Chapter 1

Environmental Sound Source Separation, Classification and Indexing

1.1 Introduction

In everyday life, people are often in complex acoustic environments, where we are surrounded by acoustic mixtures consisting of various sound types. Sounds can be roughly grouped into three clusters, namely human voice, artificial sound, and non-artificial/natural sound. Human voice refers to sounds created by people physically such as speech, cough, and singing. Artificial sounds refer to sounds created by human activities such as traffic, aircraft, and music. Non-artificial sounds include sounds created by nature such as wind, rain, land animal, insects and marine life.

Digitally stored multimedia material [1] is currently a rapidly growing resource due to the ongoing technological advancement in data storage, communications and computing. In only recent years, it has become widely possible to transfer long audio and video files via the internet with an acceptable delay. The storage capacities of portable multimedia devices and personal computers have been rapidly increasing. The amount of available digital multimedia information is beginning to overwhelm the capacity of humans to manage and organize it. Thus, computerized solutions for automatic organization of the multimedia material are an attractive approach to access the content efficiently. Because of this, information retrieval is an important field of application for automatic audio recognition. Besides audio material, the indexing and
retrieval of (audiovisual) video material also benefits from audio recognition.

Besides retrieval applications, another important type of recognition application for audio is automatic transcription. It refers to using computer-aided methods for transcription i.e., generating a sequence of class labels in order to save human labor; examples include speech dictation, music transcription and environment recognition applications. While information retrieval uses automatic recognition results as content descriptors to accomplish the search, transcription can directly use the recognition results as output. Automatic audio recognition is also central in some user interfaces and surveillance applications [1]. It is also frequently useful in application areas such as coding, enhancement and synthesis.

Sounds which cannot be categorized as either music or speech have been widely termed environmental sounds [2]. These sounds are instrumental in generating an awareness of the surrounding environment as the various active and passive sources that constitute the acoustic environment generate, reflect, and modify the sound in ways which leaves their imprints within the sound signal. This information becomes parts of our understanding in subtle ways and we are consciously unaware of this process unless we acoustically isolate ourselves from the environment, e.g., in anechoic chambers where people may feel disoriented because the information derived from room reverberation cues is missing or when we try to drive a car or cross a busy road while listening to music with our headphones.

It is at these times that we realize that these sounds generate a subconscious awareness of the surroundings, e.g., indicates the presence of invisible objects, provides location, focuses our attention if it is important, orients us with respect to the surroundings in a timely manner. The awareness of the surroundings generated through this first-order environmental sound analysis is an indispensable tool in the
repertoire of an intelligent self-aware being. If one wants to augment this awareness for disabled humans under sensory deprived conditions or replicate it in autonomous systems, such as robots, an appreciation of the role played by environmental sounds is necessary.

This insight is motivated by the human auditory system. Beginning at the ear and extending up to and beyond the auditory cortex, this auditory system extracts, maintains, and utilizes multiple representations for sounds and as a result, it has been argued, exhibits capabilities in perceiving and extracting relevant information from the environmental sounds [3]. The exact nature of the particular representations and the mechanisms employed by the biological auditory system are perhaps not as important as the principle of maintaining multiple representations relevant to the sound at hand.

1.2 Sound Classification

Sound may be defined as any pressure variation that the human ear can detect. Sound classification makes a distinction between three domains. Namely Speech sounds, Musical sounds and Environmental sounds [4]. This division only arises as a consequence of the division of scientific domains:

- Is studied by phonologists and phoneticians
- By musicologists and musicians
- By psychoacousticians

The block diagram of Hierarchical sound classification system is shown in Fig. 1.1.

The description of speech sounds is often performed for different scientific motivations than the description of musical and environmental sounds. This explains some of the heterogeneity of the approach to sound classification. For example, formant
notion, characteristic of the acoustic signal, is relevant in phonetics to describe vowels but exists as impedance in musical sounds; it is a physical characteristic of the musical instrument, not of the sounds. In other words, in music, the format is relevant to describe the mechanics of the instrument.

Formerly and acoustically the same concept is employed to describe two acoustic “realities” relevant to the science impulsing the classification. The majority of studies on sound classification (and categorization) focus not on identification and semantic categories, but on the first step of perception. Exceptions are to be found in musical acoustics and within the domain of ecological psychology for environmental sounds [5], [6].

Fig. 1.1: Block Diagram of Hierarchical Sound Classification System.
1.2.1 Speech Sounds

Speech can be characterized as sound made by the human vocal tract proposed for correspondence. Recorded speech and system produced sound that approximates speech are additionally considered speech. There are numerous approaches to part up the discourse area into sub-classifications. A conspicuous path is to classify speech by language. One can likewise arrange discourse by who or what the speaker is, by the enthusiastic substance of the speaker and by the topic of the discourse. Further down the chain of importance, we can group discourse into specific words utilized, and afterward specific phonemes.

A meaningful unit of sound within a language, known as a phoneme is defined with reference to a set of articulatory and acoustic distinctive features, the value of which is determined by its contrastive relationship with other phonemes within a given language. This could mean that the quantity and kind of meaningful sound distinctions in any language is limited by the available distinctive features, though it could also mean that the matter is functional (based on what humans perceive optimally among articulable sounds), at least in part.

The relationship between human perception and auditory objects in speech sounds is apparent when one considers the definition of vowels solely from the acoustic characteristics of its formants. The variability of production prohibits fixing the measurement of each formant. Despite the differences in the observed formant structures among speakers of the same language and even within different realizations of a given vowel by the same speaker. The concept of prototype or percept magnet for example [7] can provide insight. It is insufficient to define a vowel (or any segment) as a signal with a set of necessary and sufficient acoustic characteristics, disregarding the linguistic system in which the segment contrasts. This system, which is in the perceiver’s mind
(and not in the signal), drives a perceiver’s attention to be sensitive to relevant cues in the signal for a particular goal.

1.2.2 Musical Sounds

In classical Western music, the acoustic characteristics of a note are defined by their relation within an octave. The musical notation extracts the pitch, the duration and, in some case, the loudness from the note. These dimensions are expressed in a representation that trained musicians can read. For example, an octave distance between two sounds is noted in music with an equal distance gap regardless of the octave. In frequency measurement, the distance is always different due to the necessity to double the frequency between each octave.

Musical notation, then, takes into account perception but not acoustics of the sound. As for speech, the perception of musical sounds is driven by the musical system: we do not try to hear third tones in Western music, where it is irrelevant, even if it is possible to measure some third tones in a performance. The goal of the listener (to hear music) depends of the musical system used and drives perception.

Timbre is another musical dimension, which provides an interesting vantage point from which to frame the link between acoustic parameters and human perception. Timbre is never used to scientifically classify instruments by itself, apart from the orchestration used by composers. In this case, the classification of timbre (in reference to an instrument) depends on several factors, be they linked to religious symbolism or the role of an instruments within an orchestra but not necessary to an acoustic definition. From an aesthetic perspective, timbre is the quality of something making sound. From a musical perspective, it is the type of musical instrument that one can hear as a sound family (like a piano) or as a sub-type of instrument (as a concert piano or electric piano). From a scientific perspective, a systematic classification taking into
account the acoustic mechanism of the instrument and, at times, the sound resulting from this mechanism, without taking into account the musician itself. But this acoustic classification (and more exactly, this mechanical classification) never used the timbre notion, because this concept, used by musicians, is relevant only in musical context.

1.2.3 Environmental Sounds

Unlike musical and speech sounds, environmental sounds do not belong to a system. Therefore, the categorical membership of environmental sounds is considerably less consensual than it is for speech (where sounds can be organized into phonetic inventories and described in terms of distinctive features) or music (where sounds, at least in the western tradition, can be organized into notes, octaves, etc.).

Although physical parameters of sound (fundamental frequency, duration, intensity, etc.) can yield some insight into particular characteristics, a descriptive overview is also required to more adequately characterize the perception of environmental sounds. Varied attempts have been made to provide different kinds of classification. The human in justifying his approach of everyday sounds perception, the taxonomy he proposes offers no human psychological aspect (like “familiarity”), apart from the concept of “event”.

Another example of descriptive overview of environmental sounds is the domain-based organization of sound effects on sound effect CDs. This is an apt illustration of how the meaning of environmental sounds is also contingent on so called “contextual” factors. In most cases, those factors relate to human activities. For example, the sound of a coffee-maker may be heard as a motor operating, an electrical appliance, a certain model of coffee-maker, an alarm indicating that it is time to awaken, or the recognition of one’s own coffee-maker. The perceiver often integrates congruent features from memory and/or other sense modalities into an (ordinary) multimodal
cognitive representation.

1.2.3.1 Noise

Noise is sound that is not wanted by the perceiver, because it is unpleasant, loud, or interferes with hearing. Noise is divided into two broad categories:

- **External noise:** External noise is noise introduced in the transmission channel. (ex. Industrial, Atmospheric)
- **Internal noise:** Internal is noise introduced inside the receiver itself. (ex. Semiconductor, Thermal)

1.2.3.2 Natural Sounds

Natural sounds are sounds produced by natural sources in their ordinary soundscape. The classification incorporates the hints of any living life form, from bug hatchlings to the biggest living warm blooded creature on the planet, whales, and those produced by characteristic, non-natural sources. In many regards, the characteristic natural surroundings from which these acoustic sources exude are characterized as not vigorously affected by human intervention.

1.2.3.3 Artificial Sounds

Artificial noise is a wave or vibration, audible, electromagnetic, or other signal, generated by a human source. It can be used to experiment on a subject by controlling the frequency or amplitude of the artificial noise to ascertain how the subject interacts with external stimulation. For example, to test the sensitivity of a microphone noise-reducing filter, the test administrator could generate artificial noise in a laboratory setting to determine whether the microphone suppresses the noise (i.e. filters it out), or interprets the noise as something that is not noise (i.e. passes it through).
With regards to urban dwellings or establishments, artificial noise might be called light pollution, or commuter traffic.

### 1.3 Generic Sound Event Characteristics

Sound is a vibration that propagates as a typically audible mechanical wave of pressure and displacement, through a medium such as air or water. The listening system in human beings senses sound of frequency that lies within 20 Hz to 20 kHz. The sound captured by a microphone gives digitalized signal by proper methods. The sound duration and any modulations in the amplitude denote temporal properties that an audio event might have. The spectral properties include the frequency components and their relative strengths.

It is a difficult task to summarise the characteristics of a generic sound event, due to the different ways in which sounds occur and, unlike speech which is confined to the sounds that are produced by the human vocal tract and tongue, a sound event may be produced by many different types of interactions. An example of an attempt in the literature [8] is shown in Table 1.1, which compares speech, music and environmental sounds. It can be seen that whereas speech and music have clear definitions according to these characteristics, environmental sounds are either undefined or can cover the full range of characteristics. A different approach is used in musical instrument classification [9], [10] where the following characteristics define the sound:

- **Timbre:** The quality of the musical note; distinguishes the type of sound production. Aspects of this include the spectral shape, the time envelope (e.g. onset and offset), and the noise-tonal characteristics.
- **Pitch:** The perceived fundamental frequency of the sound.
- **Loudness:** The perceived physical strength of the sound, which depends on the
Table 1.1: Characteristics of Speech, Music and Environmental Sounds.

<table>
<thead>
<tr>
<th>Acoustical Characteristics</th>
<th>Speech</th>
<th>Music</th>
<th>Environmental Sounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Classes</td>
<td>No. of Phonemes</td>
<td>No. of Tones</td>
<td>Undefined</td>
</tr>
<tr>
<td>Length of Window</td>
<td>Short (fixed)</td>
<td>Long (fixed)</td>
<td>Undefined</td>
</tr>
<tr>
<td>Length of Shift</td>
<td>Short (fixed)</td>
<td>Long (fixed)</td>
<td>Undefined</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Narrow</td>
<td>Relatively Narrow</td>
<td>Broad</td>
</tr>
<tr>
<td>Harmonics</td>
<td>Clear</td>
<td>Clear</td>
<td>Clear</td>
</tr>
<tr>
<td>Stationarity</td>
<td>Stationary (not percussion)</td>
<td>Stationary</td>
<td>Non-Stationary Stationary</td>
</tr>
<tr>
<td>Repetitive Structure</td>
<td>Weak</td>
<td>Weak</td>
<td>Strong</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weak</td>
</tr>
</tbody>
</table>

curve of equal loudness in human perception.

They are analysed using segmentation process in which short segments are analysed at intervals. Estimation results are better in representing those characters in the signal that correspond to the segmented signal. The way of estimation of the signal characteristics that varies over time is termed as short-time analysis. Source filter model evaluates the signal characteristics that sound possess.

1.4 Basic Principles of Sound

Sound is formed by vibrations. In order to create sound, an object must disturb the air and cause it to vibrate. These vibrations take the form of small fluctuations in the air pressure which are called sound waves [11].

10
Sound moves in the air as waves. When sound waves reach to the human ear, they cause the eardrum to vibrate. Then these vibrations are transmitted through the inner ear and then translated into electrical pulses that brain interprets as sound. The height of the wave which represents the amount of the vibration of the molecules in the air is called amplitude which corresponds to the volume of the sound, meaning higher the amplitude, louder the sound is. The time period between reoccurring of the waveforms is called frequency which is the length of the wave.

1.5 Audio Signal Representation

An audio signal represents the sound as an electrical voltage. Signal flow is nothing but a route taken by an audio signal for travelling towards the speaker from the source. Audio signal is characterized by bandwidth, power and voltage. Impedance of the signal path determines the relation between power and voltage [12]. Electrical signal is used by analog processors but digital signals are mathematically dealt by the digital processors.

In order to process a sound wave digitally, sound wave needs to be transferred in to the digital bits. Transferring these analog signals to digital bits is made by an analog-to-digital sound converter.

First task in digitization process is sampling. Sampling is done by taking the values of the waveform, per a constant period of time. The number of times per second the sample is taken called sampling rate, which is defined in terms of frequency. After each sample point is set, the amplitude value of each point is taken and encoded, this process is called quantization.

Radio broadcasting uses audio processing, but it is hard to deal with because many problems are faced. A continuous stream of data is represented by analog signal. It
can also be represented by electrical parameters like voltage, current, etc [13]. The act of processing an analog signal deals with changing the signals through altering the electrical parameters. Nowadays the scenario has changed; processing a non-analog signal is being more prioritized [14]. A sequence of binary numbers represents a pressure waveform in case of digital processing. This process is more efficient.

The terms which are required in a waveform illumination of the signal is the signal itself and the noise. One can define a waveform to be any data that is raw and has to be analysed as signal [15]. The human ears accomplish the task of analysing the frequency of signal, thereby allowing the listeners to differentiate minute differences relating the parameters like time as well as frequency representing the audio [16].

1.6 Challenges in Environmental Sound Data

Classifying and getting these acoustic problems solved experience certain challenges. Unlike speech or music signals, environmental acoustic signals are difficult to model due to its high unpredictable nature. Speech or music can be categorized to structured sounds due to their formantic or harmonic structure characteristic whereas environmental sounds, on the other side, are typically unstructured, which have a broad noise-like flat spectrum and diverse variety of signal composition and are difficult to build models. Analysis of real-world audio that consists of a rich mix of naturally occurring sounds such as the environmental sound is complex. As a result, classification and processing of environmental sound is generally more cumbersome compared with that of speech or music.

- **Recognition:** Automatic recognition of environmental sound using computers is a challenging issue. This is because systems that operate in a real world environment have to process ambiguous and noisy input. Current techniques for
sound recognition are mostly designed for specific applications, such as automatic speech recognition. Therefore, they can rely on assumptions about the audio signal, such as the signal being undistorted and of a known type. However, these assumptions cannot usually be met in a real-world environment.

- **Noise**: Sound is usually adulterated with unwanted sounds which emanate from the surroundings. Echoes, singing, sound from a fan, etc. contribute to noise. Noise is simple and unwanted sound existing in the environment.

- **Channel Variability**: Variability of sound depends on location of the source because the changeability of code is a major issue. Channel variability is a problem that occurs when the sound gets contaminated by several other sounds.

- **Realization**: One can never perceive the sound same as before even if the same sound is produced, because human brain can’t accept it as similar to the previous one.

### 1.7 Issues in Classification and Indexing of Environmental Sound Data

#### 1.7.1 Acoustic Feature Extraction

Features that are extracted from audio are denoted by the acoustic features. Feature extraction is defined as the process of converting an audio signal into a sequence of feature vectors carrying characteristic information about the signal which are used as basis for various types of audio analysis algorithms. It is typical for audio analysis algorithms to be based on features computed on a window basis. These window based features can be considered as short time description of the signal for that particular moment in time. Difficulty arises due to limitations of the existing feature extraction techniques.
1.7.2 Environmental Sound Source Separation

Humans can typically and easily differentiate between different sources. Before reaching the human ear, individual natural sounds will become part of a complex mixture of various sound sources. Depending on the environment, this mixture may contain the sum of many different sources. The human auditory system has the ability to recover individual sources [17]. Sound Source separation is the process of recovering individual sources or signals, from a mixture containing a number of signals. While source separation is a difficult and ill-defined mathematical problem, the topic is highly motivated by many outstanding problems in audio signal processing and machine learning.

Environmental Sound source separation may be regarded as one aspect of a more general process of auditory organization of these simple structures, which is able to untangle an acoustic mixture in order to retrieve a perceptual description of each constituent sound source [18]. In blind separation there is no available a priori knowledge concerning the exact statistical distributions of the source signals; no available information about the nature of the process by which the source signals were combined (mixing process). In reality, some assumptions must be made regarding the source signal distributions and a model of the mixing process must be adopted. However, these assumptions remain fairly general without undermining the strength of the method.

1.7.3 Modeling

Environmental sound segmentation and classification can provide powerful tools for content management. Modern computer and multimedia application take audio as their integral part. A lot of sounds are present which constitute the overall content of the database. Classifying the environmental sound to get the idea of the category to which our content belongs is one of the major fields of research mining [18] [19].
Content-based classification and retrieval of audio sound is essentially a pattern recognition problem in which there are two basic issues: feature selection and classification based on the selected features [20]. The first step involves the reduction of audio into frames. In the second step, classification or categorization algorithms ranging from simple Euclidean distance methods to sophisticated statistical techniques are carried out over these coefficients, to develop one or more reference models for capturing the audio information. The modeling technique has to capture proper audio features and it classifies each set of features corresponding to its own class and determines the efficacy of an audio classification. The performance of an audio classification system depends primarily on the effectiveness of the models in capturing the audio information, and hence this plays a major role in determining the performance of the audio classification system.

1.7.4 Indexing and Retrieval of Environmental Sound

This is one amongst the most complex tasks that have to be employed as indexing finds its use in segmentation and classification [21]. However, even if it is complex, development of applications is very interesting because of their intuitive approach to search and retrieve. Usage of good index algorithms is preferred because it is helpful in reducing search and retrieval time.

Environmental audio recognition and retrieval is a very difficult task, still open to research. The long-term goal is to achieve results comparable to the human sense of hearing. The human auditory sense provides optimal performance, since it is able to bridge the semantic gap. Audio recognition and retrieval techniques can at best narrow the semantic gap. The representation of audio signals by numerical features is currently at a low level of abstraction that does not consider semantic information. Audio retrieval is currently only applicable to a limited domain of sounds.
The retrieval quality decreases [22] rapidly with an increasing number of classes that have to be distinguished. Besides, the quality of retrieval degrades with increase in homogeneity of the audio samples that belong to the same class. Another challenge is the presentation of queries for retrieval systems. Early approaches employed query-by-example techniques. Later, query-by-humming gained importance especially in the field of music retrieval. A retrieval task is always a trade-off between universality and assumptions-about the domain, about the media, and about the user.

Throughout this research work, the environmental sound data collected from BBC and AURORA database, which is sampled at 16kHz and encoded by 16-bit. This database consists of sample from 16 different source recordings of 5 minutes each. The recordings are categorized into general classes according to common characteristics of the scenes (kitchen noises, living room noises, laundry sounds, meeting sounds, office sounds) and events (pan boiling, steel plate, music player, paper scrap, washing machine, flush, overlapped speech, footsteps, typewriter, dust bin, etc.). The recordings are manually labelled and are separated into 2-second, 3-second and 5-second fragments. Every sound signal was stored with some properties that are also the initial conditions and criteria for the well functioning of algorithm. The sample database is split into training sets and test sets. We randomly select 80% sounds of each class for the training set. The remaining 20% sounds form the test set. Each categories having equal proportion in both the train and test set.

1.8 Objective of the Thesis

The objective of Environmental sound classification and indexing system is to classify and index environmental sound data using one or more acoustic characteristics associated with the signal. Environmental sound source separation makes use of acoustic
features such as Spectral and Mel Frequency Cepstral Coefficients (MFCC) to the process of recovering individual sources or signals, from a mixture containing a number of signals. Environmental sound classification systems extract acoustic features such as MFCC and spectral to classify the continuous audio stream into environmental sound categories using modeling techniques. The environmental sound indexing system facilitates the user to have an efficient access, search and browsing capabilities to the desired environmental content respectively.

The objective of this research is to develop automatic environmental sound source separation, classification and indexing system using acoustic features extracted from different categories of environmental audio.

The proposed method of environmental sound source separation, classification and indexing addresses the following issues:

- Acoustic feature extraction from environmental sound data for environmental sound source separation
- Recovering individual sources or signals, from a mixture containing a number of signals
- Acoustic feature extraction for environmental sound classification
- Modeling the acoustic features for environmental sound classification
- Environmental sound clips extraction for environmental indexing
- Environmental sound indexing and retrieval

Acoustic features denote the features extracted from the environmental sound. Environmental sound source separation refers to detecting the individual sources or signals, from a mixture containing a number of signals. Environmental classification system classifies the segmented audio streams into one of the predefined categories.
namely environmental sounds (door bells, alarm signal, etc.). Environmental sound clips are then extracted for each segment of the classified audio for efficient indexing and retrieval.

The framework of the proposed system is shown in Fig. 1.2.

Fig. 1.2: Framework of the proposed system.

1.9 Organization of the Thesis

The focus of the research work presented in this thesis is to develop an automatic environmental sound source separation, classification and indexing system based on
the extracted features which are used for retrieval applications. The contents of the thesis are organized as follows:

A review of methods for environmental sound source separation, classification, indexing and retrieval is given in Chapter 2. Section 2.2 reviews the method of environmental sound source separation techniques. Section 2.3 reviews the method of environmental sound classification. Section 2.4 presents a review of the existing methods for indexing and retrieval of environmental sound.

In Chapter 3, a method is proposed for environmental sound source separation of environmental sound data using the acoustic features. Extraction of acoustic features are described in Section 3.2. Section 3.3 presents the methods for environmental sound source separation using various techniques.

In Chapter 4, various modeling techniques are analyzed for environmental sound classification using acoustic features. The methods of extracting the features are described in Section 4.2. The techniques for sound classification are described in Section 4.3.

In Chapter 5, environmental sound indexing and retrieval using the acoustic features are depicted. Section 5.2 explains the environmental sound retrieval. Acoustic feature extraction for environmental sound indexing and retrieval is described in Section 5.3. Section 5.4 and Section 5.5 describe the creation of index and retrieval technique for environmental sound using GMM.

Chapter 6 summarizes the work presented in the thesis. Also the chapter highlights the contributions of the research and the directions for future research.