Nanyang Technological University

Sound Event Recognition and Classification in Unstructured Environments

A First Year Report
Submitted to the School of Computer Engineering
of the Nanyang Technological University

by

Jonathan William Dennis

for the Confirmation for Admission
to the Degree of Doctor of Philosophy

August 10, 2011

Abstract

The objective of this research is to develop feature extraction and classification techniques for the task of acoustic event recognition (AER) in unstructured environments, which are those where adverse effects such as noise, distortion and multiple sources are likely to occur. The goal is to design a system that can achieve human-like sound recognition performance on a variety of hearing tasks in different environments.

The research is important, as the field is commonly overshadowed by the more popular area of automatic speech recognition (ASR), and typical AER systems are often based on techniques taken directly from this. However, direct application presents difficulties, as the characteristics of acoustic events are less well defined than those of speech, and there is no sub-word dictionary available like the phonemes in speech. In addition, the performance of ASR systems typically degrades dramatically in such adverse, unstructured environments. Therefore, it is important to develop a system that can perform well for this challenging task.

In this work, two novel feature extraction methods are proposed for recognition of environmental sounds in severe noisy conditions, based on the visual signature of the sounds. The first method is called the Spectrogram Image Feature (SIF), and is based on the time-frequency spectrogram of the sound. This is captured through an image-processing inspired quantisation and mapping of the dynamic range prior to feature extraction. Experimental results show that the feature based on the raw-power spectrogram has a good performance, and is particularly suited to severe mismatched conditions. The second proposed method is the Spectral Power Distribution Image Feature (SPD-IF), which uses the same image feature approach, but is based on an SPD image derived from the stochastic distribution of power over the sound clip. This is combined with a missing feature classification system, which marginalises the image regions containing only noise, and experiments show the method achieves the high accuracy of the baseline methods in clean conditions combined with robust results in mismatched noise.

Acknowledgements

I would like to thank my supervisors, Dr. Tran Huy Dat (I^{2}R) and Dr. Chng Eng Siong (NTU) for their invaluable support and advice, which has helped guide me in my research. I also want to thank all my colleagues in the Human Language Technology department of I^{2}R, for their support in a variety of ways, and in particular Dr. Li Haizhou and the members of the speech team from NTU for their feedback and guidance.

Contents

List of Figures

List of Tables

List of Abbreviations

AER	Acoustic Event Recognition
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CASA	Computational Auditory Scene Analysis
DTW	Dynamic Time Warping
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
kNN	k-Nearest Neighbours
LPCC	Linear Prediction Cepstral Coefficients
LVQ	Learning Vector Quantisation
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
NTU	Nanyang Technological University
SIF	Spectrogram Image Feature
SPD-IF	Spectral Power Distribution Image Feature
STFT	Short Time Fourier Transform
SVM	Support Vector Machines
VAD	Voice Activity Detection

Chapter 1

Introduction

The environment around us is rich in acoustic information, extending beyond the speech signals that are typically the focus of automatic speech recognition (ASR) systems. This research focuses on the recognition of generic acoustic events that can occur in any given environment, and takes in consideration the typical problems associated with an unstructured environment that complicate the recognition task.

1.1 Motivation

The goal of computational sound recognition is to design a system that can achieve human-like performance on a variety of hearing tasks. Lyon refers to it as "Machine Hearing" [1] and describes what we should expect from a computer that can hear as we humans do:

> If we had machines that could hear as humans do, we would expect them to be able to easily distinguish speech from music and background noises, to pull out the speech and music parts for special treatment, to know what direction sounds are coming from, to learn which noises are typical and which are noteworthy. Hearing machines should be able to organise what they hear; learn names for recognisable objects, actions, events, places, musical styles, instruments, and speakers; and retrieve sounds by reference to those names.

This field, as described above, has not received as much attention as ASR, which is understandable given some of applications of human-like language capabilities. It is also a very broad topic, and hence has to be split into smaller research topics for any research to be fruitful.

In this research, the focus is on the recognition of acoustic events in challenging real-world, unstructured environments. There are a number of areas which set this research apart from the traditional topic of ASR. Firstly, the characteristics of acoustics events differ from those of speech, as the frequency content, duration and profile of the sounds have a much wider variety than those of speech alone. Secondly, no sub-word dictionary exists for sounds in the same way that it is possible to decompose words into their constituent phonemes. And finally, factors such as noise, reverberation and multiple sources are possible in unstructured environments, whereas speech recognition research has historically ignored these, by assuming the use of single-speakers using close-talking microphones.

The most interesting application for acoustic event recognition is the recognition of sounds from particular environments, such as for providing context for meeting rooms, improving security coverage using acoustic surveillance, or as tools to assist doctors in medical applications. Other applications include environment recognition, to deduce the current location from the received sounds, music genre and instrument recognition, and also the indexing and retrieval of sounds in a database.

1.2 Contribution

A typical acoustic event recognition system is composed of three main processes:

- ***detection:*** to find the start and end points from a continuous audio stream.
- ***feature extraction:*** to compress the detected signal into a form that represents the interesting components of the acoustic event.
- ***pattern classification:*** compares the extracted feature with previously trained models for different classes to determine the most likely label for the source.

Many state of the art systems are based on existing ASR techniques [2], which typically have a high recognition accuracy in clean conditions, but poor performance in realistic noisy environmental conditions. In this thesis, two new feature extraction approaches are proposed for recognising environmental sounds in severe noisy conditions, which utilise both the time and frequency information in the sounds to improve the recognition accuracy. The first method is called the Spectrogram Image Feature (SIF) [3], and is based on the visual signature of the sound, which is captured through an image-processing inspired quantisation and mapping of the signals dynamic range prior to feature extraction. Experimental results show that the feature based on the raw-power spectrogram gives a robust performance in mismatched conditions, although the compromise is a lower accuracy in matched conditions compared to conventional baseline methods.

The second proposed method is the Spectral Power Distribution Image Feature (SPD-IF) [4], which uses the same image feature approach, but is based on an SPD image derived from the stochastic distribution of power over the length of the signal. This is combined with a missing feature classification system, which marginalises the image regions containing the harmful noise information. Experiments show the effectiveness of the proposed system, which combines the high accuracy of the baseline methods in clean conditions with the robustness of the SIF method to produce excellent results.

1.3 Outline of the Document

The report is organised as follows:

Chapter 2 covers the background information on the field of acoustic event recognition, typical applications, and the structure of a recognition system.

Chapter 3 reviews the current state-of-the-art systems for acoustic event recognition, before discussing the important factors that affect the performance of such systems.

Chapter 4 proposes several new methods for robust environmental sound classification, and includes experimental results and discussion on their performance.

Finally, Chapter 5 concludes the report, and also includes information about the directions and schedule of the future work to be carried out.

Chapter 2

Overview of Acoustic Event Recognition

This chapter contains an overview of the field of acoustic event recognition (AER), beginning with the background and comparisons with similar areas of study. Next, a number of applications for AER are described, followed by a discussion of the different building blocks that combine to form a typical acoustic event recognition system.

2.1 Background

Perception of acoustic information is important in allowing humans to understand sounds that they hear in their environment. In this thesis, sounds are treated as "acoustic events", which have properties such as onset, duration and offset times, with a given frequency content that defines the source of the sound. For humans, speech is usually the most informative acoustic event, although the rich variety of acoustic events that occur around us in our environment carry important information and cues and should not be ignored. For example, valuable context can be gained from the acoustic events that occur in meeting rooms, such as people laughing or entering the room.

Noise is a special case of an acoustic event, because it has unique properties, such as a long duration and a spectral content that is consistent over time. In this thesis, I consider any sound that does not conform to these properties as an acoustic event that should not be treated as noise, as it could carry useful information. A simple example of what I consider noise would be background office noise, created by the fans in computers and air-conditioning. Another example could be rain, which consists of many impulsive rain drop

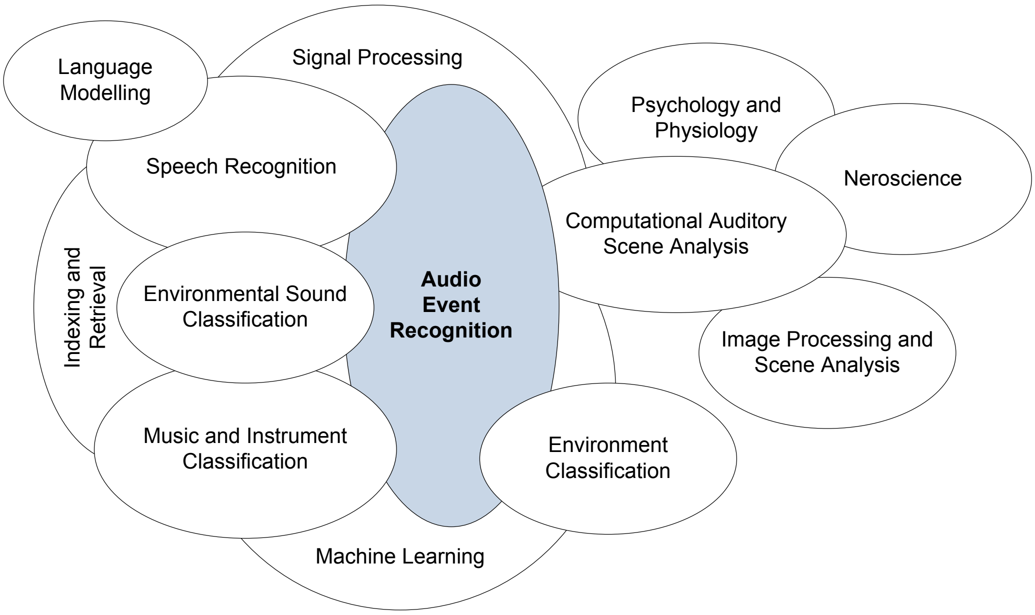
Language Modelling
Signal Processing
Psychology and Physiology
Neroscience
Speech Recognition
Computational Auditory Scene Analysis
Indexing and Retrieval
Environmental Sound Classification
Audio Event Recognition
Image Processing and Scene Analysis
Music and Instrument Classification
Environment Classification
Machine Learning

Figure 2.1: Overview of how audio event recognition (AER) intersects the wider field

sounds that combine together to produce a spectral content close to white noise. However, sounds such as keyboard clicking, footsteps or doors closing, which are often considered as impulsive noise in the field of speech recognition, are here considered acoustic events.

Historically, the field of non-speech acoustic events has not received as much attention as ASR, which is understandable when the applications of robust computer-interpreted speech are considered. The field also suffers from the difficulty in defining exactly where it lies within a larger scope, which means that relevant publications are spread across many disciplines. In addition, authors often concentrate on specific problems, with such diverse examples as identification of bird species [5], medical drill [6] and meeting room sounds [7]. Therefore, it is often difficult to compare different methodologies, as the underlying datasets are so different, unlike the comparatively closed field of speech recognition. There are however some common resources available, and these are given in Section 2.3.

Figure 2.1 shows how the Audio Event Recognition field intersects with others, such as music or environment classification, and how environmental sound classification is often based on methods from the speech and music fields. The underlying fields shown are signal processing, machine learning, and computational auditory scene analysis (CASA). CASA is a relatively new development, beginning with Bregman's pioneering work in 1994 [8], but it brings together ideas that are based on biological evidence to form a quite complete picture of human audio cognition. It is discussed further in Section 3.2.2.

Acoustical Characteristics	**Speech**	**Music**	**Environmental Sounds**
No. of Classes	No. of Phonemes	No. of Tones	Undefined
Length of Window	Short (fixed)	Long (fixed)	Undefined
Length of Shift	Short (fixed)	Long (fixed)	Undefined
Bandwidth	Narrow	Relatively Narrow	Broad Narrow
Harmonics	Clear	Clear	Clear Unclear
Stationarity	Stationary	Stationary (not percussion)	Non-Stationary Stationary
Repetitive Structure	Weak	Weak	Strong, Weak

Table 2.1: Characteristics of Speech, Music and Environmental Sounds

2.1.1 Generic Sound Event Characteristics

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 [9] is shown in Table 2.1, 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. This explains why it is common to define a narrow scope for the problem, such a specific type of sounds, so that at least some of the characteristics can be defined.

A different approach is used in musical instrument classification [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 curve of equal loudness in human perception.

Timbre is the most descriptive characteristic, and it applies equally well to generic sound events as to musical notes. One aspect in particular, the spectral shape, is characterised through Mel-Frequency Cepstral Coefficients (MFCCs) features, and commonly used in speech recognition. In addition, MFCCs attempt to capture the temporal envelope, which is locally estimated through the use of delta and acceleration components. However, the temporal evolution is only captured locally within a few frames, and although MFCCs are often combined with first order Hidden Markov Models, which assume the next state is only dependent on the current state, they may not fully capture the variation over time. This is discussed further in Section 2.1.3.

Among the other two sound characteristics, *loudness* is a universal factor that applies to every acoustic event. *Pitch* on the other hand, may not be applicable to the generic sound event, as there may be multiple, or less well defined fundamental frequencies, which makes the exact pitch difficult to measure.

Finally, it is important to note that the characteristics of a perceived sound depends heavily on the environment in which it is captured. Noise is likely to mask certain portions of the spectral shape, while reverberance depends on the impulse response of the environment and can blur the spectral shape as the signal arrives at different times depending on the shape of the surroundings. Another factor to consider is the frequency response of the microphone, which can present difficulties for AER systems, because when different microphones are attached to the system, the recorded sound can be different, particularly when using less expensive hardware.

2.1.2 Sound Taxonomy

It is a common task in Acoustic Event Classification, to segment a given audio stream, and assign each segment to one class of either speech, music, environmental sounds and silence [11]. It is therefore important to develop a taxonomy, which separates sounds into several groups and sub-groups, so that other researchers can understand the data domain.

An example of a sound taxonomy in the literature is found in [12], and is shown in Figure 2.2. Here, the author splits the class of hearable sounds into five categories, where an environmental sound would fall under the natural or artificial sound classes. The classes are chosen to follow how humans would naturally classify sounds, and examples are given under each to describe the class. This human-like classification can lead to ambiguity. For example, noise is quite subjective and depends on our individual perception. It is also notable that while both speech and music are well structured, natural and artificial sounds

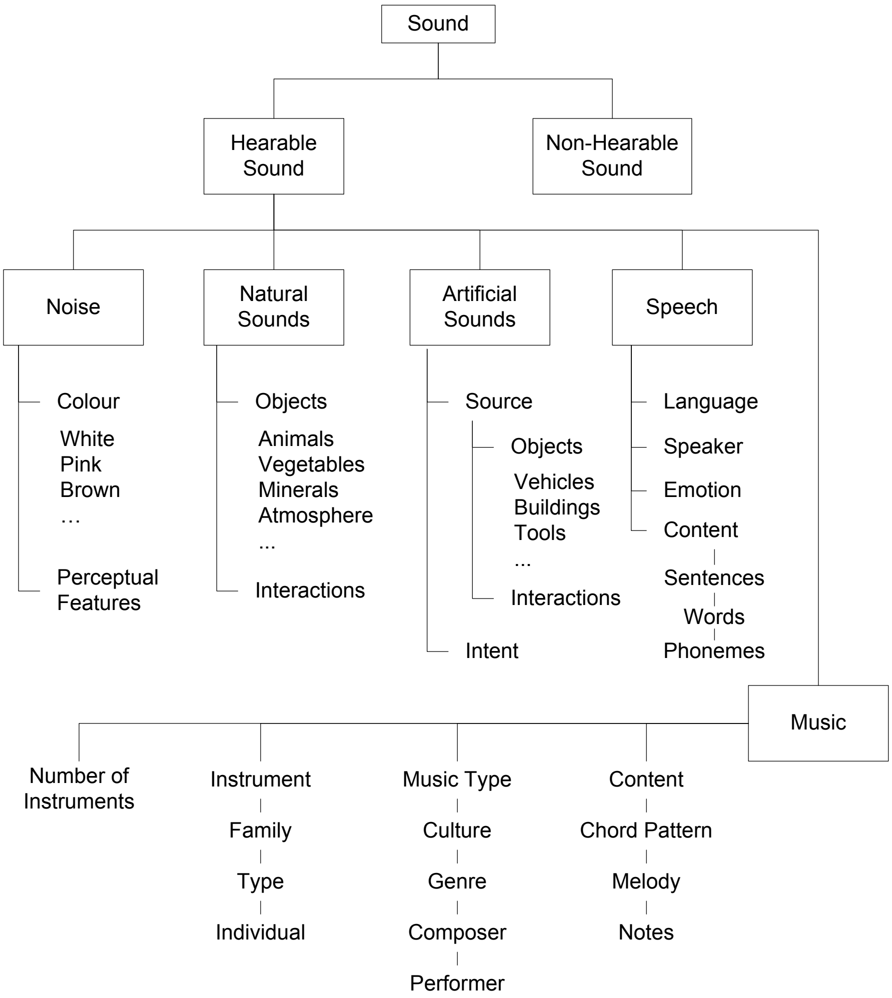

Figure 2.2: Taxonomy for Sound Classification

only act as general groupings, without much internal structure to the class.

An alternative taxonomy for environmental sounds is proposed in [13], to provide some structure to the near-infinite search-space of the classification problem. The proposed taxonomy is based on source-source collisions and physics, such that the sound is produced by two objects interacting, both of which are either solid, liquid or gas. Once the basic type of interaction is determined (e.g. solid-solid, solid-liquid, etc.), a second layer of classification takes place on the sub-groups of the above types (e.g. metal-on-wood, metal-on-water, etc.). A final classification layer then determines the actual interaction from the above sub-group. In this way, a hierarchical structure is formed, which is intended to produce an environmental sound alphabet that has similarities with the phoneme structure of speech.

2.1.3 Comparisons with Speech/Speaker Recognition

Both speech and speaker recognition are relatively mature fields when compared to the broader field of acoustic event recognition. They are all based on similar signal processing concepts, where a detector is first used to segment the continuous audio stream, features are then extracted from the segment, which is then clustered or modelled to provide information about the segment. However, the principles behind the different fields are not the same, which calls for different approaches to the problem:

- ***speech recognition:*** the task is to transcribe continuous speech to text, by classifying phonemes according to previously trained examples. Other aspects which contribute to the variability of speech are also considered interesting, such as speaking rate, emotion and accent [14].

- ***speaker recognition:*** the field contains several sub-topics including speaker verification for biometrics [15] and speaker diarisation, which is the task of identifying "who spoke when" [16]. It requires a model to be created which uniquely identifies the speech of an individual, which does not necessarily require the recognition of individual phonemes or words.

- ***language identification:*** this is commonly used for machine translation [17], and is similar to speaker recognition in that systems may model the variability of the individual languages, rather than recognise the spoken words.

- ***acoustic event recognition:*** the task is to detect and classify acoustic events into the correct sound category, for example speech, music, noise, dog barking or bells ringing, etc. The number of sound categories can be very large, hence they are often constrained by the problem in hand, such as those in a meeting room [7], or in surveillance [18, 19].

The differences between them are related to the scope of the problem at hand. In speech, for example, individual phonemes are not isolated acoustic events. Therefore, it is common to extract acoustic features frame-by-frame, and then use Hidden Markov Models (HMMs) to find the most likely sequence of phonemes for the given features [20]. In the case of speaker recognition, acoustic features are extracted as in speech, but the frames are not decoded into a sequence of words. Instead, it is common to use clustering techniques to identify different speakers, especially as the number of speakers may not be known in advance.

For acoustic event recognition, the foundation is similar, and it is common to use the same acoustic features and pattern recognition systems as found in speech and speaker recognition [21, 22]. However, the scope of the acoustic events is much broader, as it includes environmental sounds, which have a wider range of characteristics, as discussed in Section 2.1.1. In addition, the environments in which the acoustic events occur is considered to be unstructured, meaning there may be background noise, multiple sound sources, and reverberation, which makes the recognition much harder. Therefore acoustic event recognition systems are based on these principles, and often incorporate different techniques.

2.1.4 Applications

In this section, the applications of acoustic event recognition are discussed, where each application is a different topic of research within the wider field shown in Figure 2.1. Most applications are based on a classification task. That is, given a short clip of audio, a recognition system must determine which acoustic event in its training database is the closest match with this new sound. In a slightly different task, the aim is to recognise the music genre, or equivalently the background environment, rather than specific events. This typically requires a longer audio clip than for individual acoustic events. The final application is indexing and retrieval, where an acoustic event can be queried by its audio content.

Environmental Sounds

Environmental sounds are often referred to as “generic acoustic events”, as they sometimes explicitly exclude speech and music. However, I prefer to consider it more literally as “sounds that might be heard in a given environment”, where although speech or musical sounds can be included, their content would not be interpreted by such a recognition system. The environment of interest then determines the scope of the recognition problem, as will be seen in the following examples.

In [7], an evaluation of detection and classification systems in a meeting room environment is carried out. The task was organised for the Classification of Events, Activities and Relationships (CLEAR) 2007 workshop, which is part of the Computers in the Human Interaction Loop (CHIL) project [23]. Here, the acoustic events of interest include steps, keyboard typing, applause, coughing, and laughter, among others. Speech during the meetings is ignored. Several different systems were employed, with one based on the Support Vector Machine (SVM) discriminative approach, and two others based on conventional speech recognition methods using Hidden Markov Models (HMM). The detection/segmentation task was found to be the most difficult, whereas classification of detected segments yielded a reasonable accuracy. It was also found that methods taken directly from speech recognition performed very well, which is why they can be considered state-of-the-art without any modifications for sound event recognition, as discussed later in Chapter 3.

Examples of applications in surveillance can be found in [18, 19]. The first work considers the detection of scream and gunshot events in noisy environments, using two parallel Gaussian Mixture Models (GMM) for discriminating both the sounds from the noise. The authors use a feature ranking and selection method to find an optimal feature vector, which yields a precision of 90% and a false rejection rate of 8%. The second work presents a system for surveillance in noisy environments, with a focus on gun shot detection and a low false rejection rate, which is important in security situations. The authors found that the noise level of the training database has a significant impact on the results, and found it was important to select this for the situation, to give the appropriate trade-off between false detection and false rejection rates.

There are many further publications, often focussed on specific environments. One example, in [6], considers the analysis of the drill sound during spine surgery, as it provides information regarding the tissue and can detect transitions between areas of different bone densities. Another example considers recognition of bird sounds, and proposes a context neural network combined with MFCC features [5].

Environment Classification

Environment classification is the task of recognising the current surroundings, such as street, elevator or railway station, from a short audio clip, and is sometimes referred to as scene recognition. The information gained from the environment can be used for context sensitive devices, which can gain valuable information regarding location and the user's activity [24].

In [25], the authors develop an HMM-based acoustic environment classifier that incorporates both hierarchical modelling and adaptive learning. They propose a hierarchical model that first tries to match the background noise from the environment, but if a low confidence score is recorded, then the segment is matched against specific sources that might be present in the environment. In experiments using only the background environment categories, their system achieves an average of 92%, outperforming human listeners who yielded just 35%.

Another example is found in [26], which considers scene recognition for a mobile robot. The authors used a composite feature set, combined with a feature selection algorithm, and tested the performance with k-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Gaussian Mixture Model (GMM) classifiers. The best overall system is the KNN classifier, which produces a 94.3% environment classification accuracy using 16 features, and has a faster running time than the SVM and GMM methods.

Music

The two main tasks in music identification are instrument recognition and genre classification. There are similarities with environmental sounds, as the first problem involves classification of sounds from a mixture with multiple sources. Genre recognition, on the other hand, draws parallels with environment classification.

Instrument recognition requires the identification of instruments playing in a given segment, with some studies considering isolated instruments [10], while more recently multi-instrument, polyphonic music is of interest. The latter is a difficult task, as it requires the separation of overlapping sounds, and the recognition of instruments that can play multiple notes simultaneously. In [27], the authors develop a system inspired by the ideas from Computational Auditory Scene Analysis and from image segmentation by graph partitioning. They use Euclidean distance between an interpolated prototype and the input frame cluster to measure the timbre similarity with the trained instruments. On isolated instruments, the system achieves a 83% classification accuracy, although performs poorly for the clarinet and violin. In mixtures of four notes, the system is able to correctly detect 56% of the occurrences, with a precision of 64%. A more recent work by the same group is found in [28], where they use the dynamic temporal evolution of the spectral envelope to characterise the polyphonic instruments.

Music genre classification is the task of determining the type of music being played from a short clip, for example classical, blues, jazz, country, rock, metal, reggae, hip-hop, disco and pop. Perhaps the most famous work on this subject is by George Tzanetakis and Perry Cook [29], with other papers published since then, including [30, 31]. The original paper develops a number of features specifically to separate different music genres. These include timbral texture features and rhythmic content features such as peak detection. A classification accuracy of 61% is achieved, which the authors regard as comparable to the human subjects, who achieved an accuracy of 70%.

Indexing and Retrieval

There is a requirement to be able to easily access audio stored in databases that comes from a variety of sources. There are three main topics of search, which are *segmentation*, to determine the start and end point of the event, *indexing*, which is the storage of information distinguishing it from other types of events, and *retrieval*, where a user queries for all events of a given type [32].

The most important aspect is in developing a feature set that uniquely defines a type of event, and can be queried with a range of descriptors, such as physical attributes, similarity with other sounds, subjective descriptions, or semantic content such as spoken text or score [33]. This is what makes it different from a simple classification problem.

There have been a number of publications recently on the topic that focus on environmental and musical sounds. In [34], the authors try to improve the retrieval of environmental sounds by applying techniques that can compare the semantic similarity of words such as “purr” and “meow”, to improve the retrieval results. In [35], a set of “morphological descriptions”, which are descriptions of the sound’s shape, matter and variation, have been considered to form a feature for indexing and retrieval. Another work in [36], presents a “query-by-text” system that can retrieve appropriate songs based on a semantic annotation of the content, including music and sound effects.

2.2 Fundamentals of Acoustic Event Recognition

The task of the recognition system is to detect acoustic events from an audio sample, and determine the most suitable label for that event, based on training carried out on similar samples. Such a system can be categorised as being either "online" or "offline":

- ***online system***: must detect acoustic events as they occur and process them in real-time to produce the closest match.
- ***offline system***: processes audio in batches, such as to produce a transcription for an audio retrieval application.

Such "live" online systems are typically more challenging, as they need to make decisions on relatively short sound clips, for applications such as surveillance or robot environment classification, as discussed in the previous section. In addition, heavy computation often cannot be tolerated, so that such systems can be used on portable devices with minimal computing power.

A typical live recognition system is composed of the processes shown in Figure 2.3. The goal is to take a continuous audio signal, extract short audio clips, each containing a single acoustic event, then extract useful information from the event to create an acoustic feature for classification. During training, this acoustic feature is used to train an acoustic model, which captures the information for each different type of acoustic event. During testing, pattern classification takes place to match the unknown acoustic features with the acoustic model, to produce a label for the event.

The following sections discuss each process in more depth.

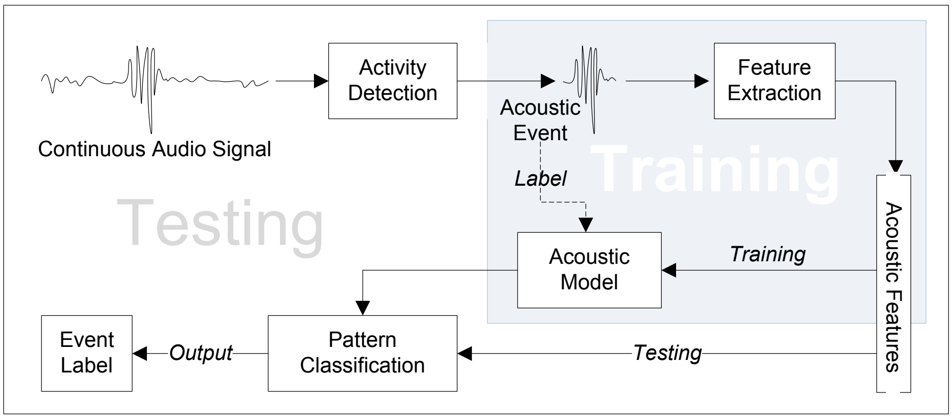

Figure 2.3: The structure of an acoustic event recognition system

2.2.1 Activity Detection

Activity detection concerns finding the start and end points of acoustic events in a continuous audio stream, so that the classification system only deals with active segments. The process can also be called *acoustic event detection* or *voice activity detection* (VAD), especially in the case of speech/non-speech segmentation. This is only required in the case of live audio, when the input is a continuous audio stream. However, while it important that a system can function in real-time, it is common to have sound clips containing isolated acoustic events for the evaluation of an acoustic event recognition system during development. In this case, the detection of the acoustic events is not required.

The outline of a typical system, taken from [37], is shown in Figure 2.4, where features are first extracted from the continuous signal, then a decision is made, followed by post-processing to smooth the detector output. Algorithms for the decision module often fall into two groups: frame-threshold-detection, or detection-by-classification [38]. The former makes a decision based on a frame-level feature to decide whether it contains activity or noise. The decision module is simply a threshold, whereby if the feature output is greater than a defined value, then the decision is positive. The detector does not make a classification of the signal, as with the detection-by-classification method. Instead, it extracts the active segments from the continuous audio stream to pass into a classification system. The simplest feature for such a system could be the frame power level, where if the total power in a given frame exceeds a threshold, the frame is marked as active. However, this is very simplistic and is prone to errors in non-stationary noise. Other possible features include pitch estimation, zero-crossing rate or higher-order statistics, with further improvements in performance reported if features use a longer time window, such as with spectral divergence features [37]. The advantages are low computational cost and real-time processing, but there are disadvantages such as the choice of threshold, which

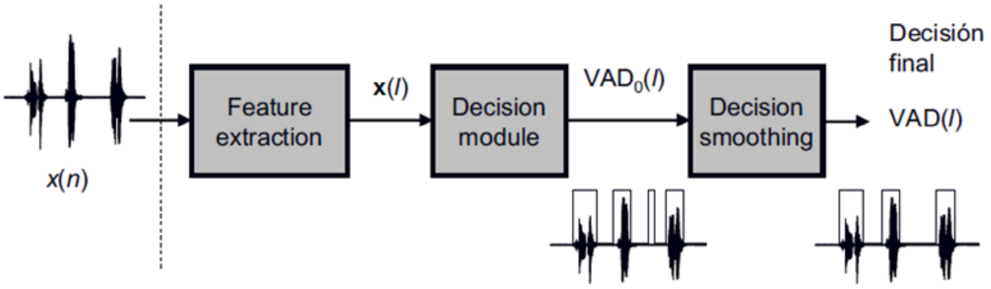

Figure 2.4: Block Diagram of an Activity Detector

is crucial and may vary over time, and the size of the smoothing window, which can be relatively long to get a robust decision.

Detection-by-classification methods do not suffer such problems, as a classifier is used to label the segment as either noise or non-noise, rather than using a threshold. In this configuration, a sliding window is passed over the signal and a set of features extracted from each window. These are passed to a classifier which is trained to discriminate between noise and other non-noise events. The classifier must go through a training phase, where it must learn what features represent noise. A simple, unsupervised decision system could be based on clustering and Gaussian Mixture Modelling (GMM). Here, a short audio segment, containing both noise and non-noise, is clustered in the training phase, so that one cluster should contain the noise, and the other cluster the events. A GMM can then be fitted to the distributions of each cluster, so that future frames will be compared with each GMM, and the most likely one chosen as the label. More advanced classification schemes are of course possible, such as incorporating decision trees and feature selection, which can improve performance [39].

2.2.2 Feature Extraction

The purpose of feature extraction is to compress the audio signal into a vector that is representative of the class of acoustic event it is trying to characterise. A good feature should be insensitive to external influences such as noise or the environment, and able to emphasise the difference between different classes of sounds, while keeping the variation within a given sound class small. This makes the task of classification easier, as it is simple to discriminate between different classes of sounds that are separable.

There are two approaches in extracting features which vary according to the time extent covered by the features [40]. They are either *global*, where the descriptor is generated over the whole signal, or *instantaneous*, where a descriptor is generated from each short time frame of around 30-60 ms over the duration of the signal. This second method is the most popular feature extraction method, and is often called the *bag-of-frames* approach [41]. The sequence of feature vectors contains information about the short-term nature of the signal, hence needs to be aggregated over time, for example using HMMs.

As speech recognition is the dominant field in audio pattern recognition, it is common for features developed for speech to be directly used for generic acoustic events. The most popular feature is the Mel-Frequency Cepstral Coefficients (MFCC), although others such as Linear Prediction Cepstral Coefficients (LPCC) are also used. However, there are many ways in which the signal can be analysed [42], hence there are a wide variety of other features that have been developed to capture the information contained in the signal. These usually fall into the following categories [40]:

- ***temporal shape:*** characterises the change of the waveform or spectral energy over time.
- ***temporal:*** capturing the temporal evolution of the signal, such as zero-crossing rate or auto-correlation coefficients.
- ***energy:*** instantaneous features to capture the various energies in the signal.
- ***spectral shape:*** instantaneous features often computed from the short time Fourier transform (STFT) of the signal. Includes MFCCs, skewness, kurtosis, roll-off, etc.
- ***harmonic:*** instantaneous features from sinusoidal harmonic modelling of the signal.
- ***perceptual:*** features that consider the human auditory process such as loudness, sharpness, etc.
- ***time-frequency:*** global features that capture the signal information in both time and frequency.

When approaching an audio event recognition task, it is clear there are a wide range of audio features to choose from. The most common solutions to the task are to either use prior knowledge about the signal and the performance of individual features to choose a feature set, or to use a feature selection algorithm, which evaluates the recognition performance for a large number of features and chooses automatically those that perform best. An example of this, in [41], uses a library of 76 elementary signal operators, and a algorithm to explore the feature space containing up to 10 operator combinations, which gives 5×10^{20} possible combinations.

2.2.3 Pattern Classification

After feature extraction, the patterns can be classified as belonging to one of classes presented during training, and a label applied to the audio segment. As mentioned earlier, it is important for features of different classes to be distinct and separable, as it is not possible for a classifier to distinguish between overlapping sound classes. The most important classification methods use Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Support Vector Machines (SVM), which are discussed in more detail below, although there are other useful methods that are summarised as follows [2]:

- ***k*-Nearest Neighbours (*k*-NN):** a simple algorithm that, given a testing pattern, uses the majority vote of the k nearest training patterns to assign a class label. It is often described as a lazy algorithm, as all computation is deferred to testing, and hence can have a slow performance for a large number of training samples. For the case of the 1-NN, the method has a 100% recall performance, which is unique.
- **Dynamic Time Warping (DTW):** this algorithm can find the similarity between two sequences, which may vary in time or speed. This works well with the *bag-of-frames* approach, as it can decode the same word spoken at different speeds. However, it has largely been superseded by HMM for Automatic Speech Recognition (ASR).
- **Artificial Neural Networks (ANN):** this method, also referred to as a Multi-Layer Perceptron (MLP), is a computational model inspired by neurons in the brain. Given a sufficient number of hidden neurons, it is known to be a universal approximator, but is often criticised for being a black-box, as the function of each neuron in the network is hard to interpret. It also suffers from difficulty in training, as the most common method of back propagation is likely to get stuck in a local minima.

Hidden Markov Models (HMM)

A Markov model consists of a set of interconnected states, where the transitions between states are determined by a set of probabilities, and the next state only depends on the current state of the model. For a Hidden Markov Model, only the output or observation from each state can be seen to an observer. Hence, knowing the transition and output probability distributions, from a sequence of observations we need to calculate the most likely sequence of states that could account for the observations [20].

Mathematically [43], we first define q_t to denote the hidden state and Y_t to denote the observation. If there are K possible states, then $q_t \in \{1, \ldots, K\}$. Y_t might be a discrete symbol, $Y_t \in \{1, \ldots, L\}$, or a feature-vector, $Y_t \in \mathbb{R}^L$.

The HMM parameters are the initial state distribution, $\pi(i) = P(q_1 = i)$, the transition matrix, $A(i,j) = P(q_t = j|q_{t-1} = i)$, and the observation probability distribution $P(Y_t|q_t)$.

Commonly for speech recognition, the observations Y_t are real vectors, and hence the observation probabilities, $P(Y_t|q_t)$, are a mixture of Gaussian components, commonly referred to as a Gaussian Mixture Model (GMM). This can be written as:

$$P(Y_t = y|q_t = i) = \sum_{m=1}^{M} P(M_t = m|q_t = i)\mathcal{N}(y; \mu_{m,i}, \Sigma_{m,i})$$

where $\mathcal{N}(y; \mu, \Sigma)$ is the Gaussian density with mean μ and covariance Σ evaluated at y:

$$\mathcal{N}(y; \mu, \Sigma) = \frac{1}{(2\pi)^{L/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

and M is the number of Gaussians, M_t is a hidden variable that specifies which mixture component to use, and $P(M_t = m|q_t = i) = C(i,m)$ is the conditional prior weight of each mixture component.

An HMM is trained using a set of observation vectors, O, to determine the above parameters, together denoted θ. The Expectation-Maximisation algorithm is used, which starts with an initial guess of θ, and then performs the E (expectation) step, followed by an M (maximisation) step. This tries to maximise the value of the expected complete-data log-likelihood:

$$\theta^{k+1} = \arg\max_{\theta} P(O|\theta^k)$$

Testing requires the decoding of an observed sequence of vectors to find the most probable state sequence that could have generated them. This process is called Viterbi decoding, and the most probable state sequence can be written as follows:

$$q_{best} = \arg\max_{q} P(O, q|\theta) = \arg\max_{q} P(O|q, \theta).P(q|\theta)$$

HMMs are popular with ASR, as they can model the time evolution of the speech vectors, and can work for sub-word language models.

Support Vector Machines (SVM)

SVM is a binary classifier that calculates the separating hyperplane between two clusters of points in a high-dimensional space. Conventional SVM considers the linear separation of two classes, however modifications add support for overlapping data, non-linear kernel mappings, and solutions for multi-class problems [44]. It is often used in online applications, due to its fast classification performance, although it is not popular for speech classification because it does not naturally model the time evolution of the signal like HMMs.

The separation problem is depicted in Figure 2.5, taken from [45]. We are given m points, each of n dimensions, which can be represented as a $m \times n$ matrix A. Each point is assigned a label in the $m \times m$ diagonal matrix D, which has ±1 indicating the class. We want to find the hyper-plane that best separates the two clusters. As shown in Figure 2.5, w is the normal to the bounding planes $x'w = \gamma + 1$ and $x'w = \gamma - 1$, which separate most of the training data. When the two classes are linearly separable, which is not the case in this figure, the two planes bound the points of each class entirely, and the separating hyperplane $x'w = \gamma$ is midway between the two.

The space between the two hyperplanes is called the "margin", and the aim is to maximise this margin to produce a better classification decision. The margin is found using geometry to be $\frac{2}{\|w\|}$. Hence, the task is to minimise $\|w\|$, while keeping points from falling into the margin. This problem can be solved by the following quadratic program:

$$\min_{(w,\gamma,y)\in R^{n+1+m}} \nu\vec{1}'y + \tfrac{1}{2}w'w \qquad s.t. \quad D(Aw - \vec{1}\gamma) + y \geq \vec{1}, \quad y \geq 0$$

where a suitable value for the SVM parameter $\nu > 0$ must be chosen [45].

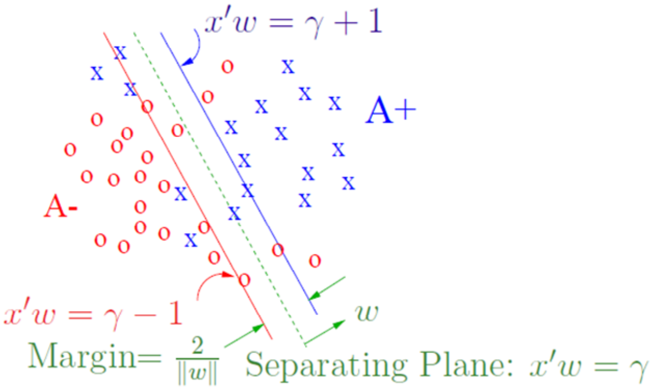

Figure 2.5: Standard SVM problem description, showing two overlapping data clusters and the separating SVM hyperplane.

2.3 Resources for Sound Recognition

2.3.1 Databases

The following is a list of the databases containing events for acoustic event recognition:

- Real World Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [46]. 105 non-speech sound classes categorised as collision, action and characteristic. Also contains data from microphone arrays, room impulse responses and diffuse background noise.

- Classification of Events, Activities and Relationships (CLEAR) 2006/2007 [7, 47]. 12 events taking place in a meeting room environment including door knock, steps, chair moving, paper work, phone ring, applause, laugh, etc. Also includes “speech” and “unknown” classes which were to be ignored in the recognition task.

- Computers in the Human Interaction Loop (CHIL) Evaluations 2005 [48]. A precursor to the CLEAR evaluations using similar acoustic events, but in an isolated event database.

- Sound Effects Libraries, including the BBC Sound Effects Library, Sound Ideas Series 6000, and Best Service Studio Box Sound Effects, which contain a wide variety of isolated acoustic events. Further, the Internet is a rich source of finding acoustic events, including the Database for Environmental Sound Research and Application (DESRA) [49].

2.3.2 Related Projects

The are various projects and research groups that are interested in the topic of acoustic event recognition. A selection of these are introduced below:

- Computers in the Human Interaction Loop (CHIL). The project aims to develop computer systems that can take care of human needs, without being explicitly being given certain tasks. The project supported acoustic classification tasks from 2004 until 2007.

- National Institute of Standards and Technology (NIST). This group supports the CLEAR and TRECVid Evaluations, with many others which are targeted at speech and speaker recognition. Recent TRECVid evaluations, although largely focussed on video, have included audio content as part of their tasks. Particularly, the surveillance event detection may be useful for evaluating acoustic surveillance, although the actual TRECVid tasks are currently not acoustic-based [50].

- Acoustic Computing for Ambient Intelligent Applications (ACAIA) group [51]. The group aims to develop systems that can advance the sensing of mobile robots in their environments, through the processing of environmental sounds and noises in hearing environments.

- Speech-acoustic scene analysis and interpretation (SHINE) research unit [52]. Their aim is to tackle and solve real-world problems including speech recognition in noisy environments, and more generally acoustic scene analysis.

- Self configuring environment aware intelligent acoustic sensing (SCENIC) project [53]. The project aims to enable acoustic systems to become aware of the environment that they operate in, to ensure they are adaptable to the response of different surroundings, such as reverberation.

Chapter 3

State of the Art

3.1 Overview

Although the development of audio recognition systems began over fifty years ago, the focus has largely been on the recognition of human speech, with the topic of acoustic event recognition not receiving as much attention from the research community. The last decade has seen an increase in the popularity of the topic, although most of the studies involved applications on a limited set of environmental sounds [5, 6, 18]. In this chapter, several popular acoustic event recognition techniques are first introduced, which are usually based on existing speech recognition techniques [2, 22, 54], inspiration from auditory models, particular time-frequency features, or the use of particular classification algorithms. The rest of the chapter discusses the factors affecting the performance of such systems, and finally, leads to conclusions on the current directions of research within the field.

3.2 Current Techniques

3.2.1 Using Speech/Music Recognition Methods

The most popular speech recognition techniques often use a combination of MFCC features with an HMM classifier [14], although other features and classifiers, such as ANNs, can also be used. However, the popular MFCC-HMM method often performs well, as it combines a compact representation of the frequency spectrum, through the MFCCs, with a classifier than can model the temporal variation of an acoustic event through the transitions between different states of a model. This method has been shown to work well with environmental sounds, as well as speech. In [25], the classification of acoustic environments is on average 92%, using MFCC features and HMM classifier. For individual acoustic events, the overall accuracy is 85%, using a database of 105 action, collision and characteristic sounds. Other examples in [54, 24], again show that the combination of MFCC and HMM performs well for environmental sound classification.

However, it was suggested by Cowling, in [2], that HMM based techniques are not suitable for acoustic event recognition, due to the lack of an "environmental sound alphabet". This statement, although partially true, is in my opinion erroneous. While it is true that there is no such sound alphabet, it is simple to train HMM models for each class of acoustic event, rather than trying to define sub-sound units that are common across different sounds. The speech recognition analogy is to train an HMM model for each word, as opposed to training a model for each sub-word unit, which is only possible due to the existence of a fixed alphabet. Therefore, for environmental sound recognition, an HMM is trained for each sound class, which is used to decode unknown sounds in the environment.

Cowling instead presents an analysis of two other speech recognition approaches: Learning Vector Quantisation (LVQ) and ANNs. The LVQ technique is based on generating a prototype for each class during training, by adapting the prototype to be closer to each successive example, then testing involves finding which prototype is closest to the observed example. It is dependent on the distance metric used, hence has not gained the popularity of the HMM approach. The results show that LVQ has similar results in both speech and non-speech tests, while ANN performs well for speech, but poorly for the non-speech tests. Cowling suggests this is due to the similarity of the sound classes used in the experiment, and that LVQ is a better classifier for similar sound types. However, this result should be interpreted carefully, as it is known that ANNs, given a sufficient number of hidden neurons, are universal approximators, and should be able to represent an arbitrary boundary between two classes.

Other authors have successfully used ANNs for acoustic event classification. In [22], MFCC features vectors are preprocessed using vector quantisation, and fed to an ANN for classification. The results show an average of 73% accuracy across ten classes of environment sounds. Another example is found in [5], where bird species recognition is based on MFCC features combined with ANN classifier. The authors acknowledge the use of MFCC features in speech recognition, but suggest they are appropriate for bird sounds due to the perceptual qualities of the Mel-filter, their data-reduction properties, and that they can characterise both aperiodic and periodic signals. The authors report an 87% accuracy

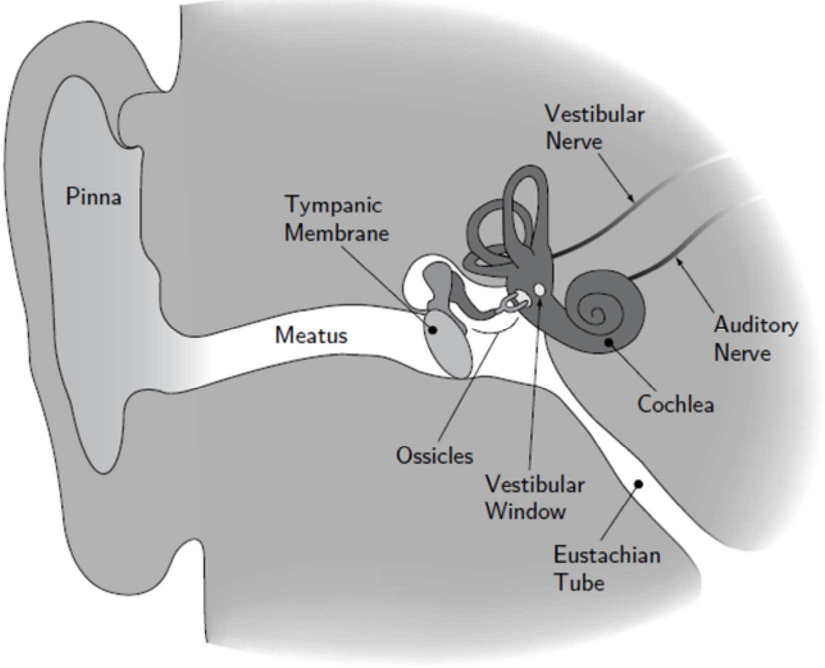

Figure 3.1: The key components of the human ear

across 14 species of birds.

3.2.2 Computational Auditory Scene Analysis (CASA)

Auditory Scene Analysis (ASA) is a term coined by Albert Bregman in his 1990's work on human perception of sound [8]. The term "scene analysis" is used in image processing to encapsulate the problem of describing the contents of a picture of a three-dimensional scene [55]. Bregman's interpretation of human perception is that of building up a picture of the "auditory scene", which is quite neatly encapsulated within the example of the "cocktail party" scenario. Here, a human listener is in a room with many conversations taking place, at different volumes, and in the presence of other acoustic distractions, is easily able to follow their conversation with a friend. Even in the presence of severe noise, the human auditory system is able to filter is out to focus on a single auditory stream. This is equally important in sound event classification, since only simulated sounds under laboratory conditions will occur in perfect, isolated conditions. In practise, a listener will always hear sounds overlapping with other acoustic events and noise, and therefore the idea of an "auditory scene" can be equally applied to sound event recognition.

Figure 3.1 (from [56]) shows the human auditory system, with the first stage beginning with the inner ear, where a frequency analysis is carried out inside the cochlear by the basilar membrane . The properties of the membrane vary along its length, causing different regions to be tuned to have different resonant frequencies. The gammatone filterbank, originally proposed in [57], provides an accurate model for the response of the basilar membrane, and is also computationally efficient. The filter bandwidths and centre frequencies are distributed along a warped frequency scale, called the Equivalent Rectangular Bandwidth (ERB) rate scale, which was found from human psycho-acoustic experiments [58]. This approach has been shown to work well for sound classification, with several different classification methods, and in some cases can perform better than conventional MFCC features [59, 60, 61].

The next stage of processing in the cochlea comes from the inner hair cells (IHCs), which convert the movement of the basilar membrane into neural activity that is transmitted to the brain via the auditory nerve [56]. A popular model of this process was proposed by Meddis, which, although potentially controversial, provides a useful simulation of the process of the hair cell firing [62]. The model converts the displacement of the basilar membrane to probability of a spike event occurring, by simulating the production and movement of neural transmitter in the cell. It is represented through the cochleagram, which is similar in appearance to the transitional spectrogram, and shows the simulated auditory nerve firing rate for each gammatone filterbank over time. This approach is commonly used in auditory perception, especially in instances of multiple concurrent sources, where biologically motivated methods are often employed [63].

An alternative inner ear model proposed by Lyon can be found in [64], which models the most important mechanical filtering effects of the cochlear and the mapping of mechanical vibrations into neural activity. The main component of the method is a parallel/cascade filterbank network of second order sections, which is a simplification of the complex behaviour of the basilar membrane using simple linear, time-invariant filtering. It is found that this model does a good job at preserving important time and frequency information, which are required for robust audio recognition, and is discussed in more detail below.

In his pioneering work on CASA, Bregman used a series of experiments, using simple tones, to discover the cues that the auditory system uses to understand sounds. He found there were two main effects present: *fusion*, where sound energy from different areas of the frequency spectrum blend together, and *sequential organisation*, where a series of acoustic events blend together into one or more streams [65]. He found that a common onset, within a few milliseconds, was the main cue for fusion, while learned schema appeared to be the main cue enabling allocation of different frequency elements to the same sound source. Therefore, CASA is now commonly used for separating speech signals from noise, which allows a spectral mask to be generated, and only those reliable elements will be used for classification [66].

3.2.3 Time-Frequency Features

The Fourier transform is the most popular signal processing method for transforming a time series into a representation of its constituent frequencies. For continuous audio, the Short-Time Fourier Transform (STFT) is often used, which uses a window function, such as the Hamming window, to split the signal into short, overlapping segments, before applying the Fourier transform. This is the approach used for extracting MFCC features, but has a drawback due to the compromise between time and frequency resolution. Increasing the length of the analysis window allows a finer decomposition of frequencies, but at the cost of reducing the resolution of the sounds over time. To overcome this, there are other methods, such as wavelet transforms, which can produce a higher resolution representation, without compromising either the time or frequency dimensions.

In general, the varied nature of the acoustic signals means that representing their frequency content alone may not sufficient for classification, and that there is also important information in the temporal domain [67]. Common environmental sounds, such as impacts where two materials come together, often have an initial burst of energy that then decays. For example, dropping a coin might cause an initial impact sound, closely followed by a ringing from the vibration of the coin. Other sounds have a more complex structure, such as that of breaking glass or footsteps. Therefore, it is important to capture the variation of the frequency content over time, in order to fully characterise the sound. MFCC features represent the frequency content of the sound at a particular instance in time, and therefore need to be combined with complex recognisers such as HMMs, which can model the temporal variation of these stationary features.

This approach works well in some circumstances, but recently another approach has been explored where the time-frequency nature of the sounds are inherently captured in the feature for classification. This has the advantage of fully characterising the evolution of the sound event, rather than simply capturing its frequency information an individual points in time. There are two common approaches to time-frequency feature extraction. The first attempts to extract sound descriptors, such as the duration, attack time, pitch and dynamic profile from the sound, to naturally describe the most characteristics aspects, while the second represents the sound event as a two-dimensional “auditory image” and extracts a mixture of local and global features.

The former approach has been used previously for both classification and indexing [35, 68], using a variety of features, that are sometimes described as “morphological descriptors”. The advantage of the descriptor approach for sound indexing is that the result is independent of the sound source, and therefore can enhance the capabilities of a sound search engine. For example, sounds that have an increasing dynamic profile over time can create tension or excitement for the listener. Hence a search for “tension” might then give a heavy weighting towards this particular descriptor, and matching sounds will have a slowly increasing profile and fewer sounds will have a flat profile which might be better used for background ambiance.

For the second approach based on image representations, the spectrogram is most commonly used as a basis to visualise the sound, although different approaches include the wavelet scalogram, for its improved time-frequency resolution, and the cochleagram, due to its biologically plausible representation. Other approaches, such as the one used by Lyon in [1], use a correlogram image to represent the auditory information. This is the short-time cross-correlation of the signal and is found to produce characteristic shapes, particularly in the case of voiced speech, which can be captured to classify the speech information. In [69], the authors use the spectrogram as a texture image, and use a random sampling of the image to extract features from image blocks to capture the local time-frequency structures. The method performs well for musical sounds, particularly when a large number of block features are used. In Chapter 4, a new approach based on the spectrogram image representation of the sound is presented, where time-frequency texture features are extracted after first quantising the dynamic range into regions.

Another, more general approach, is called Matching Pursuit (MP), which decomposes the signal into a variety of representations, contained in a base dictionary, and finds the bases with the most energy which are used to represent the signal [70]. The advantage is that it can be combined with a variety of bases such as Fourier, Haar and Gabor. When Fourier bases are used, the MP algorithm behaves like a Fourier series expansion, so has the advantage that it can also model the Fourier transform. When Gabor bases are used, which are two-dimensional sine-modulated Gaussian functions with the capability of being scaled and shifted, the MP algorithm can extract both time and frequency information. It has been noted that the Gabor bases are more effective at reconstructing a signal from only a small number of bases, hence are often preferred over the one-dimensional Haar and

Fourier bases [68]. The method has also been shown to produce good classification results for environmental sounds, where the time-frequency information is important [9].

3.2.4 Classification Methods

While the most popular classification algorithms were introduced in section 2.2.3, approaches from other fields have been used and can sometimes offer advantages over conventional methods.

One such related field is machine vision and image retrieval, which can be used for acoustic event classification if we consider the time-frequency spectrogram as a colour image. One such method is the "passive-aggressive model for image retrieval" (PAMIR) [71], which uses a ranking-based cost function that aims to minimise the loss related to the ranking performance of the model. This allows it to efficiently learn a linear mapping from a sparse feature space to a large query space. It uses the "bag-of-words" approach, which uses a collection of descriptors, or features, as the input, with no particular ordering, which here is the histogram of the occurrences of the various features in the feature space. The approach is not limited to image retrieval, and has been reported recently, in [1], to perform well with both MFCCs and auditory image feature representations.

Another related field is that of biologically plausible pattern recognition, such as those found in spiking neural networks. These approaches aim to understand the human physiology involved in the auditory system, and build systems that replicate the processing mechanisms that work very effectively for humans. The first stages of such a system were introduced in section 3.2.2, which translate the incoming sound signal into spikes that are transmitted through the auditory nerve to the brain. Much work has been carried out to study the response of the brain to different auditory inputs [72], and a variety of neural recognition systems have been proposed. One example was developed by Hopfield [73], which uses a simple feature from the sound onset, peak and offset times, combined with a spiking neural network for recognition. The system requires that a set of neurons, with certain spiking-rate decay times, are selected during training. Recognition occurs when, at a particular time after the onset of the sound, all the neurons are spiking at the same rate, causing a combined potential large enough to trigger a detector neuron. If an untrained sound is presented, the neuron firing never synchronises, and the combined output remains below the threshold for detection. The system is invariant to time warp, and, while it can only recognise a small dictionary of words, it shows that there are alternative techniques , which provide inspiration to expand the scope of the field of acoustic classification.

3.3 Factors Affecting Sound Recognition

Although the accuracy of typical acoustic event classification systems is high under clean, matched conditions, the performance often decreases in the presence of real-world effects, such as noise, reverberation and multiple sound sources. All three effects naturally occur in environments such as meeting rooms, offices and outdoor environments, and humans are amazingly adept at overcoming these issues. For example, the human brain is able to focus on a single speaker in a room full of competing conversations, which would leave most state-of-the-art speech recognition systems struggling. Such a situation is commonly referred to as the “cocktail party” problem, as introduced in Section 3.2.2, and is often the focus of CASA, which tries to emulate human-like sound segregation and classification.

3.3.1 Noise

Many speech and acoustic event recognition systems are trained with data recorded in environments with a high signal-to-noise ratio (SNR), such as speech with a close-talking microphone. Previous studies, such as [74], have shown clearly that training and testing at different SNRs results in a decreasing classification accuracy with an increasing difference in SNR between the training and testing samples. The decrease in performance is due to distortion of the feature vectors, which may have some key components masked by the noise or changed significantly enough to affect certain feature dimensions. The most popular features for speech recognition, MFCCs, show very little robustness to noise, since the noise can affect the feature unpredictably. This is because the feature dimensions do not represent a particular frequency; rather each represents a mixture of frequencies through the discrete cosine transform of the spectrum.

For MFCCs to perform well in noisy conditions, the recognition system must have been explicitly trained in similar noise and SNR conditions [74]. However, this is unrealistic for real-world applications, as the conditions where the system will finally be deployed are often not known in advance. One solution is to perform multi-conditional training, where the system is trained on a variety of different noise conditions, at different SNRs, so that the acoustic models contain some knowledge of how the signals might be received in a novel noise environment. However, this requires a large amount of training data, and often reduces the recognition accuracy under clean conditions, due to the reduced discrimination between the noisy acoustic models.

Current techniques to improve the performance in mismatched conditions typically fall into three categories:

- ***signal enhancement:*** here the aim is to enhance the noisy signal so that the extracted features are closer to the trained condition. Examples include Wiener filtering and spectral subtraction [75]. There also exists standardised toolkits, such as the ETSI Advanced Front End [76], that implement these approaches, and provide a good baseline for other noise reduction techniques.

- ***feature compensation:*** here, features are extracted from noisy signals, and the aim is to transform the statistics of the noisy features to be closer to those of the clean features. Examples include cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN) [77], which are simple but effective methods for improving the performance in mismatched conditions.

- ***model adaptation:*** here the idea is to adapt the acoustic models, trained on clean data, to better represent the noise present in the signal. Popular methods include parallel model combination (PMC) and maximum likelihood linear regression (MLLR) [78].

3.3.2 Distortion

As well as additive noise, another factor that affects recognition performance is convolutional distortion that are typically caused by room or microphone effects. For the case of room distortion, the effect is commonly referred to as reverberation, and it can alter the spectrum of the signal due to the reflected late arrival of frequency components. For microphone distortion, different types of microphone will have a different frequency response, again altering the frequency spectrum. It has been reported in the past even mild reverberation, often experienced by hands-free microphones, or a simple change of microphone, can have a severe effect on recognition performance [74].

In principle, it appears simple to remove the convolutional distortion, through estimation and filtering of the impulse response of the receiver in a particular environment. However, in practise, the problem is much more difficult, since the system is blind and doesn't have knowledge of the room impulse response or the original, undistorted signal. Several algorithms have been proposed to overcome the problem, and can generally be grouped into two categories [79]:

- ***linear prediction (LP) residual:*** the LP residual contains evidence of the reverberation effects in the peaks due to the excitation events in voiced speech and those caused by the reflections. Many approaches are possible, for example the kurtosis of the residual has been shown as a useful metric to optimise an adaptive filter [80].

- ***blind channel estimation/inversion:*** this often requires multiple channels, and is based on the cross-relation between the two channels, which enables estimation of the underlying channel vectors. Once identified, dereverberation can be achieved by inversion, although this suffers from various problems, such as "spectral nulls" that invert to give a strong peak causing narrow-band noise amplification [79].

3.3.3 Multiple Overlapping Sources

This factor is often overlooked by the speech recognition community, but is a real problem in the field of sound recognition, as most environments do not preclude the chance of two sound events occurring simultaneously. The problem is addressed in computational auditory scene analysis (CASA), where a typical "sound scene", much like an visual scene, may be composed of many different elements mixed together, and the aim is to separate this mixture into its component sources [8].

The problem faced by conventional methods for speech recognition is that training occurs in a controlled environment without any overlap, and the features typically capture the spectral content of the sound. When two sounds occur at the same time, a simple approximation of their interaction is simply to add the two together, which means that the spectrum at a given moment is made up of contributions from each source. Therefore, the conventional system would simply extract a feature from the combined spectrum of the sounds, and try to match this against the existing models. The result would be unpredictable, and may produce a match with either one of the sounds, or none at all.

A solution, within the same framework, is to use parallel, concurrent HMMs, called Factorial HMMs [81]. Here, spectral features, rather than cepstral features such as MFCCs, must be used, to preserve to additive assumption of the sources, which is broken through the use of the cepstral features. Then, the decoding task is to assume that the given observation can be a weighted sum of the current state of more than one model, hence this alters the optimisation task over the traditional Viterbi decoding. The biggest drawback with this approach is the large computation requirement, since any observation could be explained by any of the many possible model combinations.

The principle of CASA is to use a grouping algorithm to collect the various elements of a composite representation together that belong to each source, based on the observed properties and cues, such as having a common onset time [65]. Therefore, such systems naturally handle scenarios where multiple sources are active at the same time. However, their performance is often poor by the standards of speech recognition systems, and currently there is no fully workable solution to achieve human-like source separation.

3.4 Conclusion

This chapter has covered the most popular current techniques used in acoustic event recognition, and introduced the biggest challenges faced in the performance of such systems, namely noise, distortion and multiple sources.

It is clear that the field is still driven by the speech recognition community, as many of the most popular techniques for acoustic event recognition have considerable overlap with speech technology. Other state-of-the-art methods show a recurring focus on both the time and frequency content of the acoustic signal as being important to achieve the human-like classification of sounds. These methods can be based on the physiology of human hearing, or computational methods such as sound descriptors or time-frequency image representations. Overall it is clear that this is an important research direction with many interesting aspects still to explore [1].

Of the three main factors affecting the performance of recognition systems, so far most work has considered only the effects of adverse noise conditions, with investigation of the other two much less common. This is again due to the field of speech recognition, where close-talking microphones reduce the impact of distortion on the performance, and reduce the likelihood of multiple sources overlapping. The field of CASA is the main source of progress for a system that can handle all three effects, but so far progress has been slow, and the focus has largely been on speech occurring in the presence of complex sound scenes, such as on the street or in restaurants. Therefore, it is important to consider each of these issues from the perspective of acoustic event recognition, where the nature of the signals is much broader than that of speech, and where typical usage environments may be considerably more challenging.

Chapter 4

Proposed Image Methods for Robust Sound Classification

4.1 Overview

In this section, two proposed methods for robust sound event classification are introduced. The first is called the Spectrogram Image Feature (SIF) [3], and is based on a visual signature extracted from the sound's time-frequency representation. The second is the Subband Power Distribution Image Feature (SPD-IF) [4], which is based on a novel representation of distribution of spectral power against frequency. The methods are motivated by the distinctive time-frequency characteristics of sound events, which can be identified though the spectrogram image, even in the case of severe background noise.

In the first section, the SIF method is introduced, which takes inspiration from approaches in image processing, such as pseudocolouring and feature based on partitioning and colour layout, to develop a robust image feature that is extracted from the spectrogram. The image feature quantises the dynamic range of the spectrogram into regions, similar to the red, green and blue colours present in human vision, such that each region represents different information contained in the signal. Next, the layout of these quantised regions is captured by first segmenting the spectrogram, before extracting the distribution from each time-frequency block, in terms of their central moments, to use as a feature for classification.

In the next section, the SPD-IF is proposed, which combines a representation that is invariant to the effects of time-shifting, with a noise estimation algorithm and missing feature classifier, to automatically select the robust blocks from the novel subband power distribution (SPD) image. The SPD represents the empirical magnitude distributions of the spectrogram in each subband frequency component, can also be considered as a generalisation of the classical power spectral density. The SPD-IF method uses the same image feature approach as for the SIF, although applies it to the SPD image, rather than directly to the spectrogram.

In the last section, the methods are tested on a large database containing 50 sound classes, under four different noise environments, varying from clean to severe 0dB SNR. Both show a significant improvement in performance over conventional speech recognition methods, such as MFCC-HMM and MFCC-SVM, which are often applied as baseline methods for acoustic event recognition. The best performing SPD-IF method is shown to achieve an average classification accuracy of 87.5% in the 0dB mismatched noise condition.

4.2 Spectrogram Image Feature

The time-frequency spectrogram of a sound contains a large amount of information and provides a representation that can be interpreted by trained human readers [82]. Even in severe noise conditions, certain characteristic shapes are still easily visible in the spectrogram. The SIF extraction method is therefore based on this visual representation of the sound, taking inspiration from techniques commonly used in image processing.

4.2.1 SIF Feature Extraction Algorithm

The algorithm can be described as follows:

1. A grey-scale spectrogram, $G(f,t)$, is generated from the sound, as shown in path (1) in Figure 4.1a.

2. An image feature is extracted from the spectrogram, following the approach shown in Figure 4.1b. This process performs dynamic range quantisation, and image feature extraction.

These steps are described in more detail below.

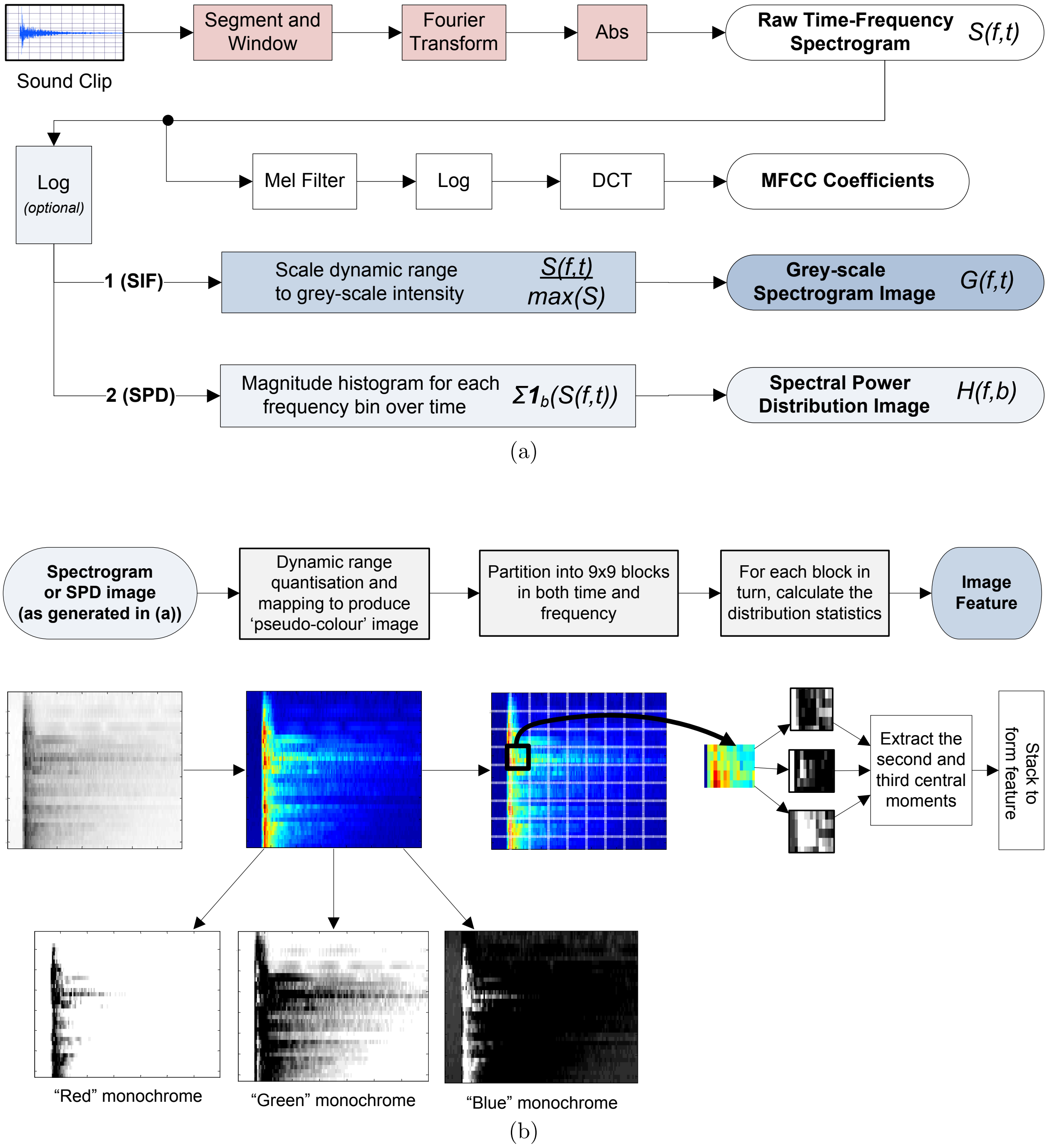
Sound Clip
Segment and Window
Fourier Transform
Abs
Raw Time-Frequency Spectrogram S(f,t)
Log (optional)
Mel Filter
Log
DCT
MFCC Coefficients
1 (SIF)
Scale dynamic range to grey-scale intensity S(f,t) / max(S)
Grey-scale Spectrogram Image G(f,t)
2 (SPD)
Magnitude histogram for each frequency bin over time Σ1_b(S(f,t))
Spectral Power Distribution Image H(f,b)
(a)
Spectrogram or SPD image (as generated in (a))
Dynamic range quantisation and mapping to produce 'pseudo-colour' image
Partition into 9x9 blocks in both time and frequency
For each block in turn, calculate the distribution statistics
Image Feature
Extract the second and third central moments
Stack to form feature
"Red" monochrome
"Green" monochrome
"Blue" monochrome
(b)

Figure 4.1: Overview of the signal processing and image feature extraction for the SIF and SPF-IF methods. (a) describes the signal processing steps for MFCC features, compared to producing the grey-scale spectrogram (for SIF) and the spectral power distribution image (for SPD-IF). (b) describes the procedure for image feature extraction, as used in the SIF and SPD-IF methods.

Grey-scale spectrogram generation

The SIF is based on the spectrogram $S(f, t)$, which is calculated using the discrete short-time Fourier transform (STFT) of a windowed frame $x_t[n]$, where f is the frequency bin, and t is time:

$$S(f,t) = \sum_{n=0}^{N-1} x_t[n] w[n] \, e^{-i2\pi \frac{f}{f_s} n} \tag{4.1}$$

where N is the number of samples per frame, $f = \frac{kf_s}{N}$ for $k = 0, \ldots, N-1$, $w(.)$ is the Hamming window, and t is the frame index, as multiples of $\frac{N}{f_s}$.

At this stage, either the raw magnitude values are used for the raw-power SIF case, or the dynamic range is compressed using a logarithm, for the log-power SIF. For the log-power case, a noise floor is used, set at 0dB: $S(f,t) \to \max\left[S(f,t), 0\right]$, since $\min[S(f,t)]$ becomes highly variable as the values tend towards $\log(0)$.

The spectrogram is then normalised into a grey-scale intensity image by dividing each element by the global maximum in the spectrogram as follows:

$$G(f,t) = \frac{S(f,t)}{\max[S(f,t)]}. \tag{4.2}$$

Dynamic range quantisation and mapping

In this first step of the image feature extraction, the dynamic range of the grey-scale spectrogram is quantised into different regions, each of which maps to a monochrome image. This operation can be seen as a generalisation of the pseudo-colourmapping procedure from image processing, where grey-scale intensities are quantised into red, green and blue (RGB) monochrome components. The mapping performed is:

$$m_c(k,t) = f_c(G(k,t)) \quad \forall c \in (c_1, c_2, ..c_M) \tag{4.3}$$

where m_c is a monochrome image, f is a nonlinear mapping function and c represents the quantisation regions. For the case of pseudo-colourmapping, c represents the three red, green and blue primary colours, from which all other colours can be derived. However, unlike in image processing, the quantisations are not limited to the $M = 3$ regions required in the colourmap. However, it was found from initial experiments that three quantisations was a good trade-off between the accuracy and computational cost, hence it is employed here. An example mapping function for the "Jet" colourmap in Matlab, is shown in Figure 4.2. Here, the grey-scale spectrogram intensities, $G(k,t)$ are mapped into monochrome

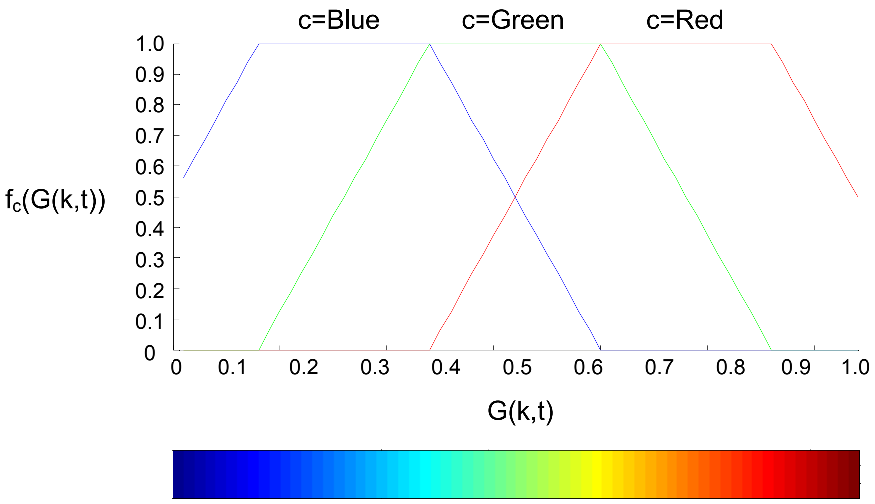

Figure 4.2: Example mapping function $f_c(.)$ for the "Jet" colourmap in Matlab

colour intensities, for each for the red, green and blue colours. It can be seen that the low intensities are largely represented by the blue colour, while the red represents the highest sound intensities, hence should be the least affected by the noise.

Feature extraction

Colour distribution statistics are a commonly used feature in image retrieval, as the colour distribution describes the image content by characterising the variation of colour intensity [83]. Hence, we can similarly exploit the sound spectrogram to characterise the signature of an audio event. Using the sound spectrogram, we can capture the local time-frequency power intensity by partitioning the spectrogram into smaller time-frequency blocks. We model the intensity distribution in each block to produce the feature, which characterises the variation of sound intensity in quantised regions over time and frequency.

We do this separately for each of the quantised regions of the dynamic range, such that the more robust, higher-power regions are better represented in the feature. Therefore, each monochrome image is partitioned into 9 × 9 blocks and the central moments of each block are used to capture the distribution, as follows:

$$\mu_k = E\left[(X - E[X])^k\right] \tag{4.4}$$

where X is the distribution, E is the expectation operator, and μ_k is the k^{th} moment about the mean. Particularly, the second and third central moments were found to produce a good overall performance. It is notable that in preliminary experiments, the classification accuracy was increased when the mean was not used as part of the feature, especially in the case of mismatched conditions. In addition, 9×9 blocks was found to be a good compromise between classification accuracy and feature complexity, hence is employed here. Overall, the final SIF is a 486 (2x3x9x9)-dimension vector, with two central moments and three quantisation regions.

Classification

Here, linear SVM is employed [84], in a One-Against-One configuration with the max-wins voting strategy. Although results are not reported from the conventional Gaussian non-linear kernel, it was found that linear SVM achieved a comparable classification accuracy, with considerably lower computational cost and hence is preferred. It should be noted that if the proposed feature extraction method is seen as as a non-linear transform $\phi\left(x\right)$ from sample x, then the method can be considered as a novel SVM kernel, where $K\left(x_i, x_j\right) = \phi\left(x_j\right)^T \phi\left(x_i\right)$.

4.2.2 Discussion

Example spectrogram images are shown on the left-hand side of Figure 4.3, while the right-hand side shows SPD-IF images which are described in detail later. As discussed earlier, it is still possible to see the underlying signal representation, even in severe 0dB noise. For humans, the pseudo-colouring of the spectrogram makes the most important elements easier to discriminate in the image. Here, the “Jet” colourmap from Figure 4.2 is used, hence the red colour carries the most important information from the high-power frequencies. The information in this red monochrome is characterised in the SIF through the image distribution feature, and should be relatively robust against the background noise.

However, there are two issues in the practical implementation of the SIF method. Firstly, the onset and offset detection of the sound clips from a continuous audio stream can be affected by the noise conditions. Therefore, the time-frequency partitioning may cause variation among the SIF dimensions, due to the time-shifting effect. Secondly, as only certain dimensions of the SIF are robust to noise, particularly those representing the high-magnitude regions, a missing feature framework can be adopted to remove the affected dimensions, which should improve the classification accuracy.

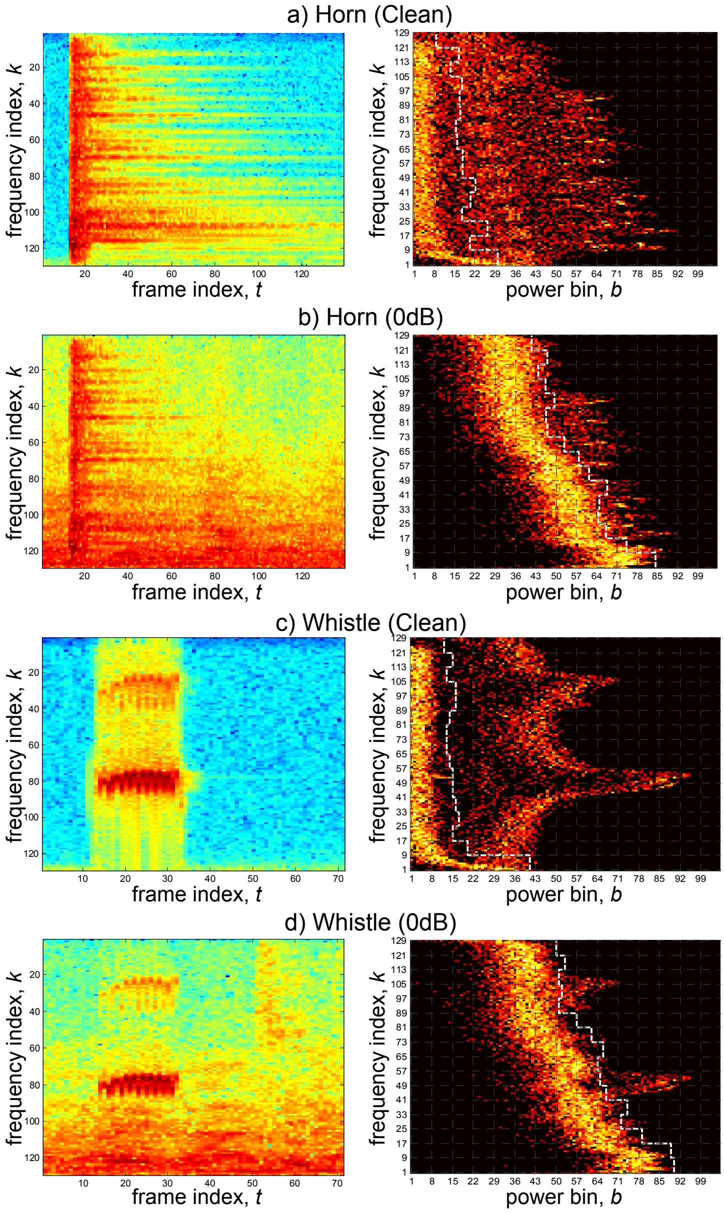

Figure 4.3: Comparisons of the log-power spectrogram (left) and SPD (right) for horn and whistle in both clean and 0dB noise conditions. For SPD, the background grid indicates the segmentation, while the white dashed line is the noise level estimate.

4.3 Spectral Power Distribution Image Feature

In this section, a new representation is described that improves upon the SIF method by being invariant to time-shifting, and easily combined with missing feature methods. This representation is called the subband power distribution (SPD), as it captures the stochastic distribution of power over the sound clip.

The SPD can be considered as a generalisation of the power spectral density (PSD), as the PSD shows the magnitude of the power present in the signal at each frequency bin [85]. However, the PSD does not represent the distribution of power over the whole signal, rather the average power in each frequency, and therefore does not contain as much information. The SPD, on the other hand, represents the stochastic distribution of the PSD in each frame over time, rather than simply the average power in each frequency.

4.3.1 SPD-IF Algorithm

Referring to path (2) in Figure 4.1a, the algorithm can be described as follows:

1. A spectrogram, $S(f, t)$, is generated from the sound signal.
2. A histogram is taken over time to produce the SPD representation, $H(f, b)$.
3. An image feature is extracted from the SPD image, following the same approach used for the SIF, as shown in Figure 4.1b.

Spectral Power Distribution Image

Starting from the spectrogram representation $S(f, t)$, calculated using Equation 4.1, the SPD representation, $H(f, b)$, is based on the histogram distribution of power for each frequency bin, f. This is calculated as follows:

$$H(f, b) = \sum_t \mathbf{1}_b(S(f, t)) \tag{4.5}$$

where $\mathbf{1}_b(S(f, t))$ is the indicator function and equals one for the b^{th} bin if the spectrogram power $S(f, t)$ lies within the range of the bin and is zero otherwise. Experimentally, it was determined that the bin size should be fixed at $\Delta = 1$ dB, using 105 bins, with the maximum bin position fixed at 125 dB.

Feature Extraction

To obtain the SPD image feature (SPD-IF), the same image feature extraction method is used, as described in Section 4.2.1 for the SIF. However, the magnitude of the SPD image, which represents the histogram bin count, is not scaled to between $[0, 1]$ as for the greyscale spectrogram in the SIF. Therefore, the quantisation regions, c, are instead fixed, as otherwise the noisy regions of the SPD could affect the scaling, as the histogram bin count in these regions is dependent on the length of the clip. Specifically, three quantisations are used with boundaries $c = \{0, 3, 6, \infty\}$, while the mean and variance are used as the two features, with $M = 16$ frequency segments, and $N = 15$ power segments. We found that these values provide a good balance between classification accuracy and feature length, which is therefore a total of $3 \times 16 \times 15 \times 2 = 1440$ dimensions, where 2 represents the mean/variance statistics extracted from each region.

4.3.2 Missing Feature Classification System

In this section, a missing feature classification system is developed based on the SPD image. Unlike the SIF, which used the full feature for classification with SVM, here only the feature components from the reliable areas of the SPD image are used. Hence, a noise estimation algorithm is used to identify the feature regions that are affected by noise, and this is combined with the k-Nearest Neighbours (kNN) classifier to marginalise these unreliable dimensions.

Reliable Feature Estimation

To generate a robust classification system, only the signal information from the SPD image should be characterised, while the noise areas rejected. Therefore, the areas affected by noise need to be estimated, and only the reliable features used for classification. Here, the first ten frames t_n are known to be silence, hence are used to estimate the noise level:

$$N(f) = \mu(S(f, t_n)) + 2\sigma(S(f, t_n)) + \Delta \tag{4.6}$$

where $\mu(.)$ is the mean, $\sigma(.)$ is the standard deviation, and Δ is a constant added to account for noise fluctuations. The noise level for each segment, $N(m)$, is estimated from the mean of $N(f)$ across the segment, and is shown by the white lines in Figure 4.3. Areas to the right of the line contain the signal information, hence are selected as the reliable feature dimensions as follows:

$$\mathbf{F}^r(m,n) = \begin{cases} \mathbf{F}(m,n), & \text{if } \underset{B(m,n)}{H(t,b)} > N(m). \\ [\,], & \text{otherwise.} \end{cases} \tag{4.7}$$

where m, n are the block indices, $N(m)$ is the noise estimate in frequency segment m, and $\mathbf{F}^r$ are the reliable dimensions.

Missing Feature Classifier

For classification, there are a variety of methods to handle missing features [86]. Here, the noisy feature dimensions are marginalised using a k-Nearest Neighbours (kNN) classifier, which is a simple, lazy classifier based on the Euclidean distance, deferring all training data until testing. Therefore, for each feature vector tested, $\mathbf{F}$, only the n^r reliable feature dimensions, $\mathbf{F}^r$, are compared with those in the training data, $\mathbf{F}_{ref}$, while the others are ignored:

$$\mathrm{d}(\mathbf{F}, \mathbf{F}_{ref}^{\{i\}}) = \frac{1}{n^r}\left[\sum_{k=1}^{n^r}(\mathbf{F}^{r(k)} - \mathbf{F}_{ref}^{\{i\}r(k)})^2\right]^{\frac{1}{2}} \quad \forall i \tag{4.8}$$

The class decision uses the maximum voting of the $k = 5$ lowest distances. The classification is fair, since the same number of dimensions are compared with each class, and the decision is made over the whole clip. Also, since only one feature per clip is calculated, the kNN classifier is fast compared to using frame-based features.

4.3.3 Discussion

From the examples in Figure 4.3, it can be seen that the dark areas are where the power histogram bin count is zero, while the lighter colours represent the bins with the highest count. The white line in the figures is an estimate of the noise, with the low magnitude areas to the left of this considered to contain no signal information.

The signal information is contained in the SPD areas to the right of the white dotted line, and has a characteristic distribution that represents the underlying sound. It is clearly visible for stationary sounds such as Whistle, while it is less obvious for others, such as Horn. The profile of the noise can be seen clearly in the 0dB examples, which is expected since the noise is diffuse and therefore has a Gaussian-like distribution with a small variance.

It is assumed that as the noise power increases, the signal information to the right of the white dotted line remains unchanged, and therefore the missing feature classifier is used to compare only these areas with the training data. The feature is therefore invariant to time-shifting, as extending the length of the clip only changes the histogram count in the noisy areas, as long as the signal is still contained within the clip.

4.4 Experiments

Experiments are conducted to compare the performance of both the SIF and SPD-IF methods with common baseline methods for sound event recognition. The methods are tested on a standard environmental sounds database, which has a variety of noise conditions added to it, to give conditions ranging from clean down to severe 0dB noise. For most experiments, training is carried out only on clean samples, while testing is carried out across all the noise conditions. An additional experiment is carried out for the baseline using multi-conditional training, which is a common, albeit data heavy, method for improving robustness to noise.

4.4.1 Experimental Setup

Database

A total of 50 sounds are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [46], giving a selection of collision, action and characteristics sounds. The isolated sound event samples have a sparse frequency spectrum, with a high signal-to-noise ratio (SNR), and are balanced to give some silence either side of the sound. The frequency spectrum of most sounds is quite sparse, with most of the power contained in only a few frequency bands. For each event, 50 files are randomly selected for training and another 30 for testing. The total number of samples are therefore 2500 and 1500 respectively, with each experiment repeated in 5 runs.

Noise Conditions

The following diffuse noise environments are added at 20, 10 and 0 dB SNR: “Speech Babble”, “Destroyer Control Room”, “Factory Floor 1” and “Jet Cockpit 1”, obtained from the NOISEX92 database. All four noises are diffuse, and have most of the energy concentrated in the lower frequencies. The exception is the Jet noise, which is diffuse but contains more high frequency components.

For each of the methods, the classification accuracy is investigated in mismatched conditions, using only clean samples for training. The average performance is reported at each SNR across all four noise environments.

Evaluation Methods

The following variations of the proposed methods are evaluated to determine the effect of each on the performance:

1. Spectrogram Image Feature (SIF)
    (a) Grey-scale spectrogram, without dynamic range quantisation/mapping
    (b) Raw and Log-power pseudo-colour quantised images
    (c) Log-power kNN-SIF, using the same missing feature (MF) approach as in Section 4.3.2 to remove noise-affected feature dimensions.
2. Spectral Power Distribution Image Feature (SPD-IF) based on:
    (a) Raw-Power and Log-Power spectrograms

Baseline Methods

For comparison with the proposed methods, several baseline methods are used, both with and without noise reduction algorithms. The following methods are implemented:

1. MFCC-HMM with 5 states and 6 Gaussians, trained with HTK [87]. A 36-dimension MFCC feature is generated, including deltas and accelerations, using the following:
    (a) HTK-HCopy with 24 filters, 12 coefficients.
    (b) Spectral Subtraction processing [75], then HCopy.
    (c) ETSI Advanced Front End (AFE) [76].
    (d) Multi-conditional training using HTK-HCopy.
2. MFCC-SVM: feature from the mean and variance of MFCCs over the clip

Method	Scaling	Details	Clean	20dB	10dB	0dB	Average
SIF	Log	Greyscale	97.3 ± 0.2	81.1 ± 5.5	53.5 ± 10.2	26.4 ± 8.8	64.6
	Log	Colour Quantised	97.3 ± 0.2	81.1 ± 5.5	53.5 ± 10.2	26.4 ± 8.8	64.6
	Raw		91.1 ± 1.0	91.1 ± 0.9	90.7 ± 1.0	80.9 ± 1.8	88.5
	Log	Missing Feature kNN	95.6 ± 0.5	90.8 ± 2.2	79.0 ± 8.4	69.1 ± 6.8	83.6
SPD-IF	Log	Missing Feature kNN	97.3 ± 0.2	**96.3 ± 0.5**	**94.1 ± 1.5**	**87.5 ± 5.2**	**93.8**
	Raw		83.2 ± 0.3	83.2 ± 0.5	82.7 ± 0.6	75.3 ± 2.5	81.1
MFCC HMM	HCopy		**99.5 ± 0.1**	78.9 ± 6.6	44.3 ± 6.9	13.6 ± 4.1	59.1
	Spec-Sub		99.2 ± 0.1	89.3 ± 4.4	68.5 ± 8.3	33.1 ± 7.6	72.5
	ETSI-AFE		99.1 ± 0.2	89.4 ± 3.2	71.7 ± 6.1	35.4 ± 7.7	73.9
	Multi-Conditional		98.9 ± 0.1	96.5 ± 1.8	88.4 ± 6.0	56.6 ± 12.8	85.1
MFCC SVM	ETSI-AFE		98.7 ± 0.1	55.3 ± 5.0	33.7 ± 3.7	13.2 ± 4.1	50.2

Table 4.1: Classification accuracy results from proposed methods vs. baseline experiments.

4.4.2 Results and Discussion

Baseline results

First, comparing the baseline results in Table 4.1, it can be seen that MFCC-SVM, which uses averaged features, performs worse than MFCC-HMM, with the same ETSI-AFE features. This is consistent since all temporal information that is lost in the frame-averaged MFCC-SVM features is fully captured in each HMM. Among the MFCC-HMM methods, the best performance comes from pre-processing carried out with the ETSI Advanced Front End, with an average accuracy of 73.9%. This is expected since the techniques used in the AFE, such as double Wiener filtering, are much more sophisticated than simple spectral subtraction, or the standard HTK-HCopy algorithm.

One of most popular methods when dealing with noise is multi-conditional training, hence another MFCC-HMM is trained with Clean, 20dB and 10dB samples using the "Destroyer" and"Jet" noises. Testing is on samples with "Babble" and "Factory" noise at 20dB, 10dB and also 0dB. As expected the results show that MFCC-HMM has greatly improved compared to mismatched conditions, although requires much more training data than for the proposed image-based methods. Hence, from here onwards, all comparisons are made with the best performing ETSI-AFE and multi-conditional MFCC-HMM baselines.

Colour Quantised vs. Grey-scale SIF

Here, the results obtained for the grey-scale and pseudo-colour quantised SIFs are compared and the effect of the quantisation is analysed. It can be seen in Table 4.1 that the quantised SIFs outperform the grey-scale SIF in both clean and mismatched conditions. This indicates that by mapping the grey-scale spectrogram into a higher dimensional space, in this case the three RGB quantisations, that the separability between sound classes has increased.

For the case of mismatched noise, the robustness of the proposed feature can be explained by fact that the noise is normally more diffuse than the sound and therefore the noise intensity is located in low-range regions of the spectrogram image. Hence, the monochrome images mapped from the higher-ranges should be mostly unchanged, despite the presence of the noise. Also, since the proposed feature is combined with the SVM classifier, where the discriminative components should be assigned a higher weighting, it should be more robust in mismatched noise conditions.

The effect of the quantisation can be shown experimentally. Since the SIF is based on the intensity distribution of the monochrome images, the distribution distance between clean and noisy samples of the same sound event should be compared. A robust feature should have a small distance, indicating that the distributions are similar. Modelling the distributions as Gaussian, the Square Hellinger Distance [88] can be used as a measure:

$$H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\, e^{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}}. \tag{4.9}$$

Since central moments are used as the SIF feature, the distributions are mean-normalised. Hence, with $\mu_1 = \mu_2 = 0$, Eq. 4.9 simplifies to a ratio of the standard deviations.

Example results are presented in Table 4.2, which show the mean distribution distances across the 9×9 SIF blocks, averaged over 50 samples using the linear-power SIF. Although

Sound Event	Green (Low)	Blue (Medium)	Red (High/Lowest)	Grey-Scale
Bottle1	0.412	0.002	0.062	0.642
Cymbals	0.377	0.041	0.095	0.343
Horn	0.350	0.069	0.069	0.306

Table 4.2: Example distribution distances for grey-scale and the three colour quantisations between clean and 0db noise conditions

the distribution distance of the green colour, representing the low region of the intensities, is relatively large, the distributions of the other colours are less affected by the noise. We suggest that this, combined with the SVM weighting, allows the dynamic range regions that are most susceptible to noise to be ignored.

Linear vs. Log Power

From the SIF experimental results in Table 4.1, it can be seen that one important experimental outcome is the improved performance of the raw-power SIF, over the log-power equivalent, since a linear representation is not commonly used in conventional methods. It can be seen that although log-power methods give a higher classification accuracy for clean conditions, the linear-power methods that show considerable robustness in mismatched conditions. The log-power representation is expected to perform better than the linear for clean conditions since the dynamic range is reduced, revealing the detail from the low power frequencies, which provides better discrimination between sound events. The linear spectrogram representation on the other hand, is sparse, with less information visible, as the strongest frequencies are an order of magnitude larger than the surrounding ones. This leads to confusion between the most similar sounds, which is reflected in the lower accuracy in clean conditions. However, in mismatched conditions, the sparse nature of the signal in the linear power representation remains visible, since the noise is diffuse and is spread over a wider range of frequencies, compared to the strong peaks of the sounds. In the case of the log-power representation, the detail of the noise is exposed and causes large changes in the colour of the spectrogram images.

For the SPD-IF, it is the log-power spectrogram that performs best, although the raw-power representation remained robust to the noise. This results is due to the noise estimation and missing feature algorithm, which ensures that the high accuracy of the log-power representation in clean conditions is preserved in the noisy conditions, since only the unaffected part of the signal is used for classification. The raw-power signal representation, on the other hand, has a very large dynamic range, where the distribution is sparse and cannot be captured as well using the image feature method, which explains the lower classification accuracies.

Comparison of proposed methods with the baseline

The results show that with an average classification accuracy of 88.5%, the best-performing linear-power SIF has a significant improvement over the ETSI-AFE MFCC-HMM baseline method, particularly in more severe mismatched conditions. In clean conditions, the performance of the SIF is just 91.1%, compared to 99.1% for the baseline, which is due to the use of a linear-power representation, as discussed above. The log-power SIF shows a higher accuracy in clean conditions, although is less effective in mismatched conditions, with a worse performance in some of the more severe noise conditions.

The best-performing SPD-IF, using a log-power dynamic range, performs even better, with a comparable result in clean conditions to the MFCC and log-power SIF baselines. However, the advantages of the SPD-IF are demonstrated in mismatched noise conditions, where even at 0dB, it achieves an accuracy of 87.5%, compared to 80.9% for the raw-power SIF and just 35.4% for the best performing MFCC-HMM method. It should be also noted that running the SPD-IF experiment using only 20 files from each sound class for training, and the remaining 60 for testing, the overall accuracy drops by less than 5%, and reduces the computation time.

The SPD-IF method even outperforms the MFCC-HMM baseline using multi-conditional training, with an average improvement in classification accuracy of over 8%

SIF vs. SPD-IF

Although based on similar concepts, the notable difference is that the SIF is based upon the spectrogram representation of the sound, as opposed to the subband power distribution used in the SPD-IF. The results show that the log-power SPD-IF performs well compared to the SIF, combining the high accuracy of the log-power SIF in clean conditions, with the robustness of the raw-power SIF in mismatched noise.

For the SIF, the raw-power representation performs better on average than the log-power representation, with an accuracy of 88.5%. This is still significantly lower than the 93.8% achieved by the SPD-IF.

In addition, using the missing feature kNN method with the log-power SIF improves the noise robustness of the method significantly. However, the feature is still affected by the problem of time-shifting, hence cannot achieve the performance shown by the log-power SPD-IF.

One factor that is not captured in the results is the importance of the time-invariance, as the SIF does not perform well in conditions where the clips are not as well balanced as in the RWCP database. In practise, the SPD-IF should be much more robust than the SIF.

4.5 Conclusion

In this chapter, two novel methods for sound event classification are proposed, motivated by the visual perception of the spectrogram image. Firstly, the Spectrogram Image Feature (SIF) is based on a feature extraction approach, inspired by image processing, that performs well for a linear-power representation that does not compress the dynamic range. The important point of the method is that since the noise is diffuse, it affects only limited part of the dynamic range, hence the quantisation of spectrogram maps the noise intensity to certain regions of the higher dimensional space, leaving the signal intensity unchanged in the other regions. Therefore when combining with a discriminative classifier, like SVM, it yields a very robust classification accuracy in mismatched conditions.

The other method is the Subband Power Distribution Image Feature (SPD-IF), which is based on the stochastic distribution of power over the whole sound clip, captured through the SPD image, which is invariant to the effects of time shifting. The image-like feature is extracted in the same way as the SIF, but in this case the feature can easily be combined with a noise estimation and missing feature classifier, to produce a robust classification system. Experiments show that it performs well in both clean and noisy conditions, considerably outperforming the best MFCC-HMM baseline using the ETSI-AFE noise reduction algorithm, and achieved a very high classification accuracy of 87.5% even in severe, 0dB mismatched noise.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The task of an acoustic event recognition system is to detect and label an unknown acoustic source occurring in an environment where noise and other effects may be present. In this report, the background information on these systems has been presented, and the important characteristics of generic acoustic events has been discussed. The most popular state-of-the-art systems are based on automatic speech recognition (ASR) systems, which typically perform well in clean, matched conditions, but their performance degrades rapidly when used in mismatched environments. Two feature extraction methods were proposed to overcome this performance degradation in mismatched noisy conditions. Both approaches consider the importance of the time-frequency representation, and use feature extraction methods inspired by similar techniques in image processing.

- The Spectrogram Image Feature (SIF) method directly extracts a feature from the sound spectrogram, by first quantising and mapping the dynamic range of the spectrogram image to a higher dimension. Experiments show that the proposed feature performs robustly in mismatched conditions, but the tradeoff is a reduced accuracy in clean conditions.

- The Spectral Power Distribution Image Feature (SPD-IF) is based on the same image feature approach as the SIF. However, the method uses a new SPD image representation, which is the stochastic distribution of power over the length of the sound clip. Experiments show that the method combines the high accuracy of the baseline experiments in clean conditions, with the robustness found in the SIF approach.

5.2 Future Work

While the proposed methods in this report have been shown to perform well in mismatched noisy conditions, there are still many aspects that still require research. The following list proposes the avenues for future research:

(i) ***Use of auditory processing:*** To achieve human-like performance in the machine hearing task, it should be beneficial to develop methods that take inspiration from the signal processing mechanisms in the human auditory system. By incorporating knowledge from the existing models of the inner ear from Lyon and Meddis [64, 62, 1], it is suggested that this can improve upon the currently proposed techniques. For instance, it is known that in the auditory nerve, there are approximately ten nerve fibres with the same best-frequency, each of which has a different spontaneous emission rate and saturation level. This is similar to thresholding into different grey-levels in image processing, hence it would be interesting to explore how the distribution of such sub-images can be captured to address recognition in noisy environments.

(ii) ***Missing feature approaches:*** For the SPD-IF method, a simple missing feature approach was used to identify and marginalise the noise affected regions of the image. Only performing classification on the reliable feature dimensions was an important element that contributed towards the robustness of the proposed method. Therefore, further research will be carried out to investigate other existing missing feature approaches, with the aim to combine these with future systems to improve classification performance.

(iii) ***Environment impulse response:*** This is an important factor which incorporates the environmental reverberation and microphone effects that are present in every real-world recognition system. Currently, performance often degrades when the system is used in a different location or with a different microphone from the one used in training. These problems are fundamental for the wider uptake of such systems, hence will be investigated by first assuming that the distortion of the high power, and therefore most important, frequency components is limited only to a change in their magnitude. In addition, it may be possible to apply image processing techniques, such as those used for be-blurring motion images, to reduce the spectral distortion prior to classification.

(iv) ***Multiple sources:*** Most existing recognition systems are not designed to handle the case where two sound events occur simultaneously, and will often produce an erroneous output. Although a few solutions have been proposed, none are currently able to address this problem in real environmental conditions. The field of computational auditory scene analysis considers this as one of the fundamental problems to be overcome, and hence further research will be carried to utilise the onset grouping constraints that define how the humans separate audio mixtures into streams. Another area is to develop a classification method that can match partial areas of the observed spectrogram with models that are trained such that only they produce a stronger match at their most important frequencies.

Publications

(i) J. Dennis, H.D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011

(ii) J. Dennis, H.D. Tran, and H. Li, "Image Representation of the Subband Power Distribution for Robust Sound Classification," in *Twelfth Annual Conference of the International Speech Communication Association*, August 2011, Accepted

References

[1] R.F. Lyon, "Machine Hearing: An Emerging Field," *IEEE Signal Processing Magazine*, 2010.

[2] M. Cowling, R. Sitte, and T. Wysocki, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," *Digital Signal Processing for Communication Systems, Sydney-Manly*, 2002.

[3] J. Dennis, H.D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011.

[4] J. Dennis, H.D. Tran, and H. Li, "Image Representation of the Subband Power Distribution for Robust Sound Classification," in *Twelfth Annual Conference of the International Speech Communication Association*, August 2011, Accepted.

[5] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor network for the monitoring of ecosystem: Bird species recognition," in *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*. IEEE, 2008, pp. 293–298.

[6] I. Boesnach, M. Hahn, J. Moldenhauer, T. Beth, and U. Spetzger, "Analysis of Drill Sound in Spine Surgery," in *Perspective in image-guided surgery: proceedings of the Scientific Workshop on Medical Robotics, Navigation, and Visualization: RheinAhrCampus Remagen, Germany, 11-12 March*. World Scientific Pub Co Inc, 2004, p. 77.

[7] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," *Multimodal Technologies for Perception of Humans*, pp. 311–322, 2007.

[8] A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, The MIT Press, 1994.

[9] N. Yamakawa, T. Kitahara, T. Takahashi, K. Komatani, T. Ogata, and H.G. Okuno, "Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition," 2010.

[10] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[11] L. Lu, H.J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 504–516, 2002.

[12] D. Gerhard and University of Regina. Dept. of Computer Science, *Audio signal classification: History and current techniques*, Citeseer, 2003.

[13] M. Cowling, "Non-speech environmental sound classification system for autonomous surveillance," 2004.

[14] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[15] D.A. Reynolds, "An overview of automatic speaker recognition technology," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings.(ICASSP'02). IEEE International Conference on*. IEEE, 2002, vol. 4.

[16] D.A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of ICASSP*, 2005, pp. 953–956.

[17] Y.K. Muthusamy, E. Barnard, and R.A. Cole, "Automatic language identification: A review/tutorial," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, 1994.

[18] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection in noisy environments," in *15th European Signal Processing Conference (EUSIPCO-07), Sep. 3-7, Poznan, Poland*, 2007.

[19] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.* IEEE, 2005, pp. 1306–1309.

[20] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[21] I. Paraskevas, SM Potirakis, and M. Rangoussi, "Natural soundscapes and identification of environmental sounds: A pattern recognition approach," in *Digital Signal Processing, 2009 16th International Conference on.* IEEE, 2009, pp. 1–6.

[22] 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.* IEEE, 2009, pp. 1–4.

[23] A. Waibel, R. Stiefelhagen, R. Carlson, J. Casas, J. Kleindienst, L. Lamel, O. Lanz, D. Mostefa, M. Omologo, F. Pianesi, et al., "Computers in the human interaction loop," *Handbook of Ambient Intelligence and Smart Environments*, pp. 1071–1116, 2010.

[24] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 321–329, 2006.

[25] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, no. 2, pp. 1–22, 2006.

[26] S. Chu, S. Narayanan, C.C.J. Kuo, and M.J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Multimedia and Expo, 2006 IEEE International Conference on.* IEEE, 2006, pp. 885–888.

[27] L.G. Martins, J.J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2007.

[28] J.J. Burred, A. Robel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *Acoustics, Speech and Signal*

Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009, pp. 173–176.

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.

[30] A. Meng, P. Ahrendt, J. Larsen, and L.K. Hansen, "Temporal feature integration for music genre classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1654–1664, 2007.

[31] R. Tao, Z. Li, Y. Ji, and EM Bakker, "Music Genre Classification Using Temporal Information and Support Vector Machine," ASCI Conference, 2010.

[32] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, indexing, and retrieval for environmental and natural sounds," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 688–707, 2010.

[33] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *Multimedia, IEEE*, vol. 3, no. 3, pp. 27–36, 2002.

[34] B. Mechtley, G. Wichern, H. Thornburg, and A. Spanias, "Combining semantic, social, and acoustic similarity for retrieval of environmental sounds," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.* IEEE, 2010, pp. 2402–2405.

[35] G. Peeters and E. Deruty, "Sound indexing using morphological description," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 675–687, 2010.

[36] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 467–476, 2008.

[37] J. Ramírez, JM Górriz, and JC Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *M. Grimm, and K. Kroschel, Robust Speech Recognition and Understanding*, pp. 1–22, 2010.

[38] A. Temko, "Acoustic Event Detection and Classification," 2008.

[39] D. Hoiem, Y. Ke, and R. Sukthankar, “SOLAR: sound object localization and retrieval in complex audio environments,” in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on.* IEEE, 2005, vol. 5.

[40] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” *CUIDADO IST Project Report*, pp. 1–25, 2004.

[41] F. Pachet and P. Roy, “Exploring billions of audio features,” in *Content-Based Multimedia Indexing, 2007. CBMI'07. International Workshop on.* IEEE, 2007, pp. 227–235.

[42] J.W. Picone, “Signal modeling techniques in speech recognition,” *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[43] K.P. Murphy, *Dynamic bayesian networks: representation, inference and learning*, Ph.D. thesis, Citeseer, 2002.

[44] C.J.C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[45] O.L. Mangasarian and E.W. Wild, “Proximal support vector machine classifiers,” in *Proceedings KDD-2001: Knowledge Discovery and Data Mining.* Citeseer, 2001.

[46] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” in *Proceedings of International Conference on Language Resources and Evaluation*, 2000, vol. 2, pp. 965–968.

[47] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo, “The CLEAR 2007 evaluation,” *Multimodal Technologies for Perception of Humans*, pp. 3–34, 2009.

[48] R. Malkin, D. Macho, A. Temko, and C. Nadeu, “First evaluation of acoustic event classification systems in CHIL project,” in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Array.* Citeseer, 2005.

[49] G. Brian and S. Valeriy, “Development of the Database for Environmental Sound Research and Application (DESRA): Design, Functionality, and Retrieval Considerations,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[50] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[51] "Acoustic computing for ambient intelligent applications," March 2011, http://acaia.org/.

[52] "Speech-acoustic scene analysis and interpretation," March 2011, http://shine.fbk.eu/en/home.

[53] "Self configuring environment aware intelligent acoustic sensing," March 2011, http://www-dsp.elet.polimi.it/ispg/SCENIC/.

[54] J.J. Xiang, M.F. McKinney, K. Fitz, and T. Zhang, "Evaluation of sound classification algorithms for hearing aid applications," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 185–188.

[55] R.O. Duda, P.E. Hart, and California 94025 Stanford Research Institute. Menlo Park, "Experiments in scene analysis," 1970.

[56] JO Pickles, *An introduction to the physiology of hearing*, Academic Press, 2008.

[57] RD Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *APU report*, vol. 2341, 1988.

[58] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[59] E. Martinez, K. Adiloglu, R. Annies, H. Purwins, and K. Obermayer, "Classification of everyday sounds using perceptual representation," in *Proceedings of the Conference on Interaction with Sound*. Fraunhofer Institute for Digital Media Techology IDMT, 2007, vol. 2, pp. 90–95.

[60] R. Anniés, E.M. Hernandez, K. Adiloglu, H. Purwins, and K. Obermayer, "Classification schemes for step sounds based on gammatone-filters," in *Neural Information Processing Systems Conference (NIPS)*, 2007.

[61] Y.R. Leng, H.D. Tran, N. Kitaoka, and H. Li, "Selective Gammatone Filterbank Feature for Robust Sound Event Recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[62] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor.," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702, 1986.

[63] S. Srinivasan and D.L. Wang, "A model for multitalker speech perception," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3213, 2008.

[64] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.* IEEE, 1982, vol. 7, pp. 1282–1285.

[65] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Citeseer, 1996.

[66] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on maxvq and casa for robust speech recognition," *Computer Speech & Language*, vol. 24, no. 1, pp. 30–44, 2010.

[67] M. Slaney and R.F. Lyon, "On the importance of time-a temporal representation of sound," *Visual representations of speech signals*, pp. 95–116, 1993.

[68] 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*, vol. 17, no. 6, pp. 1142–1158, 2009.

[69] G. Yu and J.J. Slotine, "Audio classification from time-frequency texture," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 1677–1680.

[70] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[71] D. Grangier, F. Monay, and S. Bengio, "A discriminative approach for the retrieval of images from text queries," *Machine Learning: ECML 2006*, pp. 162–173, 2006.

[72] J.W. Lewis, F.L. Wightman, J.A. Brefczynski, R.E. Phinney, J.R. Binder, and E.A. DeYoe, "Human brain regions involved in recognizing environmental sounds," *Cerebral Cortex*, vol. 14, no. 9, pp. 1008, 2004.

[73] J.J. Hopfield and C.D. Brody, “What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 1282, 2001.

[74] BH Juang, “Speech recognition in adverse environments,” *Computer speech & language*, vol. 5, no. 3, pp. 275–294, 1991.

[75] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[76] A. Sorin and T. Ramabadran, “Extended advanced front end algorithm description, Version 1.1,” *ETSI STQ Aurora DSR Working Group, Tech. Rep. ES*, vol. 202, pp. 212, 2003.

[77] X. Xiao, E.S. Chng, and H. Li, “Normalization of the speech modulation spectra for robust speech recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1662–1674, 2008.

[78] M.J.F. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1995.

[79] P.A. Naylor, *Speech dereverberation*, Springer Verlag, 2010.

[80] B.W. Gillespie, H.S. Malvar, and D.A.F. Florêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on.* IEEE, 2001, vol. 6, pp. 3701–3704.

[81] Z. Ghahramani and M.I. Jordan, “Factorial hidden markov models,” *Machine learning*, vol. 29, no. 2, pp. 245–273, 1997.

[82] V. Zue, “Notes on spectrogram reading,” *Mass. Inst. Tech. Course*, vol. 6, 1985.

[83] J.L. Shih and L.H. Chen, “Colour image retrieval based on primitives of colour moments,” in *Vision, Image and Signal Processing, IEE Proceedings-*. IET, 2002, vol. 149, pp. 370–376.

[84] O.L. Mangasarian and E.W. Wild, “Proximal support vector machine classifiers,” in *Proceedings KDD-2001: Knowledge Discovery and Data Mining.* Citeseer, 2001.

[85] W.C. Chu, *Speech coding algorithms*, Wiley Online Library, 2003.

[86] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, 2005.

[87] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*.

[88] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal processing*, vol. 18, no. 4, pp. 349–369, 1989.