.2 Support Vector Machine

Vapnik [25] introduced SVM which has the principle of separating two classes by a linear hyper-plane. Positive and negative examples in the training set causes this hyper-plane which can be shown in Figure 3.6. The support vectors are marked with grey squares which define the margin of the largest separation between two classes. There exists many hyper-planes to classify the data, but the optimal one is the hyper-plane maximizing the distance between itself and the nearest data point, namely margin.

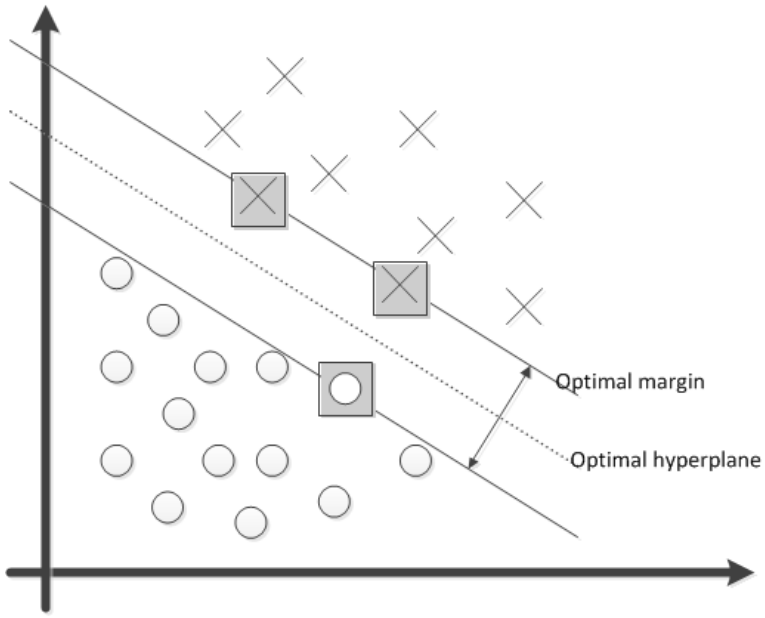

Figure 3.6: An example of a separable problem in two dimensional space.

The decision function of classifying an unknown point x is defined as:

$$f(x) = sgn(w \cdot x + b) = sgn\Big(\sum_{i=1}^{N_s} \alpha_i m_i x_i \cdot x + b\Big) \tag{3.27}$$

where (X_i, y_i) is the training vector having conditions $i = 1, ..., l$, $X_i = \{x_i, ..., x_n\}$ and $y_i \in \{+1, -1\}$. This unknown point is expected to be optimally separated by a hyper-plane formula:

$$W \cdot X + b = 0; W \in R^N \text{ and } b \in R \tag{3.28}$$

where W is the perpendicular vector to the hyper-plane and b is a constant. N_s is the support vector number, α_i is the Lagrange multiplier and $m_i \in \{-1, +1\}$ is a parameter describing which class x belongs to.

When the feature distribution of data has overlapping areas, it is not possible to separate the data in the given input space. For non-separable data like overlapping nature of audio data, kernel methods are used to map the feature vectors into a higher dimensional space where linear separation of the training set is possible. If kernel functions are used to construct the optimal hyper-plane, the decision function becomes like:

$$f(x) = sgn(w \cdot x + b) = sgn\Big(\sum_{i=1}^{N_s} \alpha_i m_i K(x_i, x) + b\Big) \tag{3.29}$$

In this thesis work, Radial Basis Kernel is used after the comparative tests with Linear Kernel. As shown in the equation 3.30, Linear Kernel provides less complexity than Radial Basis Kernel (See equation 3.31) but for our dataset, Radial Basis Kernel provides more satisfactory accuracy values.

Linear Kernel:

$$K(x, y) = x \cdot y \tag{3.30}$$

Radial Basis Kernel:

$$K(x, y) = exp(-\gamma \cdot \| x - y \|^2) \tag{3.31}$$

3.2.3 Genetic Algorithm

Genetic Algorithm (GA) provides a learning method with an analogy to biological evolution which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover. GA addresses the problem of searching a space of candidate hypothesis to find the best hypothesis. To reach the best hypothesis it evolves through generations to optimize a numerical measure called fitness.

GA iteratively updates a pool of hypotheses, namely population. On each iteration, all members of the population are evaluated according to the fitness function. A new population is then generated by probabilistically selecting the most fit individuals (chromosomes) from the current population. Some of these selected individuals are carried forward into the next generation and others are used as the basis for creating new offspring individuals by applying genetic operations such as crossover and mutation.

The prototypical genetic algorithm [26] can be parameterized as:

GA(Fitness, Fitness_threshold, p, r, m), where:

Fitness: A function that assigns an evaluation score, given a hypothesis.

Fitness_threshold: A threshold specifying the termination criterion.

p: The number of hypotheses to be included in the population.

r: The fraction of the population to be replaced by Crossover at each stop.

m: The mutation rate.

- Initialize population: P ← Generate p hypothesis at random
- Evaluate : For each h in P, compute Fitness(h)
- While

$$\max_h Fitness(h) < Fitness\_threshold$$

  do the following : Create a new generation, P_s:

  1. Select : Probabilistically select $(1 - r)p$ members of P to add P_s. The probability $P_r(h_i)$ of selecting h_i from P is given by

$$P_r(h_i) = \frac{Fitness(h_i)}{\sum_{j=1}^{p} Fitness(h_j)} \tag{3.32}$$

  2. Crossover : Probabilistically select $\frac{r-p}{2}$ pairs of hypotheses from P, according to $P_r(h_i)$ given above. For each pair, $< h_1, h_2 >$, produce two offspring by applying the Crossover operator. Add all offspring to P_s.
  3. Mutate : Choose m percent of the members of P_s with uniform probability. For each, invert one randomly selected bits in its representation.
  4. Update : $P \leftarrow P_s$,
  5. Evaluate : for each h in P, compute $Fitness(h)$

- Return the hypothesis from P that has the highest fitness.

Here a population has p hypotheses and on each iteration, the successor population P_s is formed by probabilistically selecting current hypotheses according to their fitness and by adding new hypotheses. New hypotheses are created by applying a crossover operator to pairs of most fit hypotheses and by creating single point mutations in the resulting generation of hypotheses. This process is iterated until sufficiently fit hypotheses are discovered.

3.3 Definitions

Acoustic Class (AC): Classes based on the acoustic feature of audio. An audio sample is classified into these classes using classifiers and audio features. Emergency alarm, car horn, gun-shot, explosion, automobile, motorcycle, helicopter, wind, water, rain, applause, crowd and laughter are selected as *acoustic classes*.

Semantic Class (SC): Higher level classes based on the acoustic classification results. System decides on these classes using only acoustical results. Outdoor, nature, meeting and violence are selected as *semantic classes*.

Audio Segment (AS): An audio stream of 1 second length.

Segment Group (SG): Sequence of segments between two consecutive silence segments.

Precision: In classification context, precision for a class is the number of true positives divided by the total number of elements labelled as belonging to the positive class (Equation 3.33).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.33}$$

Recall: The number of true positives divided by the total number of elements that actually belong to the positive class (Equation 3.34).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.34}$$

F-measure: Weighted harmonic mean of precision and recall (Equation 3.35).

$$\text{F-measure} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.35}$$

CHAPTER 4

PROPOSED METHOD

4.1 Classification Approach

Audio features and classifiers are two main concepts in audio classification research. Audio features are vectors or scalars containing several descriptive measures of an audio stream. Whereas classifiers are statistical or linear models representing specifically the intended classes. Audio features are utilized to create these models, which is called model training. In order to create and verify these models, audio dataset is divided into train and test sets. Train set is used for model training and test set is used for verification.

In this study, HMM (Section 3.2.1) and SVM (Section 3.2.2) are used as classifiers with combinations of MPEG-7 (Section 3.1.1), MFCC (Section 3.1.2) and ZCR (Section 3.1.3) as audio features.

Proposed classification system consists four main blocks (Figure 4.1): preprocessing, model training, acoustic and semantic classification.

In preprocessing block, a given audio clip (file) is divided into one-second segments which will be a sequence of consequent segments. Audio features are extracted from each segment and system labels the silence and non-silence segments of this segment sequence.

In order to build up the decision mechanism, the models are created from the train data set after a preprocessing step within the model training block. The best representative feature and classifier combination is utilized for model training.

In acoustic classification block, non-silence segments of the given segment sequence are clas-

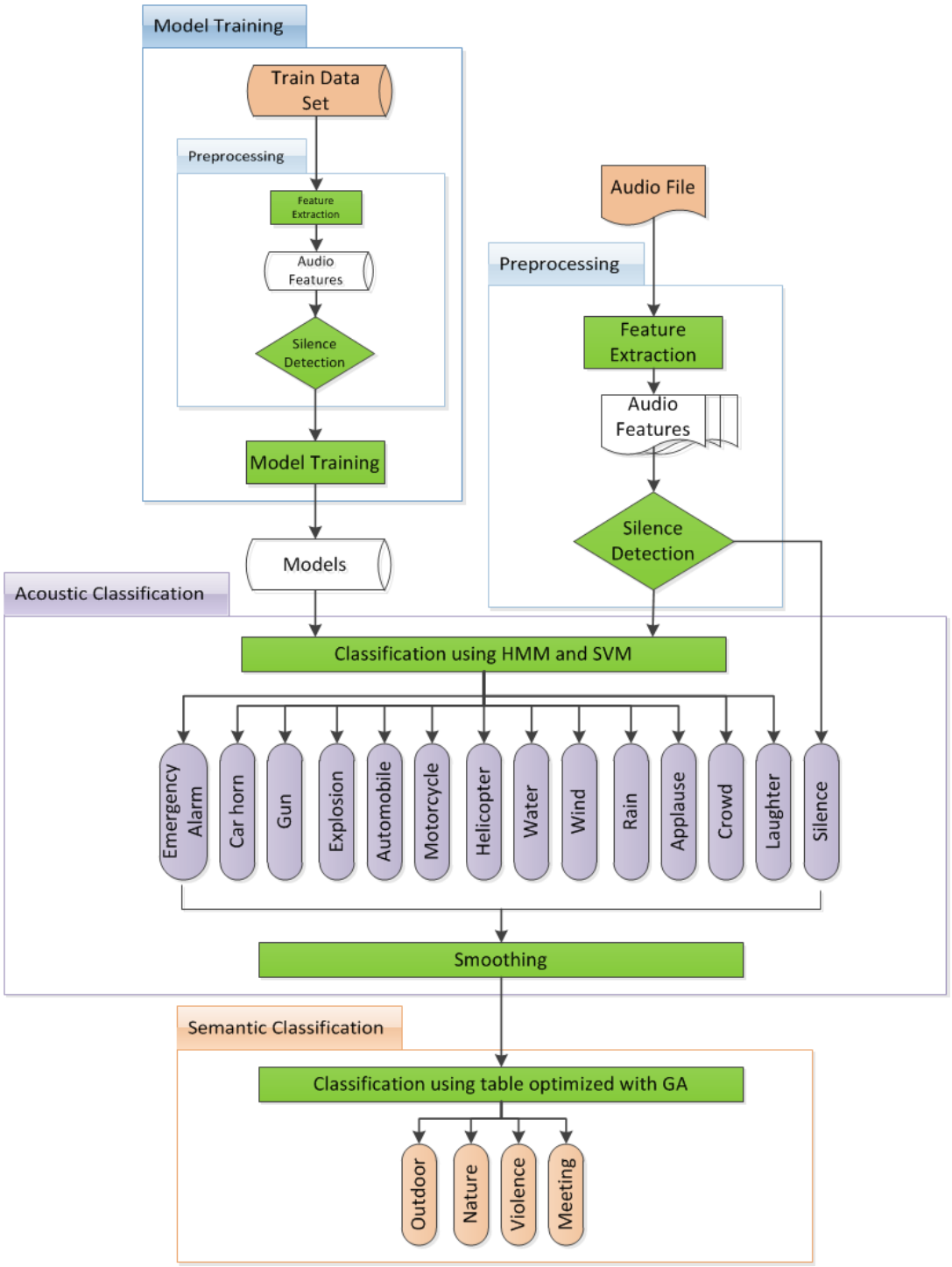

Figure 4.1: Block diagram of the proposed classification approach.

sified into emergency alarm, car horn, gun-shot, explosion, automobile, motorcycle, helicopter, wind, water, rain, applause, crowd and laughter classes. In the content of this thesis study, we named these classes as *acoustic classes* and this process is called *acoustic classification* which refers to the classification using trained models with extracted audio features. In order to discard the classification errors, given segment sequences are passed through the smoothing process.

Acoustically labelled segment sequence is the input for *semantic classification*. In semantic classification block, this sequence is classified into higher level *semantic classes*, namely outdoor, nature, violence and meeting.

4.1.1 Preprocessing

4.1.1.1 Feature Extraction

In our system, different kinds of feature sets are used in order to make comparisons in a wide range of audio features. All features are extracted using 30 ms frames and 10 ms hop-size. Feature sets used in the system are:

- A set containing MPEG-7 ASP feature. This feature represents one frame with 21-dimensional vector.
- A set containing MPEG-7 ASF feature. This feature represents one frame with 20-dimensional vector.
- A set containing MPEG-7 AH feature. This feature represents one frame with scalar value.
- A set containing MPEG-7 ASC feature. This feature represents one frame with scalar value.
- A set containing MPEG-7 ASS feature. This feature represents one frame with scalar value.
- A set containing ZCR feature. This feature represents one frame with scalar value.
- A set containing MFCC feature. This feature represents one frame with 13-dimensional vector.

- A set containing a composition of MPEG-7 ASF, ASC, ASS and AH features. This feature composition represents one frame with 23-dimensional vector.
- A set containing a composition of MPEG-7 ASF, ASC, ASS and ZCR features. This feature composition represents one frame with 23-dimensional vector.
- A set containing a composition of MFCC, MPEG-7 ASC, ASS and AH features. This feature composition represents one frame with 16-dimensional vector.
- A set containing a composition of MFCC, MPEG-7 ASC, ASS and ZCR features. This feature composition represents one frame with 16-dimensional vector.

MPEG-7 Audio Encoder Project [27] is used for extraction of Audio MPEG-7 features. MFCC features are extracted using Malcolm Stanley's Auditory Toolbox [28]. Zero Crossing Rate feature is extracted using the formula in Section 3.1.3.

4.1.1.2 Silence Detection

The aim of this step is to detect the silent segments. Kiranyaz et al. [29] proposed an approach for silence detection. Input audio stream is first divided into frames and then silence detection is performed per frame in their proposed approach. In our study, similar calculations are performed for silence detection. The minimum P_{min}, maximum P_{max}, and average P_{μ} Audio Power (AP) values are calculated from each one-second segment. Following conditions are checked to determine silent segments:

- $P_{max} >$ Minimum Audible Power Value
- $P_{max} \geq P_{min}$

If the presence of non-silent segment is confirmed, then Threshold value T is calculated according to the following equation:

$$T = P_{min} + \lambda_s(P_{\mu} - P_{min}), 0 < \lambda_s \leq 1 \tag{4.1}$$

where λ_s is the silence coefficient, which determines the silence threshold value between P_{min} and P_{μ}. If all samples of AP feature values for one-second segments are less than the calculated threshold T, then that segment is classified as silent.

4.1.2 Model Training

SVM and HMM models are created to be able to distinguish between *acoustic classes*. After the feature extraction of all files in the train data set, outputs are used to train HMM and SVM models.

4.1.2.1 Hidden Markov Model Training

HMM models are trained for each *acoustic class* resulting 13 HMM models. According to the state optimization experiment results (see Appendix D), 5-state Ergodic Hidden Markov Model is chosen for this study.

HMM parameters are estimated by using the well-known Baum-Welch algorithm (see Section 3.2.1.3). Taking the initial values for all the parameters, Baum-Welch finds the optimum values for the parameters by iterative re-estimations. This problem is called "training problem" in HMM applications. Since this algorithm finds only locally optimum values, the initial guess for the parameters is very important. For that reason K-Means Clustering algorithm is applied to estimate these values instead of randomly assigned initial values.

After specification of complete parameter set of HMM parameters, the classification problem becomes "evaluation problem", in other words, given a model and a sequence of observations, computing the probability of the observed sequence to be produced by the model. Forward-Backward algorithm (see Section 3.2.1.1) is used to select the audio class label for the given observation sequence.

4.1.2.2 Support Vector Machine Training

Support Vector Machine training is performed using multi-class SVMs and models are created with one-versus-all approach. For each class, an SVM model is trained to distinguish between itself and the rest of the classes with a weight of 12. As discussed in Section 3.2.2, Radial Basis Kernel function is used to map data into high-dimensional feature space.

4.1.2.3 Verification

During the verification process, audio segments are classified into thirteen classes. Probabilities coming from HMM and SVM models are calculated and segments are labelled with the class label whose model outputs the highest probability.

JAHMM [30] and LIBSVM [31] library packages are used for HMM and SVM classifications, respectively.

Audio feature sets mentioned in Section 4.1.1.1 are used for the experiments in order to reach the highest recall and precision values with SVM and HMM models. MPEG-7 ASF, ASS, ASC and AH feature combination with SVM classifier (see Table A.21) is the best combination for selected environmental sound classes. All experiments are explained in the following chapter.

4.1.3 Acoustic Classification

In this step an audio segment is classified into proposed *acoustic classes*. Consequent segments are smoothed in order to discard classification errors.

4.1.3.1 Classification

In HMM classification, a segment is given as an input for each HMM model. The model providing the highest score for the given segment labels the segment with the acoustic class label. Given the models $M_1, M_2, ..., M_n$ and a one-second segment containing frame sequence S; the probability P_i (the probability of a sequence to be produced from an HMM model) is calculated as follows (see Equation 3.11 for *HmmPredict* function):

$$P_i = HmmPredict(M_i, S);\ 0 \leq i < 13 \tag{4.2}$$

Then, the maximum probability from each model M_i, S pair.

$$Pmax = max(HmmPredict(M_i, S), ..., HmmPredict(M_n, S));\ 0 \leq i < 13 \tag{4.3}$$

Model M_i with the *Pmax* probability, labels sequence S.

In SVM classification, a segment is given as input for each SVM model but processed in a different manner. Since SVM is a binary classifier, the outputs of the models are binary, namely positive (1) and negative (-1) values. Given the models $M_1, M_2, ..., M_n$, one-second segment containing frame sequence $S = \{s_1, s_2, ..., s_n\}$ with size n, positive or negative value V coming from each frame prediction, positive frame count PC; probability P_i (percentage of positive frame count over total frame count) is calculated as follows (see equation 3.11 for $SvmPredict$ function):

$$V_j = SvmPredict(M_i, s_j);\ 0 \leq i < 13,\ 0 \leq j < n \tag{4.4}$$

$$PC_i = V_0 + V_1 + ... + V_j;\ V_j > 0,\ 0 \leq i < 13,\ 0 \leq j < n \tag{4.5}$$

$$P_i = PC_i\ /\ n;\ 0 \leq i < 13 \tag{4.6}$$

Then, the maximum probability from each model M_i, S pair:

$$Pmax = max(PC_0, PC_1, ..., PC_i);\ 0 \leq i < 13 \tag{4.7}$$

Sequence S is labelled by the model M_i which outputs the $Pmax$ probability.

4.1.3.2 Smoothing

After several tests, some classification errors are observed during the audio classification. An illustrative example is given in Figure 4.2. Coloured boxes indicates acoustically labelled segments in a sequence. Black coloured boxes are the silent segments. Yellow, green and red boxes indicates different acoustic labels. Then, majority of yellow boxes indicates that this segment sequence belongs to yellow type heuristically. Therefore, system assumes that the green and red labelled segments are classification errors. After the smoothing process this misclassification is discarded.

In this process two observed cases are ruled out :

- *Rule 1*: $(s_1 \neq s_0\ \ \&\ \ s_0 = s_2) \Rightarrow s_1 = s_0$. In this rule, consequent three segments are considered at a time s_0, s_1, s_2. This rule implies that if the middle segment has different

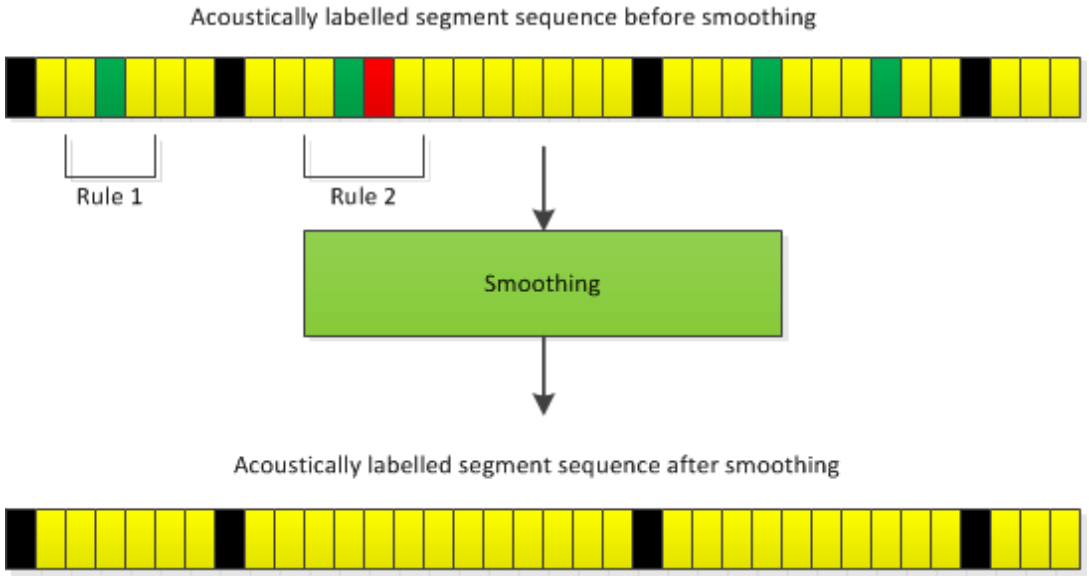

Figure 4.2: An illustrative example for smoothing.

class label than the other surrounding two segments and these segments have the same class label; then consider that the middle segment is misclassified and change the label of this segment to the label of first segment.

- *Rule 2*: $(s_0 = s_3 \wedge s_1 \neq s_0 \wedge s_1 \neq s_2 \wedge (s_1.secondlabel = s_0 \vee s_2.secondlabel = s_0)) \Rightarrow s_1 = s_0, s_2 = s_0$. In this rule, four-second sequence is considered at a time s_0, s_1, s_2, s_3 standing for the audio. This rule is applicable if middle two segments does not have same labels with each other and the surrounding segments and surrounding segments has same labels. In order to regard the middle segments as misclassified, the second labels (label of the model providing the second highest score during the classification step) should be checked. If the second labels of at least one of the middle segments are equal to the surrounding segment labels, the middle segments are considered as misclassification (See also examples in Appendix C).

Following the smoothing process, temporally adjoining segments are combined together if they share the same acoustic class label. As a result, the entire audio sequence is partitioned into homogeneous segment joints having a distinct acoustic class label. A segment joints has attributes such start time, end time, duration, acoustic class label and acoustic classification score.

4.1.4 Semantic Classification

The result of the *acoustical classification* of a segment sequence is used as input in order to distinguish between higher level classes such as outdoor, nature, violence and meeting. This classes are called *semantic classes*. Given a sequence of segments, system needs some distinctive points in order to detect the starting and ending times of the concept changes. For that reason, silent segments are assumed to be the marker segments in order to determine the *segment groups* which are the candidate segment sequences for semantic classification. In Figure 4.3 *segment groups* are illustrated. Black coloured boxes are the silent segments whereas red, green and blue coloured boxes represents the acoustically labelled segments.

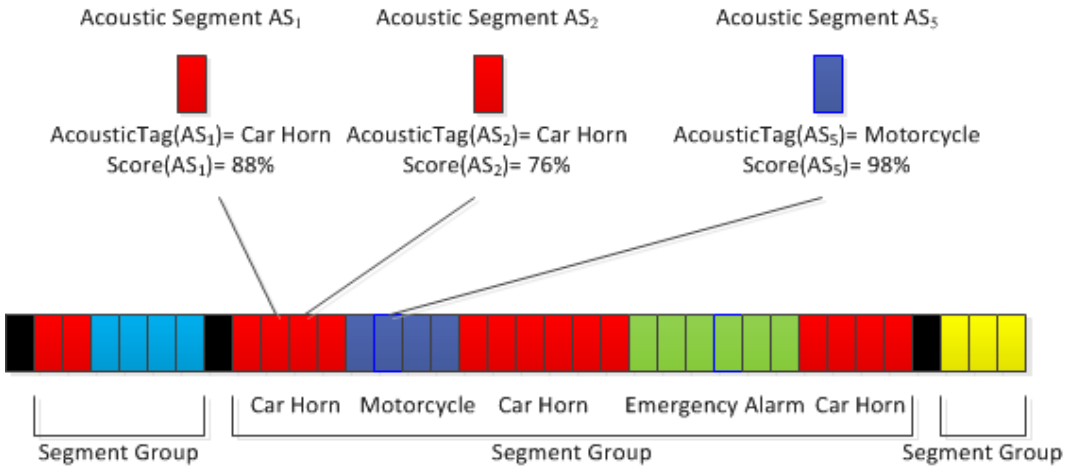

Figure 4.3: An illustrative example for segment group.

A grading technique is proposed which calculates the impacts of each acoustic class for candidate semantic classes. Given the segment group $SG = AS_0, AS_1, ..., AS_n$ where AS_i is the acoustically labelled segment and n is the total number of segments in a SG, $Grade(SG, SC)$ function returns the total impacts of each AS_i for the semantic class SC. The group is labelled by the winner semantic class which has the highest grade greater than the threshold value. The maximum grade is calculated as follows:

$$\text{Maximum Grade} = max(Grade(SG, outdoor), ..., Grade(SG, violence)) \tag{4.8}$$

and $Grade(SG, SC)$ function is defined as:

$$Grade(SG, SC) = \frac{\sum_i^n Impact(SC, AS_i) Score(AS_i)}{n} \tag{4.9}$$

where $Impact(SC, AS)$ function returns impact degree of AS for SC which are generated using Genetic Algorithms (see Section 3.2.3) shown in Table 5.3. SC represents outdoor, nature, meeting or violence semantic classes, n is the number segments in SG and $Score(AS_i)$ is the classification score of i^{th} segments coming from the classifier.

Impact table (see Table 5.3) contains the impact values of each *acoustic class* on each *semantic class*. The values in this table are calculated with Genetic Algorithm. The impact table and the threshold value are used as the chromosome. 50-sized population is evolved to reach the optimal values of the chromosome using the average F-measure value as the fitness function. After the optimization experiments the proposed semantic classification succeeded to 87% average F-measure values shown in Table 5.5. JGAP [32] package is used for the calculations.

4.2 Retrieval Approach

The proposed classification approach provides the capability of categorizing environmental sounds. With the growth of data set, gathering information becomes a problem. Retrieval is the activity of obtaining relevant information from a collection of resources. In the context of this thesis, two types of retrieval techniques are implemented to provide an easy access to the desired information: Keyword queries and query by example. These techniques are described in the following subsections.

4.2.1 Keyword Queries

In order to retrieve the acoustically and semantically labelled segments in classification results, the proposed system provides keyword-based querying capability of these classes. The queries can be expressed also with a possibility degree that are associated with the classifier result. In addition, temporal queries are supported in order to retrieve data using the duration, start and end time information:

- Acoustic Label:

"Find all *emergency alarm* segments."

- Acoustic Possibility Degree:

  "Retrieve all possible segments with an acoustic possibility degree between *80%* and *90%*."

- Semantic Label:

  "Find all *outdoor* segments."

- Semantic Possibility Degree:

  "Retrieve all possible segments with a semantic possibility degree between *50%* and *70%*."

- Start and End Time:

  "Find all segments between the 40^{th} and 50^{th} seconds."

- Duration:

  "Find all segments of length between *2* and *4* seconds."

Keyword-based and temporal queries can be combined to form more complex queries:

- Acoustic Label and Possibility Degree:

  "Retrieve all possible *emergency alarm* segments with an acoustic possibility degree between *80%* and *90%*."

- Semantic Label and Possibility Degree:

  "Retrieve all possible *outdoor* segments with a semantic possibility degree between *50%* and *70%*."

- Duration, Start and End Time:

  "Retrieve all segments of length between *2* and *4* seconds and between the 40^{th} and 50^{th} seconds."

- Acoustic and Semantic Label, Acoustic and Semantic Possibility Degree, Duration, Start and End Time:

> "Retrieve all possible *emergency alarm* segments with an acoustic possibility degree between *80%* and *90%* and semantically labelled as *outdoor* with a semantic possibility degree between *50%* and *70%*, between the 40^{th} and 50^{th} seconds and with a length between *2* and *4* seconds."

In the classification stage, the results were written into a text file. In this stage, the results are loaded to the system from this result file and stored in a table. This table contains columns for start and time, labels and possibility degrees for acoustic and semantic classes. Whereas the rows contain audio parts, namely the concatenation of consequent segments which have same acoustic class label. System filters the keywords (which are also column names) to provide the mentioned query capabilities. For the last query example above, system retrieves the rows from the table which ensures the following conditions:

1. acoustic class label = *emergency alarm*;
2. acoustic classification score $\geq$ *80%* and acoustic classification score $\leq$ *90%*;
3. semantic class label = *outdoor*;
4. semantic classification grade $\geq$ *50%* and semantic classification grade $\leq$ *70%*;
5. start time $\geq$ *40* and end time $\leq$ *50*;
6. duration (end time - start time) $\geq$ *2* and duration $\leq$ *4*.

4.2.2 Querying by Example

Query by Example (QBE) retrieval technique is basically a similarity search of the given segment among the search space. When a query audio file is given as an input to the system and relevant files are requested, both the query and each audio segment in the segmented audio file are represented as feature vectors. The system calculates the similarity measurements of the queried audio file and search space vectors and outputs a list of audio segments according to the decreasing similarity order.

For similarity measurements, MPEG-7 ASF, ASC, ASS and AH feature combination is selected as the feature vector. Similarity between two series of feature vectors is measured by employing a correlation function [33] which computes the correlation coefficient of A_{mn} and

B_{mn}, where A and B are the feature vector representations of two audio segments, m is the size of the segment and n is the feature vector size. The correlation function is defined as follows:

$$\frac{\sum_m \sum_n \left(A_{mn} - \bar{A}\right)\left(B_{mn} - \bar{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}} \tag{4.10}$$

Maximum correlation between audio search space and query audio matrices are calculated by sliding the query audio matrix over audio search space matrix and computing the correlation coefficient for every window position of 30 ms. A correlation coefficient array is generated and processed to find the window position where correlation coefficient peaks.

The system provides three different kinds of queries which are *point query*, *k-nearest neighbour query* and *range query*. *Point query* is a type of query in which system retrieves the most similar segment to the given input. In *k-nearest neighbour query*, the k best matches are retrieved. For the range query, user should provide predetermined ranges in order to retrieve the intended content. Given an audio sample user can search for similar segments in the segmentation results. For instance:

- Point Query

  "Retrieve the most similar segment to the *given audio.*

  "Retrieve the most similar segment to the *given audio* in *emergency alarm* segments labelled as *outdoor*."

- Range Query

  "Retrieve similar segments to the *given audio* with similarity ratio between *80%* and *90%*."

  "Retrieve similar segments to the *given audio* with similarity ratio between *80%* and *90%* and labelled as *emergency alarm*."

- kNN Query

  "Retrieve best 6 matches similar to the *given audio*."

  "Retrieve best 6 matches similar to the *given audio* and labelled as *nature*."

CHAPTER 5

EVALUATION

5.1 Experiments for Acoustic Classes

In the proposed system, HMM and SVM classifications are tested with the following 11 feature sets : ASP, ASF, ASC, ASS, AH, ZCR, MFCC, (ASF + ASC + ASS + AH), (ASF + ASC + ASS + ZCR), (MFCC + ASC + ASS + AH) and (MFCC + ASC + ASS + ZCR)

The data set, in Table 5.1 is used during train and test procedures of our models. 85% of dataset is used for model training while the rest is used for testing.

Table 5.1: Overview of acoustic data set.

	Duration
Emergency Alarm	23 min 39 sec
Car Horn	4 min 44 sec
Gun-shot	9 min 50 sec
Explosion	19 min 14 sec
Helicopter	7 min 24 sec
Motorcycle	9 min 57 sec
Auto-mobile	6 min 53 sec
Rain	13 min 45 sec
Wind	17 min 15 sec
Water	27 min 49 sec
Applause	10 min 30 sec
Laughter	13 min 26 sec
Crowd	9 min 45 sec
TOTAL	3 hours 4 min 11 sec

5.1.1 Dataset Collection

Audio clips used in train and test are collected from internet [34, 35], films and videos [36]. All audio files are listened and cleaned from irrelevant parts and converted to PCM (little endian 16 bit) 44100 Hz frequency and mono format. Audio clips with long duration are separated into smaller clips not exceeding 15 seconds. Battle sounds has the average shortest duration of a second because of the nature of gun-shot and explosion sounds. Other sounds has an average length of 10 seconds. Emergency alarm set contains sounds like ambulance, police, fire service sirens and other emergency alarms like nuclear and fire alarms. Explosion sound set contains explosion and bomb sounds. Gun-shot sound set contains gun-shot, riffle, fireworks, machine gun and laser gun sounds. Our data set also contains automobile, helicopter and motorcycle sounds including sounds of starting and stopping engine, sounds taken from traffic, sound inside vehicle and outside vehicle. Water sounds set contains sounds like swimming, splash, waves, ocean, sea, dropping and rowing sounds.

5.1.2 Experiments to Find Best Representative Feature Set and Classifier

In the following paragraphs, experiments to find the best representative feature set for each classifier are explained and the classification test results are discussed. See Appendix A for confusion matrices.

5.1.2.1 Experiments with HMM

In the experiment with ASP feature, the system success is quiet low with 36.2% average F-measure shown Table A.1.

For ASF feature, test results can be seen in Table A.2. Since this feature depicts the flatness of audio, it is good at differentiating between impulse-like and noise-like sounds. This feature is certainly successful for Emergency Alarm, Wind and Helicopter acoustic classes. The average F-measure is 55.7% which is relatively better than ASP feature but not satisfactory.

Tests results for ASC and ASS features can be observed in Table A.3 and Table A.4 respectively. Since these features are one dimensional, the success is quiet low when they are used standalone.

For MFCC feature, test results can be seen in Table A.5. This feature provides better F-measure (65.3 %) than ASF (see Table A.2) feature. As seen in the results, F-measure for each class is smoothly distributed over classes compared to ASF results.

AH and ZCR feature test results are shown in Table A.6 and Table A.7 respectively. They are both one dimensional features and results are not satisfactory to classify selected classes.

ASF and MFCC features provides relatively better result than ASP, ASC, ASS, AH and ZCR features. In the experiments with ASF and MFCC features the calculated F- measure values are 55.7% and 65.3% respectively. In order to get better results, combinations are experimented: (ASF + ASC + ASS + ZCR) , (ASF + ASC + ASS + AH), (MFCC + ASC + ASS + ZCR) and (MFCC + ASC + ASS +AH).

Combining ASF, ASC, ASS and ZCR features increased the F-measure value to 62.1% where these features has 55.7%, 27.4%, 20.2% and 27.8% F values respectively when they are tested stand alone. Test results are shown in Table A.8.

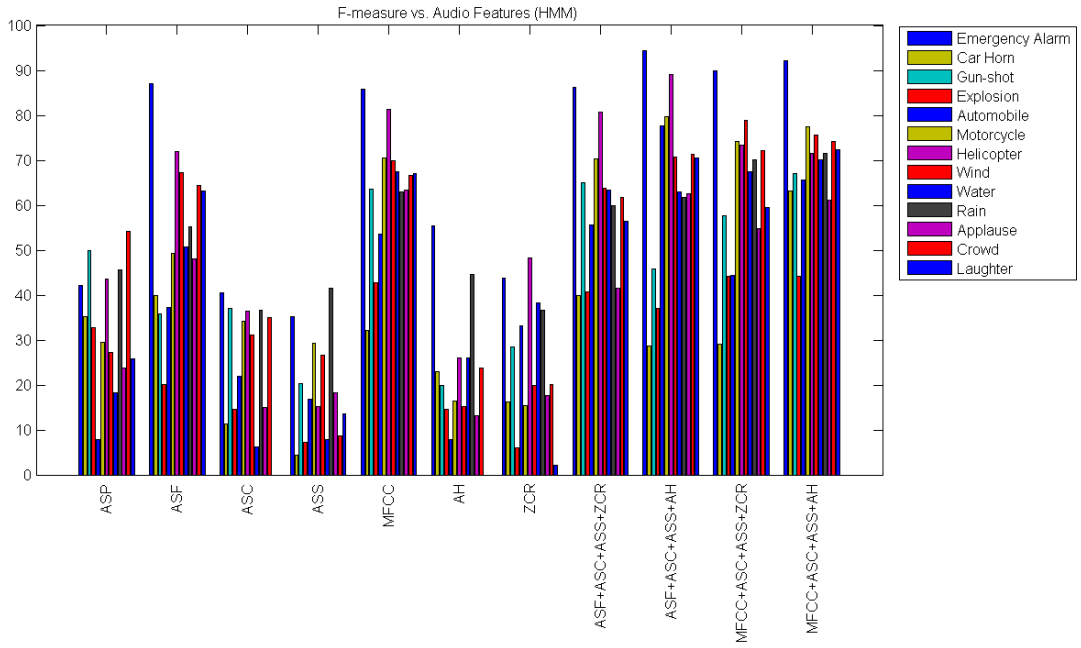

Figure 5.1: Results of HMM classification experiments.

In the experiment with ASF, ASC, ASS and AH features the F-measure value is calculated as 69.4% (see Table A.9) which is more successful then 62.1% (ASF+ASC+ASS+ZCR re-

sult). When the results of two tests are examined, contribution of AH feature is better in distinguishing Emergency Alarm, Motorcycle, Automobile, Helicopter and Laughter sounds.

Combinations of MFCC+ASC+ASS+ZCR and MFCC+ASC+ASS+AH feature sets are also tested. Table A.10 and Table A.11 shows that MFCC+ASC+ASS+AH set has the highest F-measure value of 70.6% among all feature sets. MFCC+ASC+ASS+ZCR combination has relatively lower F-measure value of 63.9% then MFCC+ASC+ASS+AH combination.

In Figure 5.1 results of eleven experiments are shown. One dimensional features, ASC, ASS, AH and ZCR, have quite lower classification success compared to higher dimensional features. Classification performance of feature combinations is obviously higher than the standalone features. Line graph in Figure 5.2 represents the performance of each feature set on selected classes.

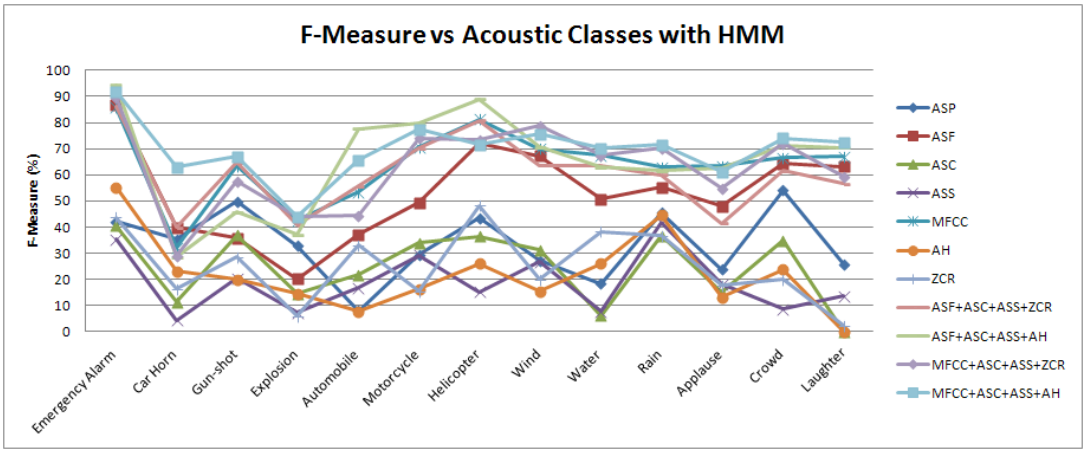

Figure 5.2: Performances of feature sets on acoustic classes using HMM classifier.

5.1.2.2 Experiments with SVM

ASF feature has 52.8% average F-measure value (see Table A.13). This feature is less successful in HMM tests (see Table A.2). Tests with both classifiers shows that, this feature is not satisfactory for the classification of proposed classes in case of standalone usage.

MFCC feature is the most successful feature for SVM with F-measure value of 55.9% if

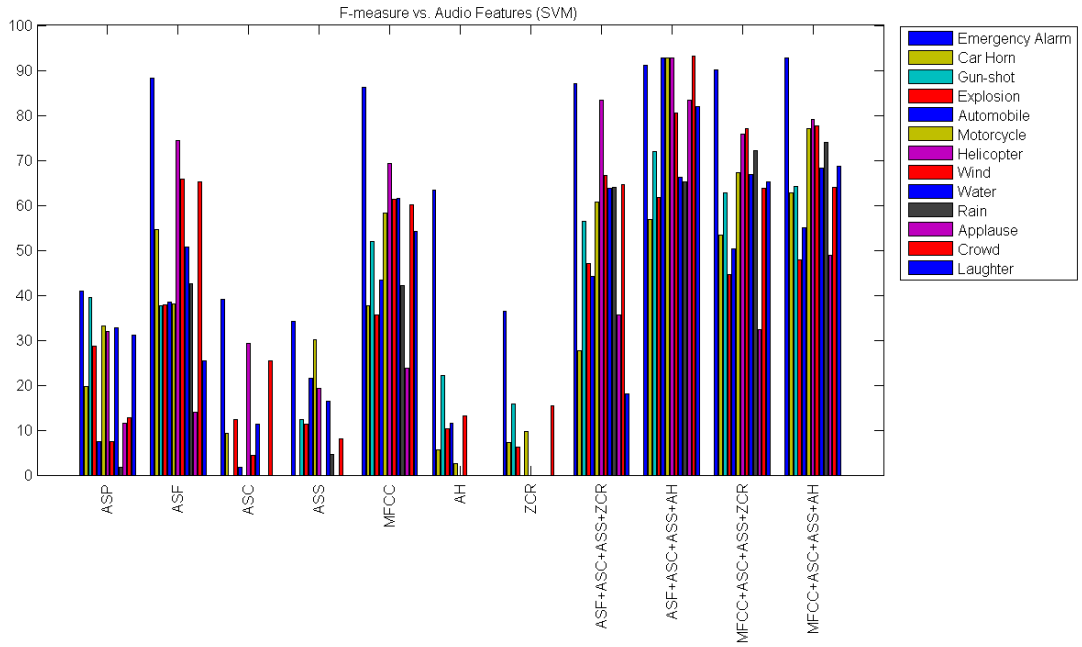

Figure 5.3: Results of SVM classification experiments.

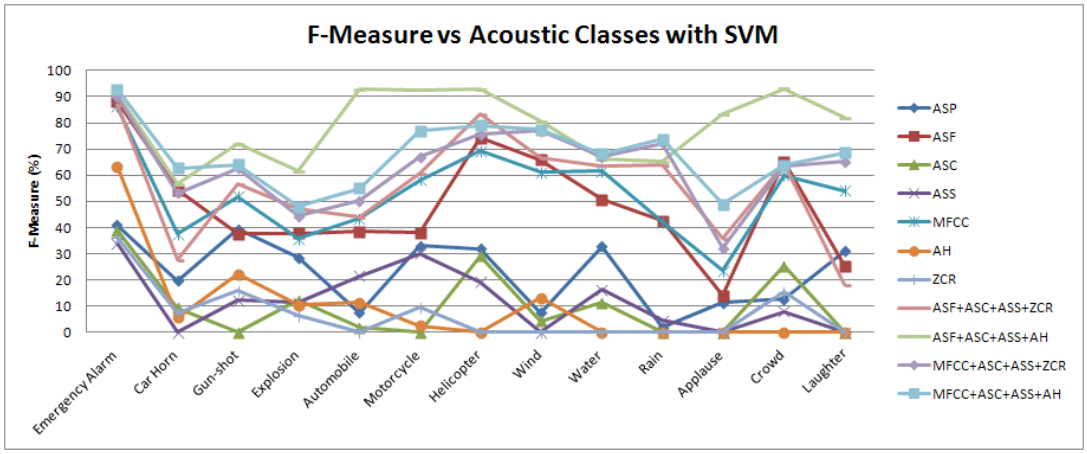

Figure 5.4: Performances of feature sets on acoustic classes using SVM classifier.

it is used standalone in the experiments. For this experiment, MFCC feature provides less F-measure value than HMM test. To compare, the results can be seen in Table A.16 and A.5.

For ASC, ASS, AH and ZCR features, SVM model is producing worse results than HMM does. The results for these tests can be seen in Tables A.14, A.15, A.17 and A.18, respectively.

Since audio features does not provide satisfactory results when used standalone, tests for four feature combinations ASF+ASC+ASS+AH, ASF+ASC+ASS+ZCR, MFCC+ASC+ASS+AH and MFCC+ASC+ASS+ZCR are repeated for SVM. ASF+ASC+ASS+AH combination is the best representative feature set of our data set with an average F-measure value of 80.6% (See in Table A.21). MFCC+ASC+ASS+AH combination also has considerably satisfactory F-measure of 69.4%. The results for ASF+ASC+ASS+ZCR and MFCC+ASC+ASS+ZCR are shown in Tables A.22 and A.20.

Experiments are summarized in with bar and line graphics in Figure 5.3 and Figure 5.4. Experimental results shows that usage of one dimensional features on HMM and SVM classifiers provides unsatisfactory results while feature combinations provide successful results for both classifiers.

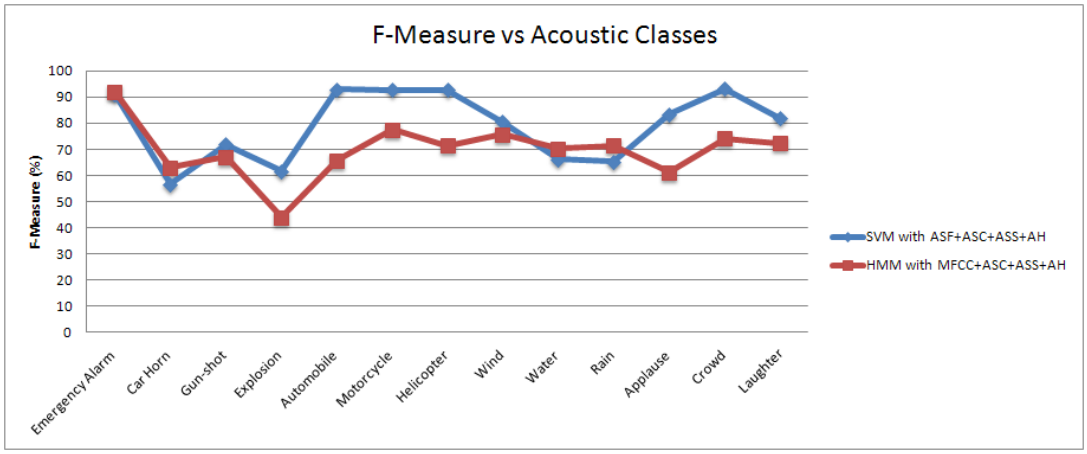

Figure 5.5: Comparison of HMM and SVM classification performances.

In Figure 5.5, avaregae F-measures of the classifiers with their best representative features are compared. SVM classifier with ASF+ASC+ASS+AH combination is performing better than HMM with MFCC+ASC+ASS+AH combination.

5.2 Experiments for Semantic Classes

5.2.1 Dataset Collection

Audio clips used in semantic classification are collected from internet videos [36]. The duration of each audio is differentiating between 20 seconds and 1 minute. Outdoor sounds contains town and city traffic, motor racing, traffic jam, sounds or cars and trucks, traffic sound effects, police and fire engine sounds. Nature sounds in the data base are consisted of sea cliff, wind ,rain, jungle forest and water stream sounds. For meeting class public house, sports crowd, interior crowd, business meeting and applause sounds are collected. Scenes from war movies and severity scenes collected from internet are used for violence sound collection. Durations of data set can be seen in Table 5.2.

Table 5.2: Overview of semantic data set.

	Duration
Outdoor	28 min 10 sec
Nature	27 min 45 sec
Meeting	20 min 13 sec
Violence	29 min 35 sec
TOTAL	1 hour 45 min 43 sec

5.2.2 GA Experiment

During the GA experiments the average F-measure is set as the fitness function and the threshold and the impact table is set as the chromosome. The values in the impact table are defined to be between 0.0 and 0.1 and iterations are started with a population containing 50 chromosomes. The optimized results are shown in Table 5.3 which reaches 87.4% average F-measure with the optimized threshold value of 51. As seen in impact table the effects of Emergency Alarm, Car Horn, Automobile, Motorcycle and Helicopter acoustic classes is over 0.94 for Outdoor semantic class. For Nature semantic class Wind, Water and Rain classes have the highest contribution, since they are the most suitable classes to exist in nature. Applause, Crowd and Laughter acoustic classes have the highest contributions to Meeting class among the others. Violence semantic class is impacted from Gun-shot and Explosion acoustic classes

mostly. It can be observed that all acoustic classes have some impact on all semantic classes. This sounds problematic because the reader may ask why helicopter has an impact value of 0.32 on Nature semantic class. The classification success of the proposed acoustic classification is 80.6% so the classifier can misclassify the segments or the test audio is consisted of segments which does not belong to any of the proposed acoustic classes. Confusion matrices of the given experiments can be seen in Appendix B.

Table 5.3: Impact table for semantic classes.

Impact Table for semantic classes													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Outdoor	0.99	0.95	0.39	0.38	0.99	0.96	0.94	0.10	0.39	0.18	0.20	0.13	0.03
Nature	0.09	0.15	0.00	0.04	0.05	0.15	0.32	0.84	0.68	0.63	0.16	0.36	0.07
Meeting	0.03	0.17	0.01	0.08	0.11	0.33	0.37	0.32	0.09	0.21	0.76	0.82	0.62
Violence	0.29	0.03	0.93	0.80	0.02	0.22	0.34	0.08	0.23	0.05	0.01	0.30	0.38

Table 5.4: Recall, precision and f-measure values for semantic classification with GA.

	Recall	Precision	F-measure
Outdoor	77.4 %	88.9 %	82.8 %
Nature	98.5 %	82.9 %	90.1 %
Meeting	89.0 %	98.0 %	93.3 %
Violence	81.9 %	82.5 %	82.2 %
Average	86.7 %	88.0 %	87.4 %

The proposed method for semantic classification is providing high F-measure value for the given dataset. But it is not sufficient for very short audio stream since the main concern is the density of the acoustic classes. For that reason a comparable classification is also performed as shown in the following section.

5.2.3 SVM Experiment

The results of acoustic classification for each segment is used as feature vectors of the SVM classifiers which is consisted of acoustic class id and classification accuracy. To be able to compare different approaches, WEKA [37] tool is used for the minor scale tests. This classification provided an avarage F-measure of 72.8% which is lower than the GA experiments result.

Table 5.5: Recall, precision and f-measure values for semantic classification with SVM.

	Recall	Precision	F-measure
Outdoor	51.5 %	80.9 %	62.8 %
Nature	94.4 %	77.0 %	84.8 %
Meeting	75.1 %	82.5 %	78.6 %
Violence	71.2 %	59.3 %	64.7 %
Average	73.4 %	74.8 %	72.8 %

In Figure 5.6 performances of GA and SVM are compared for semantic classes. The proposed method with GA performs better than SVM for all semantic classes.

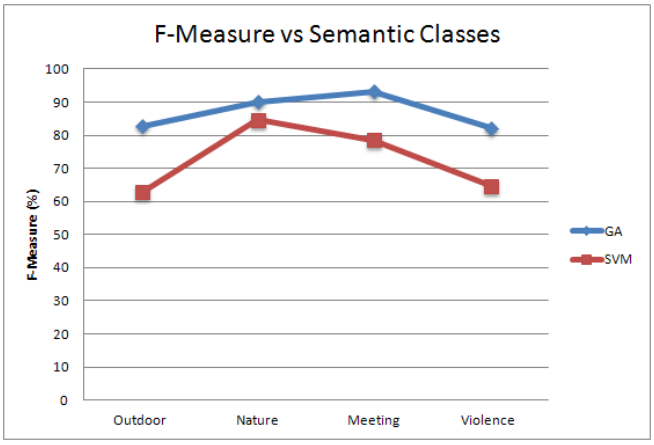

Figure 5.6: Comparison of GA and SVM classification performances.

5.3 Experiment for Content Retrieval

5.3.1 Dataset Collection

All audio clips used in content retrieval experiment are collected from acoustic dataset in Table 5.1. One minute and 15 seconds of audio clips for each acoustic class are selected for search space and query audio, respectively. In total, 13 minutes of audio clip is used for search space whereas 3 minutes 15 seconds of audio clip is used for the tests.

For keyword-based queries, an additional experiment is not considered, since the retrieval success of these queries are bounded with the classification results.

5.3.2 QBE Experiment

Table 5.6: Accuracy values for QBE retrieval using ASF, ASC, ASS and AH feature combination.

QBE Retrival	
	Accuracy
Emergency Alarm	31.2 %
Car Horn	55.7 %
Gun-shot	66.2 %
Explosion	61.6 %
Automobile	57.9 %
Motorcycle	60.9 %
Helicopter	41.0 %
Wind	36.3 %
Water	39.0 %
Rain	70.3 %
Applause	36.6 %
Crowd	32.6 %
Laughter	48.0 %
Average	49.0 %

QBE retrieval experiment is performed with categorization based approach using acoustically labelled segments. For each acoustic category, the most similar five segments are retrieved and accuracy is calculated as a ratio of number of correctly retrieved segments over number of total retrieved segments.

CHAPTER 6

USER INTERFACE

In order to present the capabilities of the proposed system, an environmental sound classification application is developed in the context of this thesis study. Details of the usage are given in the following sections.

6.1 Main Window

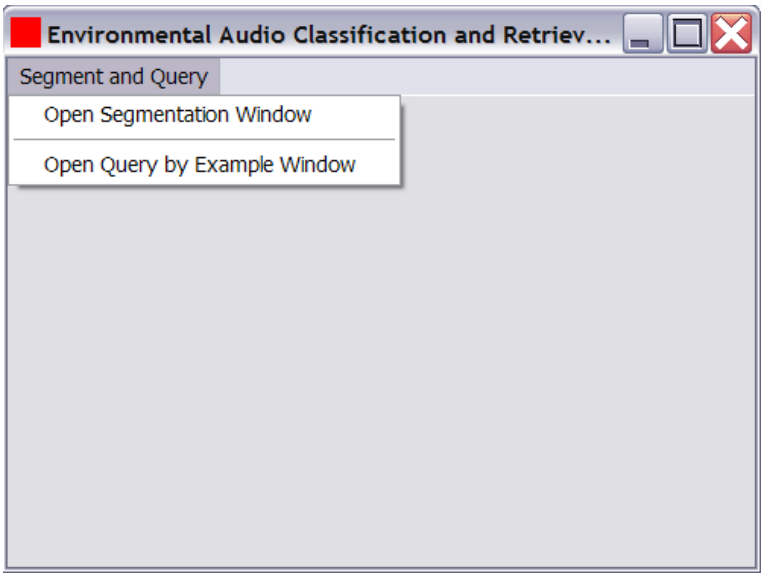

Figure 6.1: Screenshot of the main window.

The screenshot of the main window is shown in Figure 6.1. User can choose to open segmentation and query by example windows.

6.2 Segmentation Window

This window provides the interface to classify an audio clip and to run temporal and keyword queries. User can select the audio file from “Select Media” button shown in Figure 6.2. After the selection, audio play window appears under the selected file box. By clicking on the “SEGMENT” button, user starts the classification and system saves the classification results in a file under the directory of the given audio. Results are automatically displayed in the result table after the classification. Previous results can be loaded and displayed in the result table using “Select Result File” button.

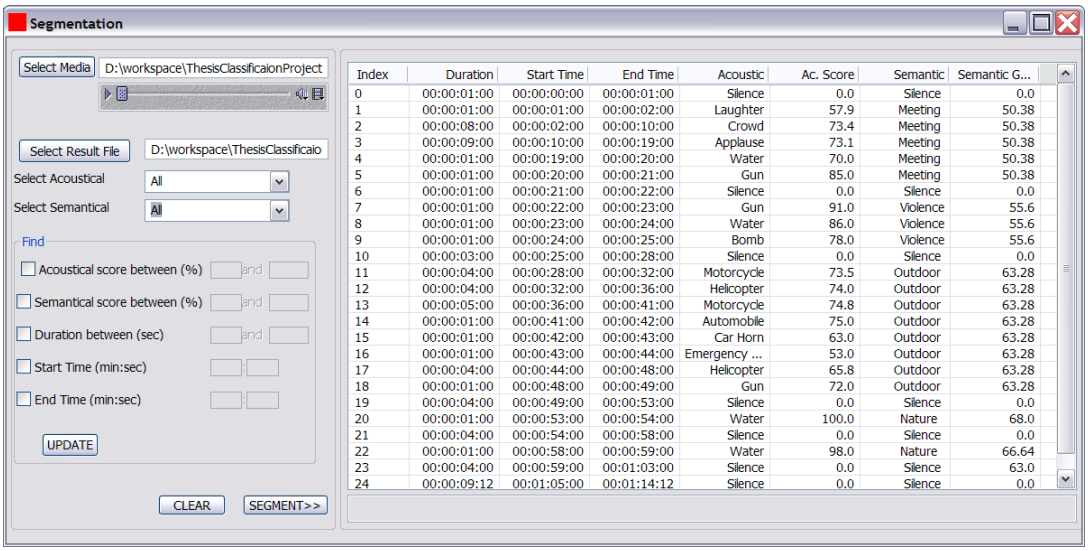

Figure 6.2: Screenshot of segmentation window after segmentation.

Using “Select Acoustical” and “Select Semantical” drop boxes, keywords are selected for queries of acoustic and semantic classes. In “Find” pane, acoustical and semantic scores, duration, start and end times of the query can be set. “UPDATE” button updates the table after the “Find” pane settings. In Figure 6.3 an example query, “Retrieve all possible *emergency alarm* segments with an acoustic possibility degree between *70%* and *80%* and semantically labelled as *outdoor* with a semantic possibility degree between *50%* and *70%*, between the 28^{th} and 49^{th} seconds and with a length between *0* and *5* seconds.”, is illustrated.

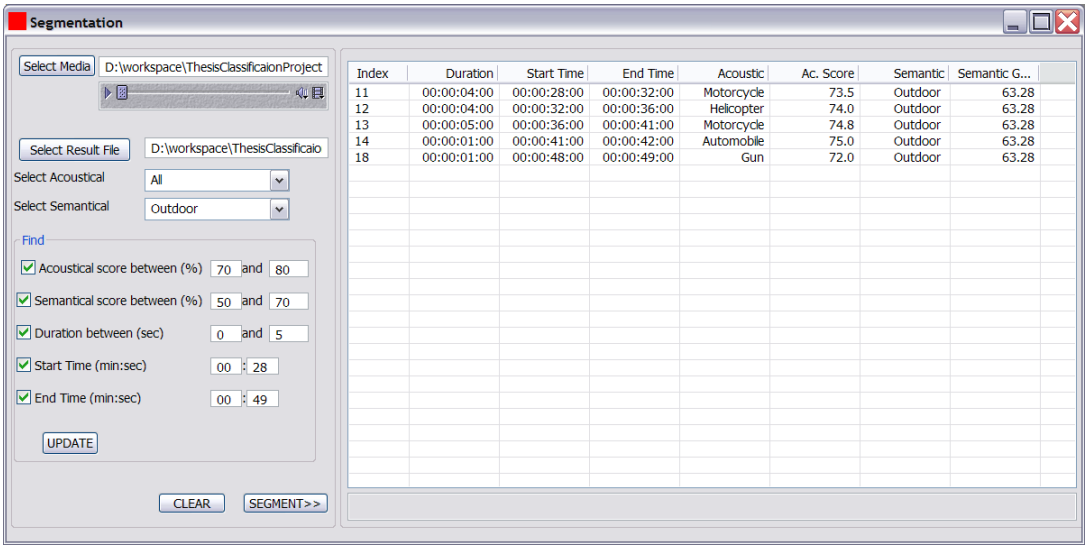

Figure 6.3: Screenshot of segmentation window with temporal and keyword-based query example.

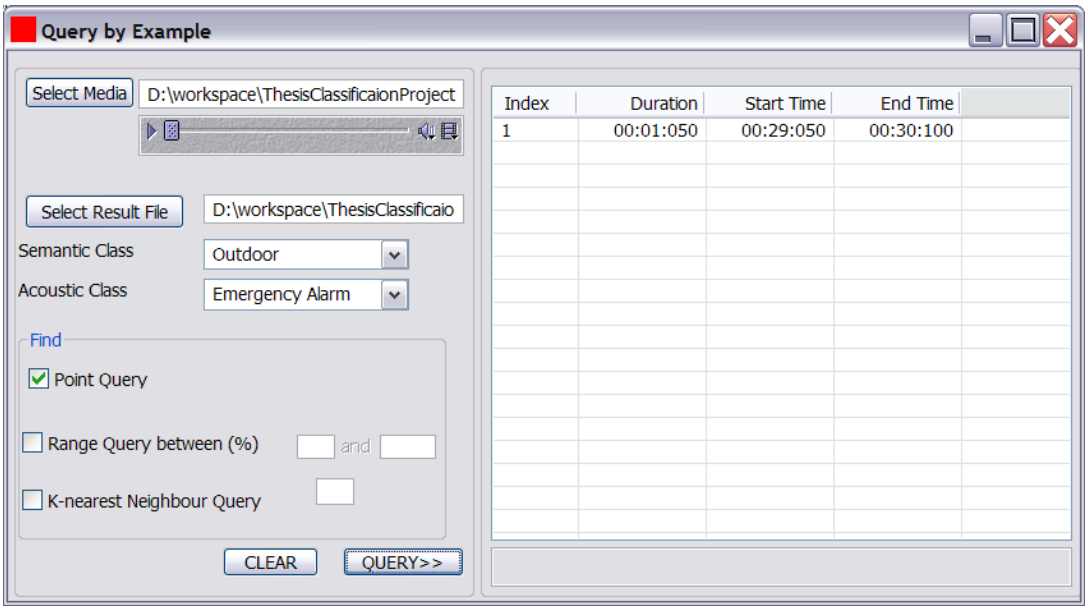

Figure 6.4: Screenshot of QBE window with point query example.

6.3 Query by Example Window

This window provides the interface to run QBE queries. Point, Range and k-Nearest Neighbour queries are supported in this window. User can select the audio file from "Select Media" button shown in Figure 6.4. User should also provide a result file of the search space using "Select Result File" button. Using the drop boxes, user can limit the search space with selected acoustical and semantic classes. Point, Range and kNN queries can be selected in "Find" pane and results are shown in the result table. In the Figure 6.4, Figure 6.5 and Figure 6.6, examples of QBE queries are shown.

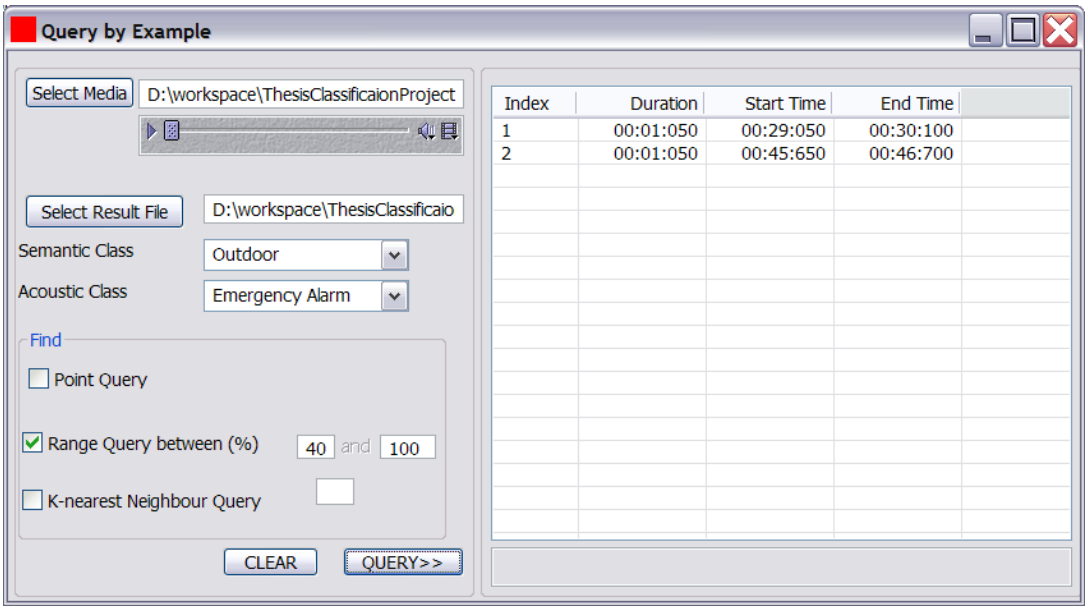

Figure 6.5: Screenshot of QBE window with range query example.

6.4 Implementation

This application is implemented using Java programming language. FFmpeg [38] multimedia framework is used in order to convert the given media file (wav, flv, mp3, avi formats) to the accepted format (mono wav, 44100 kbps, pcm 16 bit little endian encoding). For feature extraction, MPEG-7 Audio Encoder Project [27] is employed. When the input audio is selected using "Select Media" button, conversation and feature extraction of the given audio is done

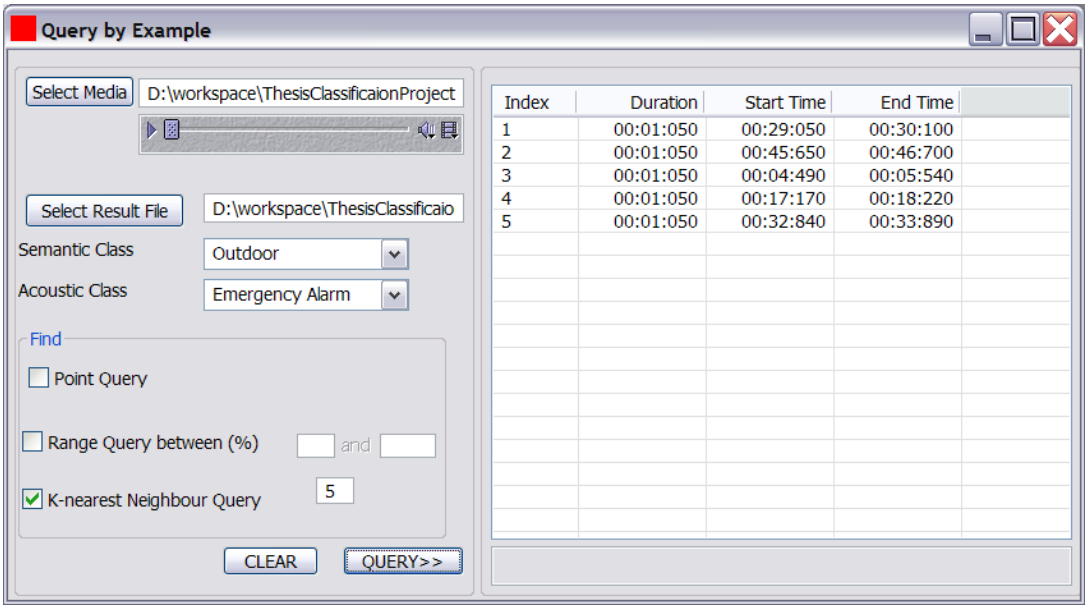

Figure 6.6: Screenshot of QBE window with KNN query example.

automatically. “SEGMENT” button starts the acoustical and semantic classifications. During the acoustic classification LIBSVM [31] is utilized. For semantic classification impact table values are used (see Section 4.1.4).

This application has query interfaces but no specific query languages are defined. The query should be created using the buttons and panes provided by the interface. Classification results are loaded from the result file to the result table in application windows. For keyword and temporal queries, system basically filters the selected options and show them in the result table. Therefore, the keyword and temporal query performance is strongly bounded with classification performance.

CHAPTER 7

CONCLUSION

A novel environmental sound classification system is proposed in this study. Environmental sounds are classified into *acoustic classes* and *semantic classes* in a two-stage approach in which the results of acoustic classification are the inputs of the semantic classification.

In order to discover the best classification performance, 22 experiments are performed using MPEG-7 audio features, MFCC and ZCR -in a standalone and combined manner- with HMM and SVM classifiers. These experiments shows that the best representative feature is MPEG-7 ASF, ASC, ASS and AH feature combination applied on SVM classifier with the average F-measure of 80.6%. The second best classification result of 70.6% is obtained from the experiment of MFCC, MPEG-7 ASC, ASS and AH combination with HMM classifier.

In acoustic classification stage, one-second audio segments is classified into selected *acoustic classes* such as emergency alarm, car horn, gun-shot, explosion, automobile, motorcycle, helicopter, wind, water, rain, applause, crowd and laughter. These acoustically classified segments are used as input for the semantic classification of outdoor, nature, meeting and violence *semantic classes*. Instead of model training, a new approach is proposed resulting 87.4% F-measure. To have ground truth for the proposed approach an SVM classification is also experimented but the average F-measure is calculated as 72.8% which is far-behind the proposed approach.

In environmental sound classification, better results are reported in [7, 8, 10, 11, 21, 22]. However, relatively complex and more audio classes are considered in this study. For instance, the following environmental sounds are considerably similar and distinguishing between them is even difficult for human perception:

- gun-shot and explosion,
- motorcycle, helicopter and automobile,
- water, rain and wind,
- crowd and laughter

We believe that, as the number of considered classes increases, system becomes relatively complex and the success of the classifier decreases. A solution to this problem is to extend the data set for similar sound classes.

In order to retrieve relevant data from the classification results, several types of queries are supported. User can query with keyword, temporal information and similar audio. QBE retrieval experiment shows that accuracy for all classes is 49.0%. These capabilities are implemented on the environmental audio classification and retrieval tool. This tool provides an efficient environment for the user in order to classify audio samples and retrieve the content.

This study contributes an initial idea for two-stage classification providing satisfying results. To improve this idea further, following items are counted as future work:

- The data sets for acoustic and semantic classifications can be enhanced in order to increase the model training success.
- New *semantic classes* and *acoustic classes* might be classified to extend the coverage of the system.
- The experiments for acoustic classification can be repeated with other classification techniques.
- A combination of HMM and SVM classifiers can be introduced for acoustic classification in order to achieve better results.

REFERENCES

[1] E. Dogan, M. Sert, and A. Yazici. Content-based classification and segmentation of mixed-type audio by using mpeg-7 features. In *Advances in Multimedia, 2009. MMEDIA'09. First International Conference on*, pages 152–157. IEEE, 2009.

[2] W.H. Liao and Y.K. Lin. Classification of non-speech human sounds: feature selection and snoring sound analysis. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 2695–2700. IEEE, 2009.

[3] H.G. Kim, N. Moreau, and T. Sikora. *MPEG-7 audio and beyond*. Wiley Online Library, 2005.

[4] T. Zhang and C.C.J. Kuo. Hierarchical system for content-based audio classification and retrieval. In *Conference on Multimedia Storage and Archiving Systems III, SPIE*, volume 3527, pages 398–409, 1998.

[5] L. Lu, H. Jiang, and H.J. Zhang. A robust audio classification and segmentation method. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 203–211. ACM, 2001.

[6] F. Beritelli and R. Grasso. A pattern recognition system for environmental sound classification based on mfccs and neural networks. In *Signal Processing and Communication Systems, 2008. ICSPCS 2008. 2nd International Conference on*, pages 1–4. IEEE, 2008.

[7] G. Muhammad, Y.A. Alotaibi, M. Alsulaiman, and M.N. Huda. Environment recognition using selected mpeg-7 audio features and mel-frequency cepstral coefficients. In *Digital Telecommunications (ICDT), 2010 Fifth International Conference on*, pages 11–16. IEEE, 2010.

[8] I. Feki, A. Ben Ammar, and A.M. Alimi. Audio stream analysis for environmental sound classification. In *Multimedia Computing and Systems (ICMCS), 2011 International Conference on*, pages 1–6. IEEE, 2011.

[9] J.F. Wang, J.C. Wang, T.H. Huang, and C.S. Hsu. Home environmental sound recognition based on mpeg-7 features. In *Micro-NanoMechatronics and Human Science, 2003 IEEE International Symposium on*, volume 2, pages 682–685. IEEE, 2003.

[10] R. Cai, L. Lu, H.J. Zhang, and L.H. Cai. Highlight sound effects detection in audio stream. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–37. IEEE, 2003.

[11] S. Chu, S. Narayanan, and C.C.J. Kuo. Environmental sound recognition with time–frequency audio features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1142–1158, 2009.

[12] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. In *Proc. of the X European Signal Processing Conference*, 2000.

[13] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano. Environmental sound source identification based on hidden markov model for robust speech recognition. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[14] S.H. Shin, T. Hashimoto, and S. Hatano. Automatic detection system for cough sounds as a symptom of abnormal health condition. *Information Technology in Biomedicine, IEEE Transactions on*, 13(4):486–493, 2009.

[15] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang. Comparing mfcc and mpeg-7 audio features for feature extraction, maximum likelihood hmm and entropic prior hmm for sports audio classification. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 5, pages V–628. IEEE, 2003.

[16] H.G. Kim, N. Moreau, and T. Sikora. Audio classification based on mpeg-7 spectral basis representations. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):716–725, 2004.

[17] Y. Song, W.H. Wang, and F.J. Guo. Feature extraction and classification for audio information in news video. In *Wavelet Analysis and Pattern Recognition, 2009. ICWAPR 2009. International Conference on*, pages 43–46. IEEE, 2009.

[18] L. Lu, F. Ge, Q. Zhao, and Y. Yan. Detecting cheering events in sports games. In *Education Technology and Computer (ICETC), 2010 2nd International Conference on*, volume 1, pages V1–223. IEEE, 2010.

[19] R. Dong, D. Hermann, E. Cornu, and E. Chau. Low-power implementation of an hmm-based sound environment classification algorithm for hearing aid application. In *Proc. EUSIPCO*, 2007.

[20] MA Guvensan and ZC Taysi. Environmental sound classification for recognition of house appliances. In *Signal Processing and Communications Applications Conference (SIU), 2010 IEEE 18th*, pages 431–434. IEEE, 2010.

[21] W.H. Choi, S.I. Kim, M.S. Keum, W. Han, H. Ko, and D.K. Han. Acoustic and visual signal based context awareness system for mobile application. In *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, pages 627–628. IEEE, 2011.

[22] R. Gerard, J. Jordi, S. Mattia, H. Perfecto, and S. Xavier. Ecological acoustics perspective for content-based retrieval of environmental sounds. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, 2011.

[23] Y. Wang, Z. Liu, and J.C. Huang. Multimedia content analysis-using both audio and visual clues. *Signal Processing Magazine, IEEE*, 17(6):12–36, 2000.

[24] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[25] V. Vapnik. The support vector method of function estimation. *Nonlinear Modeling: Advanced Black-Box Techniques*, 55:86, 1998.

[26] J.R. Anderson, R.S. Michalski, R.S. Michalski, J.G. Carbonell, and T.M. Mitchell. *Machine learning: An artificial intelligence approach*, volume 2. Morgan Kaufmann, 1986.

[27] H. Crysandt. Mpeg-7 audio encoder project. http://mpeg7audioenc.sourceforge.net/index.html, last accessed date: 20.09.2012.

[28] M. Stanley. Auditory toolbox. https://engineering.purdue.edu/ malcolm/interval/1998-010/, last accessed date: 20.09.2012.

[29] S. Kiranyaz, A.F. Qureshi, and M. Gabbouj. A generic audio classification and segmentation approach for multimedia indexing and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):1062–1081, 2006.

[30] J.M. François. Jahmm - hidden markov model (hmm). http://www.run.montefiore.ulg.ac.be/ francois/software/jahmm/, last accessed date: 20.09.2012.

[31] C.C. Chang and C.J. Lin. Libsvm – a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm/, last accessed date: 20.09.2012.

[32] T. Pohlmann. Jgap - java genetic algorithms package. http://jgap.sourceforge.net/, last accessed date: 20.09.2012.

[33] E. Dogan. Content-based audio management and retrieval system for news broadcasts. 2009.

[34] FreeSound. http://www.freesound.org, last accessed date: 20.09.2012.

[35] SoundBible. http://www.soundbible.org, last accessed date: 20.09.2012.

[36] YouTube. http://www.youtube.com, last accessed date: 20.09.2012.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[38] F. Bellard. Ffmpeg. http://ffmpeg.org//, last accessed date: 20.09.2012.

APPENDIX A

CLASSIFICATION RESULTS AND CONFUSION MATRICES OF ACOUSTICAL EXPERIMENTS

A.1 Classification Results

Table A.1: Recall, precision and f-measure values for ASP feature with HMM classification.

ASP feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	37.3 %	48.2 %	42.1 %
Car Horn	50.0 %	27.2 %	35.2 %
Gun-shot	36.0 %	81.3 %	50.0 %
Explosion	39.0 %	28.4 %	32.8 %
Automobile	9.0 %	7.0 %	7.9 %
Motorcycle	31.1 %	28.1 %	29.5 %
Helicopter	69.4 %	31.7 %	43.6 %
Wind	25.0 %	29.8 %	27.2 %
Water	13.1 %	30.2 %	18.3 %
Rain	53.2 %	39.8 %	45.6 %
Applause	31.6 %	19.1 %	23.8 %
Crowd	72.5 %	43.2 %	54.2 %
Laughter	18.9 %	41.1 %	25.9 %
Average	37.4 %	35.0 %	36.2 %

Table A.2: Recall, precision and f-measure values for ASF feature with HMM classification.

ASF feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	81.0 %	93.9 %	87.0 %
Car Horn	30.0 %	60.0 %	40.0 %
Gun-shot	26.3 %	56.9 %	35.9 %
Explosion	28.1 %	15.7 %	20.2 %
Automobile	32.4 %	43.8 %	37.3 %
Motorcycle	45.0 %	54.4 %	49.3 %
Helicopter	64.4 %	81.7 %	72.0 %
Wind	58.6 %	79.0 %	67.3 %
Water	59.8 %	44.1 %	50.8 %
Rain	82.2 %	41.7 %	55.3 %
Applause	40.0 %	60.0 %	48.0 %
Crowd	90.3 %	50.0 %	64.3 %
Laughter	74.3 %	55.0 %	63.2 %
Average	54.8 %	56.6 %	55.7 %

Table A.3: Recall, precision and f-measure values for ASC feature with HMM classification.

ASC feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	43.6 %	37.9 %	40.5 %
Car Horn	16.6 %	8.6 %	11.3 %
Gun-shot	25.2 %	69.0 %	37.0 %
Explosion	21.8 %	10.8 %	14.5 %
Automobile	24.6 %	19.7 %	21.9 %
Motorcycle	30.3 %	38.9 %	34.1 %
Helicopter	41.5 %	32.6 %	36.5 %
Wind	30.4 %	32.1 %	31.2 %
Water	4.0 %	13.7 %	6.2 %
Rain	33.6 %	40.4 %	36.7 %
Applause	23.7 %	11.1 %	15.1 %
Crowd	83.8 %	22.1 %	35.0 %
Laughter	0.0 %	0.0 %	0.0 %
Average	29.2 %	25.9 %	27.4 %

Table A.4: Recall, precision and f-measure values for ASS feature with HMM classification.

ASS feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	30.8 %	41.1 %	35.2 %
Car Horn	3.3 %	6.2 %	4.3 %
Gun-shot	13.9 %	38.1 %	20.4 %
Explosion	9.3 %	6.0 %	7.3 %
Automobile	22.0 %	13.6 %	16.8 %
Motorcycle	38.5 %	23.7 %	29.3 %
Helicopter	17.7 %	13.3 %	15.2 %
Wind	29.5 %	24.4 %	26.7 %
Water	5.0 %	16.9 %	7.8 %
Rain	46.7 %	37.5 %	41.6 %
Applause	28.8 %	13.3 %	18.2 %
Crowd	11.2 %	7.0 %	8.6 %
Laughter	14.8 %	12.6 %	13.6 %
Average	20.9 %	19.5 %	20.2 %

Table A.5: Recall, precision and f-measure values for MFCC feature with HMM classification.

MFCC feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	77.2 %	96.6 %	85.8 %
Car Horn	33.3 %	31.2 %	32.2 %
Gun-shot	51.1 %	83.9 %	63.5 %
Explosion	68.7 %	30.9 %	42.7 %
Automobile	53.2 %	53.9 %	53.5 %
Motorcycle	63.1 %	80.2 %	70.6 %
Helicopter	84.7 %	78.1 %	81.3 %
Wind	60.8 %	82.3 %	70.0 %
Water	77.1 %	60.0 %	67.5 %
Rain	64.4 %	61.6 %	63.0 %
Applause	65.0 %	61.9 %	63.4 %
Crowd	93.5 %	51.7 %	66.6 %
Laughter	64.8 %	69.5 %	67.1 %
Average	65.9 %	64.8 %	65.3 %

Table A.6: Recall, precision and f-measure values for AH feature with HMM classification.

AH feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	54.5 %	56.3 %	55.4 %
Car Horn	40.0 %	16.2 %	23.0 %
Gun-shot	12.8 %	44.7 %	19.9 %
Explosion	17.1 %	12.6 %	14.5 %
Automobile	11.6 %	5.9 %	7.8 %
Motorcycle	13.1 %	21.9 %	16.4 %
Helicopter	36.7 %	20.2 %	26.1 %
Wind	12.1 %	20.5 %	15.3 %
Water	18.8 %	42.0 %	26.0 %
Rain	42.9 %	46.4 %	44.6 %
Applause	20.3 %	9.6 %	13.1 %
Crowd	51.6 %	15.5 %	23.8 %
Laughter	0.0 %	0.0 %	0.0 %
Average	25.5 %	24.0 %	24.7 %

Table A.7: Recall, precision and f-measure values for ZCR feature with HMM classification.

ZCR feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	47.8 %	40.3 %	43.7 %
Car Horn	13.3 %	21.0 %	16.3 %
Gun-shot	18.1 %	65.7 %	28.4 %
Explosion	4.6 %	8.3 %	6.0 %
Automobile	35.0 %	31.3 %	33.1 %
Motorcycle	16.3 %	14.5 %	15.4 %
Helicopter	85.5 %	33.5 %	48.2 %
Wind	16.5 %	25.3 %	20.0 %
Water	42.6 %	34.8 %	38.3 %
Rain	34.5 %	38.9 %	36.6 %
Applause	13.5 %	25.8 %	17.7 %
Crowd	33.8 %	14.3 %	20.1 %
Laughter	1.3 %	6.6 %	2.2 %
Average	27.9 %	27.7 %	27.8 %

Table A.8: Recall, precision and f-measure values for ASF, ASC, ASS and ZCR feature combination with HMM classification.

ASF+ASC+ASS+ZCR feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	86.8 %	85.9 %	86.3 %
Car Horn	30.0 %	60.0 %	40.0 %
Gun-shot	56.3 %	76.5 %	64.9 %
Explosion	53.1 %	33.0 %	40.7 %
Automobile	54.5 %	56.7 %	55.6 %
Motorcycle	62.2 %	80.8 %	70.3 %
Helicopter	85.5 %	76.5 %	80.8 %
Wind	56.0 %	73.8 %	63.7 %
Water	70.5 %	57.4 %	63.3 %
Rain	62.6 %	57.7 %	60.0 %
Applause	35.0 %	51.2 %	41.5 %
Crowd	93.5 %	46.0 %	61.7 %
Laughter	52.7 %	60.9 %	56.5 %
Average	61.4 %	62.8 %	62.1 %

Table A.9: Recall, precision and f-measure values for ASF, ASC, ASS and AH feature combination with HMM classification.

ASF+ASC+ASS+AH feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	94.7 %	94.2 %	94.4 %
Car Horn	16.7 %	100.0 %	28.6 %
Gun-shot	33.1 %	74.6 %	45.8 %
Explosion	48.4 %	30.1 %	37.1 %
Automobile	79.2 %	76.3 %	77.7 %
Motorcycle	75.4 %	84.4 %	79.6 %
Helicopter	85.6 %	92.7 %	88.9 %
Wind	66.1 %	76.0 %	70.7 %
Water	70.1 %	56.9 %	63.0 %
Rain	77.6 %	51.2 %	61.7 %
Applause	67.8 %	57.9 %	62.5 %
Crowd	98.4 %	55.9 %	71.3 %
Laughter	74.3 %	67.1 %	70.5 %
Average	68.2 %	70.5 %	69.4 %

Table A.10: Recall, precision and f-measure values for MFCC, ASC, ASS and ZCR feature combination with HMM classification.

MFCC+ASC+ASS+ZCR feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	88.9 %	90.8 %	89.8 %
Car Horn	26.6 %	32.0 %	29.0 %
Gun-shot	46.9 %	74.8 %	57.7 %
Explosion	70.3 %	32.1 %	44.1 %
Automobile	41.5 %	47.7 %	44.4 %
Motorcycle	76.2 %	72.0 %	74.1 %
Helicopter	77.1 %	70.0 %	73.3 %
Wind	68.9 %	91.9 %	78.8 %
Water	75.1 %	61.1 %	67.4 %
Rain	69.1 %	71.1 %	70.1 %
Applause	53.3 %	56.1 %	54.7 %
Crowd	85.4 %	62.3 %	72.1 %
Laughter	55.4 %	64.0 %	59.4 %
Average	64.2 %	63.5 %	63.9 %

Table A.11: Recall, precision and f-measure values for MFCC, ASC, ASS and AH feature combination with HMM classification.

MFCC+ASC+ASS+AH feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	90.4 %	93.9 %	92.1 %
Car Horn	60.0 %	66.6 %	63.1 %
Gun-shot	59.2 %	77.3 %	67.0 %
Explosion	67.1 %	32.8 %	44.1 %
Automobile	59.7 %	73.0 %	65.7 %
Motorcycle	73.7 %	81.8 %	77.5 %
Helicopter	70.3 %	72.8 %	71.5 %
Wind	70.4 %	81.8 %	75.7 %
Water	70.5 %	69.8 %	70.2 %
Rain	77.5 %	66.4 %	71.5 %
Applause	62.7 %	59.6 %	61.1 %
Crowd	90.3 %	62.9 %	74.1 %
Laughter	72.9 %	72.0 %	72.4 %
Average	71.1 %	70.0 %	70.6 %

Table A.12: Recall, precision and f-measure values for ASP feature with SVM classification.

ASP feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	36.3 %	47.2 %	41.0 %
Car Horn	23.3 %	17.0 %	19.7 %
Gun-shot	27.8 %	68.0 %	39.5 %
Explosion	39.0 %	22.7 %	28.7 %
Automobile	6.4 %	8.9 %	7.5 %
Motorcycle	45.9 %	25.9 %	33.1 %
Helicopter	37.2 %	28.0 %	31.9 %
Wind	4.3 %	31.2 %	7.5 %
Water	57.8 %	22.8 %	32.8 %
Rain	0.9 %	100.0 %	1.8 %
Applause	11.6 %	11.4 %	11.5 %
Crowd	8.0 %	31.2 %	12.8 %
Laughter	22.9 %	48.5 %	31.1 %
Average	24.7 %	35.6 %	29.2 %

Table A.13: Recall, precision and f-measure values for ASF feature with SVM classification.

ASF feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	80.0 %	98.7 %	88.3 %
Car Horn	40.0 %	85.7 %	54.5 %
Gun-shot	25.0 %	75.2 %	37.6 %
Explosion	39.0 %	36.7 %	37.8 %
Automobile	28.5 %	59.4 %	38.5 %
Motorcycle	29.5 %	53.7 %	38.0 %
Helicopter	70.3 %	79.0 %	74.4 %
Wind	54.3 %	84.0 %	65.9 %
Water	76.6 %	38.0 %	50.8 %
Rain	50.4 %	36.7 %	42.5 %
Applause	8.3 %	45.4 %	14.0 %
Crowd	77.4 %	56.4 %	65.3 %
Laughter	16.2 %	60.0 %	25.5 %
Average	45.8 %	62.2 %	52.8 %

Table A.14: Recall, precision and f-measure values for ASC feature with SVM classification.

ASC feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	48.4 %	32.8 %	39.1 %
Car Horn	33.3 %	5.4 %	9.3 %
Gun-shot	0.0 %	0.0 %	0.0 %
Explosion	32.8 %	7.6 %	12.3 %
Automobile	1.2 %	2.7 %	1.7 %
Motorcycle	0.0 %	0.0 %	0.0 %
Helicopter	33.0 %	26.1 %	29.2 %
Wind	5.2 %	3.7 %	4.3 %
Water	9.6 %	13.9 %	11.4 %
Rain	0.0 %	0.0 %	0.0 %
Applause	0.0 %	0.0 %	0.0 %
Crowd	64.5 %	15.8 %	25.4 %
Laughter	0.0 %	0.0 %	0.0 %
Average	17.5 %	8.3 %	11.3 %

Table A.15: Recall, precision and f-measure values for ASS feature with SVM classification.

ASS feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	26.0 %	49.4 %	34.1 %
Car Horn	0.0 %	0.0 %	0.0 %
Gun-shot	7.1 %	48.7 %	12.4 %
Explosion	20.3 %	7.9 %	11.4 %
Automobile	44.1 %	14.2 %	21.5 %
Motorcycle	35.2 %	26.3 %	30.1 %
Helicopter	31.3 %	14.0 %	19.3 %
Wind	0.0 %	0.0 %	0.0 %
Water	19.7 %	14.0 %	16.4 %
Rain	4.6 %	4.6 %	4.6 %
Applause	0.0 %	0.0 %	0.0 %
Crowd	11.2 %	6.1 %	8.0 %
Laughter	0.0 %	0.0 %	0.0 %
Average	15.3 %	14.2 %	14.8 %

Table A.16: Recall, precision and f-measure values for MFCC feature with SVM classification.

MFCC feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	82.0 %	91.1 %	86.3 %
Car Horn	53.3 %	29.0 %	37.6 %
Gun-shot	40.9 %	70.7 %	51.9 %
Explosion	50.0 %	27.8 %	35.7 %
Automobile	42.8 %	44.0 %	43.4 %
Motorcycle	47.5 %	75.3 %	58.2 %
Helicopter	73.7 %	65.4 %	69.3 %
Wind	53.0 %	72.6 %	61.3 %
Water	83.7 %	48.8 %	61.6 %
Rain	28.9 %	77.5 %	42.1 %
Applause	20.0 %	29.2 %	23.7 %
Crowd	93.5 %	44.2 %	60.1 %
Laughter	51.3 %	57.5 %	54.2 %
Average	55.4 %	56.4 %	55.9 %

Table A.17: Recall, precision and f-measure values for AH feature with SVM classification.

AH feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	62.5 %	64.2 %	63.4 %
Car Horn	3.3 %	20.0 %	5.7 %
Gun-shot	17.3 %	30.4 %	22.1 %
Explosion	87.5 %	5.5 %	10.3 %
Automobile	9.0 %	15.9 %	11.5 %
Motorcycle	1.6 %	5.2 %	2.5 %
Helicopter	0.0 %	N %	0.0 %
Wind	8.6 %	27.7 %	13.2 %
Water	0.0 %	0.0 %	0.0 %
Rain	0.0 %	0.0 %	0.0 %
Applause	0.0 %	0.0 %	0.0 %
Crowd	0.0 %	0.0 %	0.0 %
Laughter	0.0 %	0.0 %	0.0 %
Average	14.6 %	13.0 %	13.7 %

Table A.18: Recall, precision and f-measure values for ZCR feature with SVM classification.

ZCR feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	67.5 %	25.0 %	36.4 %
Car Horn	36.6 %	4.0 %	7.3 %
Gun-shot	13.9 %	18.6 %	15.9 %
Explosion	12.5 %	4.2 %	6.3 %
Automobile	0.0 %	0.0 %	0.0 %
Motorcycle	9.0 %	10.4 %	9.6 %
Helicopter	0.0 %	0.0 %	0.0 %
Wind	0.0 %	0.0 %	0.0 %
Water	0.0 %	0.0 %	0.0 %
Rain	0.0 %	0.0 %	0.0 %
Applause	0.0 %	0.0 %	0.0 %
Crowd	33.8 %	10.0 %	15.4 %
Laughter	0.0 %	0.0 %	0.0 %
Average	13.3 %	5.5 %	7.8 %

Table A.19: Recall, precision and f-measure values for MFCC, ASC, ASS and AH feature combination with SVM classification.

MFCC+ASC+ASS+AH feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	93.0 %	92.5 %	92.8 %
Car Horn	53.3 %	76.1 %	62.7 %
Gun-shot	55.4 %	76.1 %	64.1 %
Explosion	64.0 %	38.3 %	47.9 %
Automobile	49.3 %	62.2 %	55.0 %
Motorcycle	77.0 %	77.0 %	77.0 %
Helicopter	81.3 %	76.8 %	79.0 %
Wind	67.8 %	90.6 %	77.6 %
Water	78.1 %	60.6 %	68.2 %
Rain	66.3 %	83.5 %	73.9 %
Applause	37.2 %	70.9 %	48.8 %
Crowd	90.3 %	49.5 %	64.0 %
Laughter	72.9 %	65.0 %	68.7 %
Average	68.2 %	70.7 %	69.4 %

Table A.20: Recall, precision and f-measure values for MFCC, ASC, ASS and ZCR feature combination with SVM classification.

MFCC+ASC+ASS+ZCR feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	87.8 %	92.2 %	90.0 %
Car Horn	53.3 %	53.3 %	53.3 %
Gun-shot	54.1 %	74.6 %	62.7 %
Explosion	67.1 %	33.3 %	44.5 %
Automobile	53.2 %	47.6 %	50.3 %
Motorcycle	61.4 %	74.2 %	67.2 %
Helicopter	74.5 %	77.1 %	75.8 %
Wind	65.5 %	93.8 %	77.1 %
Water	74.1 %	61.0 %	66.9 %
Rain	64.4 %	82.1 %	72.2 %
Applause	26.6 %	41.0 %	32.3 %
Crowd	93.5 %	48.3 %	63.7 %
Laughter	67.5 %	63.2 %	65.3 %
Average	64.9 %	64.7 %	64.8 %

Table A.21: Recall, precision and f-measure values for ASF, ASC, ASS and AH feature combination with SVM classification.

ASF+ASC+ASS+AH feature with SVM			
	Recall	Precision	FValue
Emergency Alarm	84.3 %	99.4 %	91.1 %
Car Horn	74.2 %	46.0 %	56.8 %
Gun-shot	61.2 %	87.2 %	71.8 %
Explosion	79.4 %	50.4 %	61.7 %
Automobile	92.2 %	93.5 %	92.8 %
Motorcycle	87.1 %	99.1 %	92.7 %
Helicopter	95.8 %	89.8 %	92.7 %
Wind	75.0 %	87.0 %	80.6 %
Water	81.4 %	55.9 %	66.3 %
Rain	53.9 %	82.7 %	65.2 %
Applause	75.0 %	93.4 %	83.3 %
Crowd	96.8 %	89.7 %	93.1 %
Laughter	91.9 %	73.9 %	81.9 %
Average	80.6 %	80.6 %	80.6 %

Table A.22: Recall, precision and f-measure values for ASF, ASC, ASS and ZCR feature combination with SVM classification.

ASF+ASC+ASS+ZCR feature with SVM			
	Recall	Precision	F-measure
Emergency Alarm	80.0 %	95.5 %	87.1 %
Car Horn	16.6 %	83.3 %	27.7 %
Gun-shot	42.8 %	83.2 %	56.5 %
Explosion	76.5 %	34.0 %	47.1 %
Automobile	41.5 %	47.0 %	44.1 %
Motorcycle	50.8 %	75.6 %	60.7 %
Helicopter	77.1 %	91.0 %	83.4 %
Wind	60.3 %	74.4 %	66.6 %
Water	67.0 %	60.8 %	63.7 %
Rain	59.8 %	68.8 %	64.0 %
Applause	28.3 %	48.5 %	35.7 %
Crowd	83.8 %	52.5 %	64.5 %
Laughter	12.1 %	34.6 %	18.0 %
Average	53.6 %	65.3 %	58.9 %

A.2 Confusion Matrices

Confusion matrix is a specific table visualizing performance of typically supervised learning algorithms. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. All correct guesses are located in the diagonal of the table which are surrounded by boxes. Following 22 confusion matrices are results of the acoustical classification experiments.

Table A.23: Confusion matrix for ASP feature with HMM classification.

Confusion Matrix for ASP feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[71]	13	2	9	12	4	14	5	16	0	16	23	5
Car Horn	3	[15]	0	1	2	0	1	3	1	0	2	0	2
Gun-shot	14	9	[83]	35	9	33	6	4	9	3	15	1	9
Explosion	1	2	3	[25]	7	12	4	2	1	1	6	0	0
Automobile	1	3	1	3	[7]	16	15	16	7	6	1	1	0
Motorcycle	0	6	7	7	20	[38]	15	5	3	16	5	0	0
Helicopter	0	0	0	0	3	2	[82]	8	1	13	6	3	0
Wind	0	0	0	1	4	2	52	[29]	0	19	9	0	0
Water	10	2	0	0	22	23	38	19	[26]	27	7	19	4
Rain	0	1	0	2	10	4	27	3	0	[57]	3	0	0
Applause	7	1	6	5	2	1	0	1	12	0	[19]	6	0
Crowd	1	1	0	0	1	0	4	0	8	1	1	[45]	0
Laughter	39	2	0	0	0	0	0	2	2	0	9	6	[14]

Table A.24: Confusion matrix for ASF feature with HMM classification.

Confusion Matrix for ASF feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[154]	4	0	0	1	0	1	1	1	0	0	13	15
Car Horn	0	[9]	4	3	0	3	0	0	0	4	0	0	7
Gun-shot	0	0	[70]	58	11	19	6	4	43	33	4	8	10
Explosion	1	0	10	[18]	2	5	6	0	10	10	1	1	0
Automobile	0	0	7	0	[25]	12	1	5	20	0	0	5	2
Motorcycle	0	0	12	5	10	[55]	3	1	30	0	2	3	1
Helicopter	2	0	9	20	1	2	[76]	5	3	0	0	0	0
Wind	2	0	4	1	0	0	0	[68]	10	25	6	0	0
Water	0	0	3	8	3	4	0	2	[118]	50	3	2	4
Rain	0	0	0	0	0	0	0	0	17	[88]	0	0	2
Applause	3	1	3	1	2	0	0	0	12	1	[24]	10	3
Crowd	1	0	1	0	0	1	0	0	2	0	0	[56]	1
Laughter	1	1	0	0	2	0	0	0	1	0	0	14	[55]

Table A.25: Confusion matrix for ASC feature with HMM classification.

Confusion Matrix for ASC feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[82]	16	1	12	0	4	0	8	10	18	26	10	1
Car Horn	0	[5]	11	2	2	2	1	0	2	1	3	1	0
Gun-shot	10	10	[67]	45	11	18	13	19	15	7	17	25	8
Explosion	6	0	4	[14]	10	3	7	8	3	1	3	5	0
Automobile	3	2	0	1	[19]	14	13	3	1	0	7	14	0
Motorcycle	8	0	0	1	41	[37]	8	5	0	0	0	22	0
Helicopter	1	0	0	14	4	7	[49]	18	4	0	1	19	1
Wind	6	2	0	17	0	2	15	[35]	0	11	3	22	2
Water	53	13	3	11	3	1	22	7	[8]	8	30	34	4
Rain	11	8	5	6	2	3	16	1	1	[36]	5	10	3
Applause	12	1	5	4	2	0	4	1	5	0	[14]	10	1
Crowd	1	0	0	0	2	3	2	2	0	0	0	[52]	0
Laughter	23	1	1	2	0	1	0	2	9	7	17	11	[0]

Table A.26: Confusion matrix for ASS feature with HMM classification.

Confusion Matrix for ASS feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[58]	3	15	15	39	4	7	14	3	3	16	1	10
Car Horn	2	[1]	9	1	0	4	0	2	0	0	2	5	4
Gun-shot	31	8	[37]	34	16	11	14	7	14	14	19	22	38
Explosion	4	0	5	[6]	4	5	9	8	4	10	3	2	4
Automobile	3	0	1	1	[17]	17	5	4	2	5	19	3	0
Motorcycle	5	1	1	1	13	[47]	9	9	3	1	12	19	1
Helicopter	1	0	6	1	5	29	[21]	23	3	10	2	17	0
Wind	17	1	9	7	5	9	8	[34]	3	11	8	2	1
Water	4	0	7	15	3	38	34	26	[10]	24	12	15	9
Rain	0	1	2	7	3	2	16	6	7	[50]	1	4	8
Applause	6	0	2	5	5	5	13	3	1	0	[17]	2	0
Crowd	0	0	1	6	0	20	12	2	6	4	3	[7]	1
Laughter	10	1	2	1	15	7	9	1	3	1	13	0	[11]

Table A.27: Confusion matrix for MFCC feature with HMM classification.

Confusion Matrix for MFCC feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	146	7	6	0	0	2	2	3	0	0	7	11	5
Car Horn	0	10	3	0	1	3	1	0	1	0	4	0	7
Gun-shot	1	7	136	60	0	2	9	2	26	8	6	2	7
Explosion	1	0	8	44	1	0	5	2	1	1	0	1	0
Automobile	0	0	1	5	41	6	6	2	13	0	0	3	0
Motorcycle	0	1	0	19	9	77	3	2	9	1	0	1	0
Helicopter	0	0	5	9	0	0	100	1	0	0	0	3	0
Wind	0	0	0	0	12	2	1	70	14	12	1	3	0
Water	0	0	0	4	7	1	0	0	152	20	0	11	2
Rain	0	0	0	0	0	1	0	1	26	69	0	10	0
Applause	0	4	1	1	4	2	0	2	3	1	39	3	0
Crowd	0	1	0	0	1	0	1	0	1	0	0	58	0
Laughter	3	2	2	0	0	0	0	0	7	0	6	6	48

Table A.28: Confusion matrix for AH feature with HMM classification.

Confusion Matrix for AH feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	102	32	4	0	12	0	2	1	1	0	0	33	0
Car Horn	1	12	6	2	4	1	0	0	0	0	1	2	1
Gun-shot	14	10	34	26	59	9	31	8	16	11	20	15	12
Explosion	5	0	3	11	17	6	11	0	1	0	3	7	0
Automobile	13	0	1	2	9	7	18	6	0	0	12	7	2
Motorcycle	10	11	7	10	13	16	12	3	1	0	16	16	7
Helicopter	6	0	5	12	0	6	43	1	3	0	2	38	1
Wind	11	6	0	8	23	6	9	14	0	15	0	21	2
Water	2	0	4	8	7	11	40	8	37	25	50	2	2
Rain	0	0	10	6	2	1	8	4	14	46	5	3	8
Applause	2	1	2	0	1	3	7	10	12	1	12	8	0
Crowd	3	0	0	0	0	2	25	0	0	0	0	32	0
Laughter	12	2	0	2	5	5	6	13	3	1	3	22	0

Table A.29: Confusion matrix for ZCR feature with HMM classification.

Confusion Matrix for ZCR feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	90	0	0	7	4	18	30	0	15	6	2	16	0
Car Horn	1	4	13	1	0	0	1	3	4	2	0	0	1
Gun-shot	12	6	48	13	17	21	23	13	47	26	11	23	5
Explosion	8	0	4	3	5	9	19	4	2	0	1	8	1
Automobile	12	0	0	0	27	3	18	2	2	0	2	11	0
Motorcycle	8	0	0	1	19	20	39	23	5	0	0	7	0
Helicopter	4	0	0	0	0	9	101	2	0	0	0	2	0
Wind	1	0	3	3	7	20	39	19	15	5	0	2	1
Water	41	0	1	5	1	14	5	2	84	18	5	20	1
Rain	4	8	2	1	2	5	0	6	29	37	0	8	5
Applause	14	1	0	1	2	5	6	0	8	0	8	14	0
Crowd	10	0	0	0	1	7	19	0	4	0	0	21	0
Laughter	18	0	2	1	1	6	1	1	26	1	2	14	1

Table A.30: Confusion matrix for ASF, ASC, ASS and ZCR feature combination with HMM classification.

Confusion Matrix for ASF+ASC+ASS+ZCR feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[165]	6	2	1	2	0	2	0	0	0	0	9	3
Car Horn	1	[9]	9	5	0	0	0	1	0	1	0	0	4
Gun-shot	1	0	[150]	48	2	1	11	2	22	8	2	8	11
Explosion	0	0	9	[34]	4	1	2	0	12	1	0	0	1
Automobile	0	0	2	0	[42]	11	6	1	0	0	1	14	0
Motorcycle	0	0	4	4	19	[76]	5	5	5	0	0	4	0
Helicopter	0	0	0	0	1	3	[101]	5	0	0	0	8	0
Wind	0	0	0	4	0	0	1	[65]	26	10	2	7	1
Water	4	0	6	6	0	0	0	2	[139]	28	10	1	1
Rain	0	0	7	1	0	0	0	5	24	[67]	3	0	0
Applause	3	0	7	0	3	0	3	2	10	0	[21]	7	4
Crowd	0	0	0	0	1	2	1	0	0	0	0	[58]	0
Laughter	18	0	0	0	0	0	0	0	4	1	2	10	[39]

Table A.31: Confusion matrix for ASF, ASC, ASS and AH feature combination with HMM classification.

Confusion Matrix for ASF+ASC+ASS+AH feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[178]	0	0	1	0	0	0	0	0	0	2	1	6
Car Horn	3	[5]	3	3	0	0	0	1	8	0	3	0	4
Gun-shot	0	0	[88]	54	3	7	1	17	41	32	6	7	10
Explosion	0	0	12	[31]	0	5	1	4	9	1	0	1	0
Automobile	0	0	1	0	[61]	3	0	1	2	0	2	7	0
Motorcycle	0	0	3	2	10	[92]	2	0	3	0	2	8	0
Helicopter	0	0	1	4	4	1	[101]	0	0	0	0	7	0
Wind	0	0	2	0	0	0	1	[76]	14	14	2	6	0
Water	0	0	7	3	0	0	1	1	[139]	30	10	1	5
Rain	0	0	0	4	0	0	0	0	20	[83]	0	0	0
Applause	0	0	1	1	2	0	2	0	5	1	[40]	5	2
Crowd	0	0	0	0	0	1	0	0	0	0	0	[61]	0
Laughter	8	0	0	0	0	0	0	0	3	1	2	5	[55]

Table A.32: Confusion matrix for MFCC, ASC, ASS and ZCR feature combination with HMM classification.

Confusion Matrix for MFCC+ASC+ASS+ZCR feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[169]	1	4	5	0	0	0	0	1	0	6	4	0
Car Horn	1	[8]	9	0	2	0	3	0	1	0	2	0	4
Gun-shot	2	10	[125]	55	10	14	9	0	16	3	5	2	15
Explosion	1	0	7	[45]	1	0	4	2	2	1	0	0	1
Automobile	0	0	0	1	[32]	16	13	1	12	0	0	1	1
Motorcycle	0	0	2	2	11	[93]	3	0	2	1	1	7	0
Helicopter	0	0	7	17	0	0	[91]	1	0	0	0	2	0
Wind	1	0	0	7	1	0	0	[80]	21	1	0	5	0
Water	0	2	2	7	1	3	2	1	[148]	23	4	3	1
Rain	0	0	7	0	0	1	0	1	24	[74]	0	0	0
Applause	2	1	3	1	5	2	0	1	10	0	[32]	2	1
Crowd	1	0	0	0	3	0	5	0	0	0	0	[53]	0
Laughter	9	3	1	0	1	0	0	0	5	1	7	6	[41]

Table A.33: Confusion matrix for MFCC, ASC, ASS and AH feature with HMM classification.

Confusion Matrix for MFCC+ASC+ASS+AH feature with HMM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[170]	3	3	5	0	0	0	2	1	0	2	2	0
Car Horn	0	[18]	6	0	0	1	1	0	0	0	1	0	3
Gun-shot	2	2	[157]	46	3	4	7	1	18	6	10	2	7
Explosion	1	0	11	[43]	0	0	4	2	0	1	0	0	2
Automobile	0	0	0	1	[46]	10	8	0	10	0	0	2	0
Motorcycle	0	0	1	5	6	[90]	5	0	5	1	3	6	0
Helicopter	1	0	8	16	0	0	[83]	3	2	0	0	5	0
Wind	2	0	1	4	1	0	0	[81]	9	9	1	3	4
Water	0	0	1	11	0	2	2	10	[139]	24	1	2	5
Rain	0	0	6	0	0	1	1	0	9	[83]	0	7	0
Applause	1	2	9	0	3	1	2	0	2	1	[37]	1	0
Crowd	1	1	0	0	2	1	1	0	0	0	0	[56]	0
Laughter	3	1	0	0	2	0	0	0	4	0	7	3	[54]

Table A.34: Confusion matrix for ASP feature with SVM classification.

Confusion Matrix for ASP feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	69	0	3	27	0	4	6	0	62	0	11	4	4
Car Horn	0	7	10	1	2	2	2	0	3	0	3	0	0
Gun-shot	11	26	64	42	12	27	7	2	25	0	1	4	9
Explosion	4	0	5	25	4	8	8	0	4	0	4	0	2
Automobile	2	1	1	3	5	24	7	3	27	0	4	0	0
Motorcycle	0	0	3	7	17	56	6	0	23	0	10	0	0
Helicopter	0	0	0	0	1	4	44	1	64	0	4	0	0
Wind	0	0	1	1	8	11	36	5	53	0	1	0	0
Water	1	0	1	1	3	48	17	2	114	0	8	1	1
Rain	0	4	0	0	4	30	20	3	42	1	3	0	0
Applause	14	1	5	3	0	1	0	0	27	0	7	0	2
Crowd	3	0	0	0	0	1	4	0	47	0	2	5	0
Laughter	42	2	1	0	0	0	0	0	7	0	3	2	17

Table A.35: Confusion matrix for ASF feature with SVM classification.

Confusion Matrix for ASF feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[152]	0	0	0	0	0	2	1	0	0	0	8	5
Car Horn	0	[12]	0	0	0	4	0	0	4	5	0	0	0
Gun-shot	0	1	[67]	36	1	9	9	3	67	34	1	3	3
Explosion	0	0	2	[25]	0	4	6	0	18	4	1	1	0
Automobile	0	0	0	0	[22]	12	1	5	18	8	0	4	0
Motorcycle	0	0	6	1	13	[36]	4	0	37	0	2	0	0
Helicopter	0	0	6	1	0	0	[83]	3	0	0	0	0	0
Wind	0	0	6	2	1	0	0	[63]	19	15	1	2	0
Water	0	0	1	3	0	0	0	0	[151]	24	1	0	0
Rain	0	0	1	0	0	0	0	0	52	[54]	0	0	0
Applause	0	1	0	0	0	0	0	0	26	1	[5]	9	0
Crowd	0	0	0	0	0	1	0	0	2	0	0	[48]	0
Laughter	2	0	0	0	0	1	0	0	3	2	0	10	[12]

Table A.36: Confusion matrix for ASC feature with SVM classification.

Confusion Matrix for ASC feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[91]	17	0	7	1	0	4	26	24	0	0	18	0
Car Horn	2	[10]	0	4	0	0	3	11	0	0	0	0	0
Gun-shot	7	69	[0]	35	8	1	27	55	25	0	0	38	0
Explosion	4	3	0	[21]	7	1	4	16	0	0	0	8	0
Automobile	11	0	0	39	[1]	0	16	1	2	0	0	7	0
Motorcycle	8	0	0	77	0	[0]	7	1	10	0	0	19	0
Helicopter	2	0	0	22	9	3	[39]	9	3	0	0	31	0
Wind	15	10	0	46	6	1	19	[6]	2	0	0	10	0
Water	71	25	0	11	2	0	10	22	[19]	0	0	37	0
Rain	14	43	0	8	1	2	5	6	8	[0]	0	20	0
Applause	21	1	0	1	0	0	4	4	13	0	[0]	15	0
Crowd	1	0	0	3	0	0	11	0	7	0	0	[40]	0
Laughter	30	7	0	1	2	0	0	2	23	0	0	9	[0]

Table A.37: Confusion matrix for ASS feature with SVM classification.

Confusion Matrix for ASS feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	49	0	1	13	45	10	13	0	32	11	0	14	0
Car Horn	8	0	0	0	5	4	4	0	5	4	0	0	0
Gun-shot	18	1	19	28	10	12	35	3	69	49	0	21	0
Explosion	0	0	2	13	3	4	10	1	22	3	0	6	0
Automobile	1	0	0	5	34	14	10	0	11	1	0	1	0
Motorcycle	2	0	0	3	36	43	28	0	5	2	0	3	0
Helicopter	0	0	0	8	11	32	37	2	18	0	0	10	0
Wind	17	0	0	12	14	4	27	0	16	1	0	24	0
Water	0	0	2	40	31	20	35	1	39	21	0	8	0
Rain	0	4	15	38	0	0	7	3	24	5	0	11	0
Applause	0	0	0	1	17	4	12	0	15	2	0	8	0
Crowd	0	0	0	2	6	11	24	0	11	1	0	7	0
Laughter	4	0	0	0	26	5	22	0	10	7	0	0	0

Table A.38: Confusion matrix for MFCC feature with SVM classification.

Confusion Matrix for MFCC feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[155]	0	8	0	0	0	3	1	0	0	15	6	1
Car Horn	1	[16]	3	0	3	0	1	0	0	0	1	0	5
Gun-shot	4	26	[109]	34	7	3	13	1	41	4	1	8	13
Explosion	2	0	18	[32]	0	1	7	0	2	1	0	1	0
Automobile	0	0	0	8	[33]	11	4	3	6	0	0	12	0
Motorcycle	0	0	4	24	23	[58]	4	0	5	1	0	3	0
Helicopter	0	0	3	8	0	0	[87]	10	3	0	0	7	0
Wind	0	0	0	0	5	0	12	[61]	32	3	1	1	0
Water	0	0	7	6	0	0	0	2	[165]	0	0	11	6
Rain	0	4	1	0	0	2	0	5	62	[31]	0	0	2
Applause	5	6	0	3	3	2	0	1	14	0	[12]	13	1
Crowd	0	0	0	0	1	0	2	0	1	0	0	[58]	0
Laughter	3	3	1	0	0	0	0	0	7	0	11	11	[38]

Table A.39: Confusion matrix for AH feature with SVM classification.

Confusion Matrix for AH feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[117]	2	1	29	15	13	0	10	0	0	0	0	0
Car Horn	12	[1]	7	8	0	0	0	1	0	0	0	0	1
Gun-shot	16	2	[46]	192	5	1	0	1	0	0	0	0	2
Explosion	6	0	0	[56]	1	0	0	1	0	0	0	0	0
Automobile	8	0	0	55	[7]	5	0	2	0	0	0	0	0
Motorcycle	9	0	1	102	3	[2]	0	5	0	0	0	0	0
Helicopter	0	0	0	117	0	0	[0]	0	0	0	0	0	0
Wind	6	0	15	73	8	3	0	[10]	0	0	0	0	0
Water	3	0	25	165	2	1	0	0	[0]	0	0	0	0
Rain	0	0	49	58	0	0	0	0	0	[0]	0	0	0
Applause	1	0	6	49	0	1	0	2	0	0	[0]	0	0
Crowd	0	0	1	58	2	0	0	1	0	0	0	[0]	0
Laughter	4	0	0	54	1	12	0	3	0	0	0	0	[0]

Table A.40: Confusion matrix for ZCR feature with SVM classification.

Confusion Matrix for ZCR feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	127	1	1	15	0	22	0	0	0	0	0	22	0
Car Horn	3	11	4	3	0	7	0	0	0	0	0	2	0
Gun-shot	60	66	37	21	0	19	0	0	0	0	0	62	0
Explosion	11	1	11	8	0	9	0	0	0	0	0	24	0
Automobile	30	0	34	12	0	0	0	0	0	0	0	1	0
Motorcycle	24	1	42	31	0	11	0	0	0	0	0	13	0
Helicopter	15	0	31	65	0	5	0	0	0	0	0	2	0
Wind	12	20	38	26	0	10	0	0	0	0	0	9	0
Water	90	78	0	2	0	8	0	0	0	0	0	19	0
Rain	19	72	0	2	0	4	0	0	0	0	0	10	0
Applause	31	5	0	1	0	3	0	0	0	0	0	19	0
Crowd	31	1	0	3	0	6	0	0	0	0	0	21	0
Laughter	55	13	0	0	0	1	0	0	0	0	0	5	0

Table A.41: Confusion matrix for ASF, ASC, ASS and ZCR feature with SVM classification.

Confusion Matrix for ASF+ASC+ASS+ZCR feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[152]	1	0	1	1	0	0	0	1	0	1	1	3
Car Horn	0	[5]	7	0	0	4	0	0	1	0	0	0	5
Gun-shot	0	0	[114]	53	3	1	5	4	20	13	1	5	6
Explosion	0	0	8	[49]	0	0	0	3	0	0	0	0	0
Automobile	0	0	0	0	[32]	13	0	1	1	0	0	16	0
Motorcycle	0	0	0	2	28	[62]	3	0	9	0	0	6	0
Helicopter	0	0	0	0	0	1	[91]	7	0	0	0	4	0
Wind	0	0	0	12	1	0	0	[70]	10	9	2	6	1
Water	0	0	8	16	0	0	0	7	[132]	7	13	1	1
Rain	0	0	0	10	1	0	0	1	31	[64]	0	0	0
Applause	0	0	0	1	1	0	0	1	10	0	[17]	5	1
Crowd	0	0	0	0	1	1	1	0	0	0	0	[52]	0
Laughter	7	0	0	0	0	0	0	0	2	0	1	3	[9]

Table A.42: Confusion matrix for ASF, ASC, ASS and AH feature with SVM classification.

Confusion Matrix for ASF+ASC+ASS+AH feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	[164]	22	0	0	0	0	0	0	2	0	0	0	5
Car Horn	0	[23]	2	0	0	0	0	0	0	0	0	0	2
Gun-shot	0	5	[170]	32	2	1	9	4	34	2	0	1	11
Explosion	0	0	8	[54]	0	0	1	1	3	0	0	0	0
Automobile	0	0	0	0	[71]	0	0	0	5	0	0	1	0
Motorcycle	0	0	0	1	2	[108]	1	0	10	0	0	2	0
Helicopter	0	0	2	0	0	0	[115]	1	2	0	0	0	0
Wind	0	0	0	2	0	0	0	[87]	18	8	1	0	0
Water	1	0	10	9	1	0	1	5	[166]	3	1	1	6
Rain	0	0	3	8	0	0	0	2	39	[61]	0	0	0
Applause	0	0	0	1	0	0	0	0	14	0	[45]	0	0
Crowd	0	0	0	0	0	0	1	0	1	0	0	[61]	0
Laughter	0	0	0	0	0	0	0	0	3	0	1	2	[68]

Table A.43: Confusion matrix for MFCC, ASC, ASS and ZCR feature with SVM classification.

Confusion Matrix for MFCC+ASC+ASS+ZCR feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	167	1	5	5	0	0	0	1	1	0	7	3	0
Car Horn	0	16	3	2	1	0	1	0	0	0	3	0	4
Gun-shot	5	6	144	42	6	3	12	0	20	1	0	5	16
Explosion	2	0	9	43	0	0	5	3	0	0	0	0	2
Automobile	0	0	1	0	41	15	3	0	5	0	0	12	0
Motorcycle	0	0	8	2	25	75	1	0	2	0	0	8	0
Helicopter	0	0	7	16	0	0	88	0	0	0	0	7	0
Wind	0	0	0	6	4	0	2	76	12	10	3	3	0
Water	0	0	15	13	3	1	0	0	146	4	1	9	5
Rain	0	0	0	0	0	6	0	0	31	69	0	1	0
Applause	5	4	1	0	3	1	1	1	15	0	16	11	2
Crowd	0	0	0	0	3	0	1	0	0	0	0	58	0
Laughter	2	3	0	0	0	0	0	0	7	0	9	3	50

Table A.44: Confusion matrix for MFCC, ASC, ASS and AH feature with SVM classification.

Confusion Matrix for MFCC+ASC+ASS+AH feature with SVM													
	Emergency Alarm	Car Horn	Gun-shot	Explosion	Automobile	Motorcycle	Helicopter	Wind	Water	Rain	Applause	Crowd	Laughter
Emergency Alarm	175	0	10	0	2	0	0	1	0	0	0	0	0
Car Horn	2	16	2	2	0	0	1	2	0	0	1	0	4
Gun-shot	1	4	147	35	3	4	13	0	27	2	0	4	18
Explosion	1	0	11	41	0	0	7	3	0	0	0	0	1
Automobile	0	0	0	0	38	18	1	0	6	0	0	14	0
Motorcycle	1	0	4	3	6	94	2	0	3	1	0	8	0
Helicopter	0	0	4	13	0	0	96	0	3	0	0	2	0
Wind	0	0	1	2	4	0	2	78	10	7	5	6	0
Water	0	0	13	11	2	1	0	1	154	4	0	6	5
Rain	0	0	0	0	0	4	0	0	31	71	0	1	0
Applause	2	1	1	0	3	1	1	1	13	0	22	13	1
Crowd	1	0	0	0	3	0	2	0	0	0	0	56	0
Laughter	6	0	0	0	0	0	0	0	7	0	3	3	54

APPENDIX B

CONFUSION MATRICES OF SEMANTIC EXPERIMENTS

Table B.1: Confusion matrix for semantic classification experiment with GA.

Experiment with GA				
	Outdoor	Nature	Meeting	Violence
Outdoor	964	92	186	2
Nature	0	1332	0	19
Meeting	0	109	886	0
Violence	120	72	18	974

Table B.2: Confusion matrix for semantic classification experiment with SVM.

Experiment with SVM				
	Outdoor	Nature	Meeting	Violence
Outdoor	639	139	64	402
Nature	15	1276	3	57
Meeting	25	103	747	120
Violence	111	140	91	844

APPENDIX C

SMOOTHING EXAMPLES

Table C.1: An example segment sequence before smoothing.

An example segment sequence just after the acoustic classification			
Classified Acoustic Class	Classification Score	Second Possible Class	Second Possible Score
Water	0.71	Helicopter	0.55
Applause	0.82	Water	0.68
Rain	0.69	Water	0.52
Water	0.77	Automobile	0.53
Water	0.69	Helicopter	0.64

Table C.2: An example segment sequence after smoothing.

Segment sequence in Table C.1 after smoothing process			
Classified Acoustic Class	Classification Score	Second Possible Class	Second Possible Score
Water	0.71	Helicopter	0.55
Water	0.74	Water	0.68
Water	0.74	Water	0.52
Water	0.77	Automobile	0.53
Water	0.69	Helicopter	0.64

APPENDIX D

HMM STATE COUNT OPTIMIZATION EXPERIMENT

In order to explore the best state count for HMM classification, a small group of experiment is conducted with ASF+ASC+ASS+AH feature. The dataset for acoustic classes is used for this experiment. 5-state HMM provides better F-measure compared to 4-state, 6-state and 7-state HMMs.

Table D.1: Classification performance of 4-state HMM.

	Recall	Precision	FValue
Emergency Alarm	90.9 %	90.0 %	90.4 %
Car Horn	20.0 %	85.7 %	32.4 %
Gun	27.5 %	68.8 %	39.3 %
Bomb	43.7 %	24.3 %	31.2 %
Automobile	50.6 %	44.8 %	47.5 %
Motorcycle	57.3 %	67.9 %	62.2 %
Helicopter	62.7 %	87.0 %	72.9 %
Wind	66.0 %	71.6 %	68.7 %
Water	66.4 %	55.9 %	60.7 %
Rain	71.9 %	45.2 %	55.5 %
Applause	45.7 %	50.0 %	47.7 %
Crowd	90.3 %	44.0 %	59.2 %
Laughter	62.1 %	48.9 %	54.7 %
Average	58.1 %	60.3 %	59.2 %

Table D.2: Classification performance of 5-state HMM.

ASF+ASC+ASS+AH feature with HMM			
	Recall	Precision	FValue
Emergency Alarm	94.7 %	94.2 %	94.4 %
Car Horn	16.7 %	100.0 %	28.6 %
Gun-shot	33.1 %	74.6 %	45.8 %
Explosion	48.4 %	30.1 %	37.1 %
Automobile	79.2 %	76.3 %	77.7 %
Motorcycle	75.4 %	84.4 %	79.6 %
Helicopter	85.6 %	92.7 %	88.9 %
Wind	66.1 %	76.0 %	70.7 %
Water	70.1 %	56.9 %	63.0 %
Rain	77.6 %	51.2 %	61.7 %
Applause	67.8 %	57.9 %	62.5 %
Crowd	98.4 %	55.9 %	71.3 %
Laughter	74.3 %	67.1 %	70.5 %
Average	68.2 %	70.5 %	69.4 %

Table D.3: Classification performance of 6-state HMM.

	Recall	Precision	FValue
Emergency Alarm	93.0 %	88.8 %	90.9 %
Car Horn	20.0 %	75.0 %	31.5 %
Gun	40.0 %	67.5 %	50.2 %
Bomb	43.7 %	27.1 %	33.5 %
Automobile	57.1 %	54.3 %	55.6 %
Motorcycle	57.3 %	77.7 %	66.0 %
Helicopter	67.7 %	86.0 %	75.8 %
Wind	65.2 %	73.5 %	69.1 %
Water	73.0 %	57.3 %	64.2 %
Rain	73.8 %	53.3 %	61.9 %
Applause	49.1 %	50.8 %	50.0 %
Crowd	91.9 %	49.1 %	64.0 %
Laughter	62.1 %	61.3 %	61.7 %
Average	61.1 %	63.2 %	62.1 %

Table D.4: Classification performance of 7-state HMM.

	Recall	Precision	FValue
Emergency Alarm	89.8 %	89.8 %	89.8 %
Car Horn	23.3 %	38.8 %	29.1 %
Gun	35.4 %	69.6 %	47.0 %
Bomb	53.1 %	26.5 %	35.4 %
Automobile	45.4 %	66.0 %	53.8 %
Motorcycle	72.1 %	75.2 %	73.6 %
Helicopter	80.5 %	89.6 %	84.8 %
Wind	61.7 %	84.5 %	71.3 %
Water	70.5 %	56.9 %	63.0 %
Rain	77.5 %	51.8 %	62.1 %
Applause	38.9 %	54.7 %	45.5 %
Crowd	93.5 %	49.5 %	64.8 %
Laughter	64.8 %	55.8 %	60.0 %
Average	62.0 %	62.2 %	62.1 %

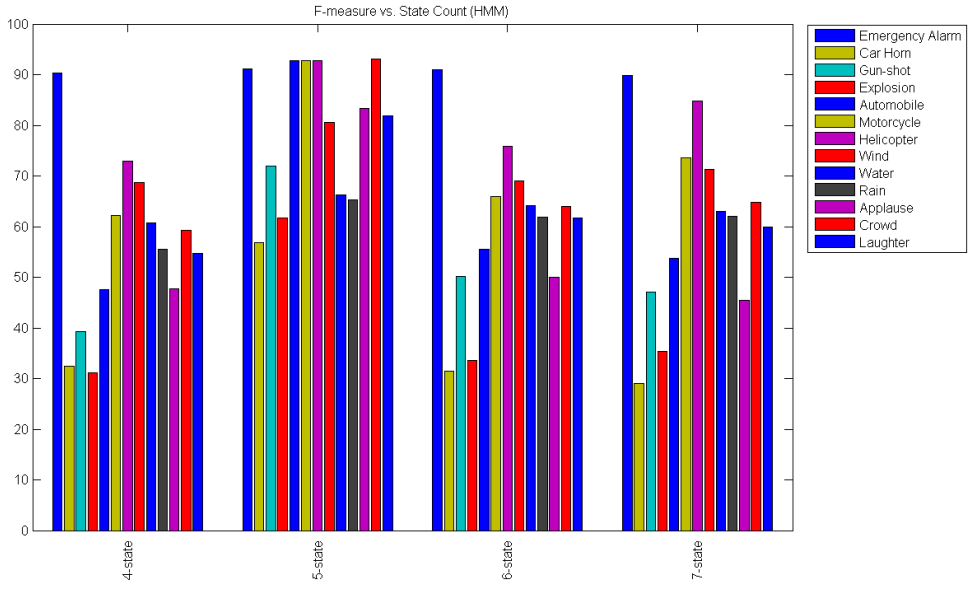

Figure D.1: Results of state count optimization experiment.