NANYANG
TECHNOLOGICAL
UNIVERSITY

SOUND EVENT RECOGNITION IN UNSTRUCTURED ENVIRONMENTS USING SPECTROGRAM IMAGE PROCESSING

JONATHAN WILLIAM DENNIS

SCHOOL OF COMPUTER ENGINEERING

2014

SOUND EVENT RECOGNITION IN UNSTRUCTURED ENVIRONMENTS USING SPECTROGRAM IMAGE PROCESSING

JONATHAN WILLIAM DENNIS

School of Computer Engineering

A thesis submitted to the School of Computer Engineering
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

2014

Acknowledgements

I would like to sincerely thank my supervisors, Dr. Tran Huy Dat (Institute for Infocomm Research, I^2R) and Dr. Chng Eng Siong (Nanyang Technological University, NTU), firstly for introducing me to the field of audio processing, and secondly 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^2R, for their support in a variety of ways. In particular, to Dr. Li Haizhou, as through his knowledge and insight in the field, I have been inspired to improve and better my own research. Also to the members of the speech processing group from NTU, as their discussions across a range of research topics has ultimately given me guidance in my own research.

Finally, this work could not have been achieved without the love and support of my family back home, my wife who helped me through in many ways, and our son who brightens up our lives.

Contents

Abstract

The objective of this research is to develop feature extraction and classification techniques for the task of sound event recognition (SER) in unstructured environments. Although this field is traditionally overshadowed by the popular field of automatic speech recognition (ASR), an SER system that can achieve human-like sound recognition performance opens up a range of novel application areas. These include acoustic surveillance, bio-acoustical monitoring, environmental context detection, healthcare applications and more generally the rich transcription of acoustic environments. The challenge in such environments are the adverse effects such as noise, distortion and multiple sources, which are more likely to occur with distant microphones compared to the close-talking microphones that are more common in ASR. In addition, 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. Therefore, the performance of ASR systems typically degrades dramatically in these challenging unstructured environments, and it is important to develop new methods that can perform well for this challenging task.

In this thesis, the approach taken is to interpret the sound event as a two-dimensional spectrogram image, with the two axes as the time and frequency dimensions. This enables novel methods for SER to be developed based on spectrogram image processing, which are inspired by techniques from the field of image processing. The motivation for such an approach is based on finding an automatic approach to "spectrogram reading", where it is possible for humans to visually recognise the different sound event signatures in the spectrogram. The advantages of such an approach are twofold. Firstly, the sound event image representation makes it possible to naturally capture the sound information in a two-dimensional feature. This has advantages over conventional one-dimensional frame-based features, which capture only a slice of spectral information

within a short time window. Secondly, the problem of detecting sound events in mixtures containing noise or overlapping sounds can be formulated in a way that is similar to image classification and object detection in the field of image processing. This makes it possible to draw on previous works in the field, taking into account the fundamental differences between spectrograms and conventional images.

With this new perspective, three novel solutions to the challenging task of robust SER are developed in this thesis. In the first study, a method for robust sound classification is developed called the Spectrogram Image Feature (SIF), which is based on a global image feature extracted directly from the time-frequency spectrogram of the sound. This in turn leads to the development of a novel sound event image representation called the Subband Power Distribution (SPD) image. This is derived as an image representation of the stochastic distribution of spectral power over the sound clip, and can overcome some of the issues of extracting image features directly from the spectrogram. In the final study, the challenging task of simultaneous recognition of overlapping sounds in noisy environments is considered. An approach is proposed based on inspiration from object recognition in image processing, where the task of finding an object in a cluttered scene has many parallels with detecting a sound event overlapped with other sources and noise. The proposed framework combines keypoint detection and local spectrogram feature extraction, with a model that captures the geometrical distribution of the keypoints over time, frequency and spectral power. For each of the proposed systems detailed experimental evaluation is carried out to compare the performance against a range of state-of-the-art systems.

List of Publications

(i) J. Dennis, H. D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *IEEE Signal Processing Letters*, vol. 18, pp. 130–133, Feb. 2011.

(ii) J. Dennis, H. D. Tran, and H. Li, "Image Representation of the Subband Power Distribution for Robust Sound Classification," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2437–2440, Aug. 2011.

(iii) J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping Sound Event Recognition using Local Spectrogram Features with the Generalised Hough Transform," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Sept. 2012.

(iv) J. Dennis, H. D. Tran, and E. S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 367–377, Feb. 2013.

(v) J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, "Temporal Coding of Local Spectrogram Features for Robust Sound Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

(vi) J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, pp. 1085–1093, July 2013.

List of Figures

List of Tables

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. While speech is arguably the most informative sound event, this research focuses on the recognition of more general sound events, such as doors closing or bells ringing, which provide information and context for the environment beyond that contained in the speech. This field of research is called sound event recognition (SER), and can open up a range of novel application areas such as acoustic surveillance, bio-acoustical monitoring, environmental context detection, healthcare applications and more generally the rich transcription of acoustic environments. The objective of this thesis is to develop novel feature extraction and classification techniques, which address the significant challenges associated with the unstructured environments that complicate the sound event recognition task.

1.1 Motivation

The goal of computational sound event 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."

The field described above has not received as much attention as ASR, despite the range of novel applications that can be derived. Although they are based on similar signal processing concepts, there are a number of aspects that set research on SER apart from the traditional topic of ASR. Firstly, the characteristics of sound events differ from those of speech, with a much wider variety in frequency content, duration and profile compared to speech alone. Secondly, no sub-word dictionary exists for sounds in the same way as for speech, where it is possible to decompose words into their constituent phonemes. Finally, noise, distortion and overlapping sources are often present in the unstructured environments in which sound events occur. Therefore, it is important to develop new methods for SER that can perform well for this challenging task.

Despite the differences between sound events and speech, many previous works on SER have been based on existing ASR techniques. The most common approach is to use a system based on one-dimensional frame-level features, such as Mel Frequency Cepstral Coefficients (MFCCs), which are modelled using Gaussian Mixture Models (GMMs), with temporal sequencing captured by a Hidden Markov Model (HMM). However, the drawbacks of using this approach for SER are twofold. Firstly, there is a wide variation in the time-frequency structure of different sound events, and this information may not be best captured by an HMM which assumes independence between adjacent observations in time. Secondly, the frame-based features only capture the sound event information within a narrow time window, and commonly represent the full frequency spectrum. This causes problems in mismatched conditions, where the features may contain elements from noise or multiple sources and it is challenging to separate them. Therefore, while such systems typically have a high recognition accuracy in clean conditions, they perform poorly in the unstructured environments that are commonly found in SER applications.

Recently, there has been a trend towards developing techniques for SER that capture the two-dimensional time-frequency sound event information [2, 3]. This is well suited for sound events, since they have a more characteristic time-frequency signature compared to speech, and such techniques have shown an improved performance compared to using traditional ASR methods. Extending this idea, a related research field that similarly operates with two-dimensional data is image processing. Here, two of the basic challenges are the classification of visual scenes and the detection of objects in cluttered images. By representing the sound event through its time-frequency spectrogram image, the task of SER can be recast into one that shares a number of similarities with the challenges faced in image processing. This provides the motivation to develop novel methods for SER that are inspired by the related techniques in image processing, taking into account the fundamental differences between spectrograms and conventional images.

1.2 Contributions

In this thesis, three novel SER methods are proposed that can overcome the drawbacks of conventional audio processing systems when applied to the challenging unstructured environments found in SER. The inspiration for this comes from the idea of "spectrogram reading" [4], where it is possible for humans to visually recognise the elements in the spectrogram belonging to different sources, even in the presence of noise or multiple sources. This suggests that there is sufficient information in the sound event spectrogram to perform SER by finding an automatic approach to spectrogram reading. In this thesis, this is referred to as spectrogram image processing. With this new perspective, the following novel solutions are developed to address the challenging task of robust SER in unstructured environments:

Spectrogram Image Feature (SIF) [5] In the first study, a method is developed for robust sound classification, which is based on a global image feature extracted from the time-frequency spectrogram of the sound. Using a technique similar to pseudo-colourmapping in image processing, the dynamic range of the spectral power is quantised into regions, such that the characteristic, high-power spectral peaks can be extracted separately to produce a robust feature. The experi-

ments on a large database of environmental sounds show that this technique can outperform a range of well-performing baselines, including those trained using multi-conditional data.

Sub-band Power Distribution (SPD) image [6, 7] Here, a novel sound event image representation is developed that can overcome some of the issues of extracting image features directly from the spectrogram. The SPD is designed to integrate directly within the SIF framework described above, and improve the performance in both clean and mismatched conditions. To this end, the SPD is derived as an image representation of the stochastic distribution of spectral power over the sound clip. The advantage of this representation is that the boundary between signal and noise can be easily found, such that it simplifies the task of generating a missing feature mask for the SPD. When combined with a missing feature classification system based on the SIF framework, the experiments show that the method can achieve the high accuracy of the baseline methods in clean conditions, while obtaining significantly more robust results in mismatched noise conditions.

Local Spectrogram Features (LSFs) [8–10] In this final study, the challenging task of simultaneous recognition of overlapping sounds in noisy environments is considered. The proposed approach is based on the idea of object recognition from image processing, where the task of finding an object in a cluttered scene has many parallels with detecting a sound event overlapped with other sources and noise. The framework combines keypoint detection and local spectrogram feature (LSF) extraction, with a model that captures the geometrical distribution of the keypoints over time, frequency and spectral power. During recognition, the generalised Hough transform (GHT) is used as a voting mechanism to generate sound event hypotheses from the LSFs the spectrogram. The hypotheses are then sparse and separable in the Hough accumulator space. The final step is verification, where the combination of hypotheses that best explains the observed spectrogram is determined. Experiments show that this approach performs well across a range of mismatched conditions. The LSF approach also has the advantage over existing methods that it does not assume a fixed overlap between a fixed number of sounds, hence can detect a arbitrary combination of

overlapping sounds, including multiple instances of the same sound.

The work on the SIF is published in the journal: IEEE Signal Processing Letters [5], while the work on the SPD is published in the journal: IEEE Transactions on Audio, Speech and Language Processing journal [6], and in the conference: Interspeech 2011 [7]. The work on the LSF is published in the journal: Pattern Recognition Letters [8], and in two conferences: Interspeech 2012 [9] and ICASSP 2013 [10].

1.3 Organisation of this Thesis

This thesis is organised as follows:

In Chapter 2, a thorough overview of the field of sound event recognition is provided, followed by a review of the current state-of-the-art SER systems, and a discussion of the limiting factors that affect the performance of such systems.

Chapter 3 then introduces the spectrogram image processing approach for sound event recognition, and provides a background of the relevant state-of-the-art techniques in image classification and object detection. The spectrogram image feature (SIF) method is then proposed for robust sound classification, and detailed experiments are carried out to analyse the performance.

In Chapter 4, the SPD image is proposed to improve upon the SIF framework, as the SPD representation simplifies the process of separating the noise and signal compared to the spectrogram. A missing feature classification system is developed, and the experimental results demonstrate the strong performance of the method compared to the best-performing baseline techniques.

Chapter 5 then introduces the problem of simultaneous recognition of overlapping sounds, and reviews the current state-of-the-art techniques in this area. A discussion of the current limitations leads to the development of the LSF approach, which takes inspiration from object detection in image processing to develop a model of the sound event that can recognise an arbitrary combination of overlapping sounds.

Finally, the thesis concludes in Chapter 6 with a summary of the contributions and possible future research directions.

Chapter 2

Introduction to Sound Event Recognition

The field of sound event recognition (SER) is a novel application area that has developed from the more general fields of audio processing and pattern recognition. While a range of techniques have been developed to address the specific challenges of this domain, the field is generally less well understood than the popular field of automatic speech recognition (ASR). Therefore, Section 2.1 of this chapter aims to provide an introduction to the field of SER, by first describing what is meant by a "sound event", then introducing the wide range of applications that are possible, the challenges that are faced, and an overview of a typical SER system. Then, Section 2.2 reviews a range of state-of-the-art methods for SER, including those based on novel audio features and auditory modelling, followed by a discussion on their limitations. Finally, Section 2.3 carries out a set of experiments to establish a baseline comparison for the proposed techniques that are introduced later in this thesis.

2.1 Overview

Until recently, the field of non-speech SER has not received as much attention as ASR, as the desire to achieve robust computer-interpreted speech has outweighed the benefits of understanding the surrounding acoustic scene [1]. However, as the field of ASR has grown, advances in audio processing have expanded the possibilities of SER,

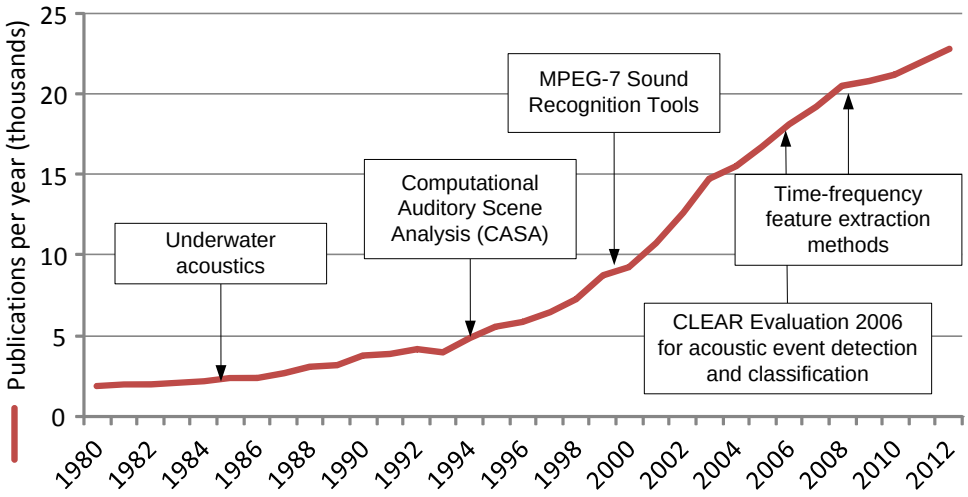


Figure 2.1: Number of publications per year on the topic of sound event recognition or classification -(brain,language) found in Google Scholar.

and opened up a wide range of applications that have begun to be realised. These include such diverse examples as identification of bird species [11], analysis of medical drill sounds [12] and detection of meeting room sounds [13], among many others. On one hand, the wide range of SER applications has introduced many opportunities for developing novel approaches to the challenges faced. However, it also introduces its own problem, in that the field of SER suffers from the difficulty in defining exactly where it lies within a larger scope. Therefore, relevant publications are spread across many different disciplines, and it is often difficult to compare different methodologies as the underlying datasets are so different. This is unlike the comparatively closed field of ASR, in which organisations have developed and provided standard corpuses, such as Aurora-2 [14] and Switchboard [15], to test speech recognition performance.

Historically, the early investigations into SER as a separate topic from ASR began in the 1980s, with the analysis of underwater acoustic patterns [16, 17]. At this point, the number of publications written on the topic of SER was relatively small, and since then it has been growing steadily, as shown in Fig. 2.1. After the initial investigations into underwater acoustics, further applications were quickly identified, such as for recognising heart sounds for cardiac diagnosis [18, 19]. However, the next step forward for the field came in the mid-90's in the form of Bregman's Auditory Scene Analysis (ASA) [20, 21]. This described the human perception of the auditory

scene without specific focus on speech, which provided motivation for segmenting and recognising sounds for novel purposes [22]. Later however, the field of Computational Auditory Scene Analysis (CASA) would shift its focus to the problem of robust speech recognition and separation [23], moving away from the generic framework described by Bregman [24]. The next major advancement for SER was in 2001 with the inclusion of sound recognition descriptors in the MPEG-7 toolkit [25]. This provided a unified framework for feature extraction and indexing of audio using trained sound classes in a pattern recognition framework [26]. Then, in 2006, the first challenge designed specifically for sound event recognition in meeting room environments was held [27], where a system based on perceptual features achieved the highest classification performance [13]. More recently, the field has been growing rapidly, and many novel state-of-the-art approaches have been proposed. In particular, there has been growing interest in the time-frequency analysis of the sound events [2], as traditional approaches in ASR are not well suited to capture the non-stationary nature and sharp discontinuities found in many sounds [3].

The research topic of SER draws inspiration from a number of underlying fields, including signal processing, machine learning, ASR, and CASA, as shown in Fig. 2.2. The field of CASA is in turn inspired by both biological evidence from the human auditory system, and from research on scene analysis from the field of image processing. In this thesis however, the direct link is made between image processing and SER by interpreting sound events as a two-dimensional spectrogram image representation and taking inspiration directly from methods designed for conventional images. This allows novel approaches to be developed that can capture the two-dimensional information in the sound, without being constrained by conventional methods and systems from ASR. This idea is further developed in Chapter 3 with the introduction of spectrogram image processing, while the rest of this section aims to first provide a detailed overview of the field of SER.

2.1.1 What is a Sound Event?

In this thesis, the focus is on "sound events" that have properties such as an onset, duration and offset time. They also have a characteristic frequency content that typically can be used to identify the source of the sound. Examples of such classes of sound

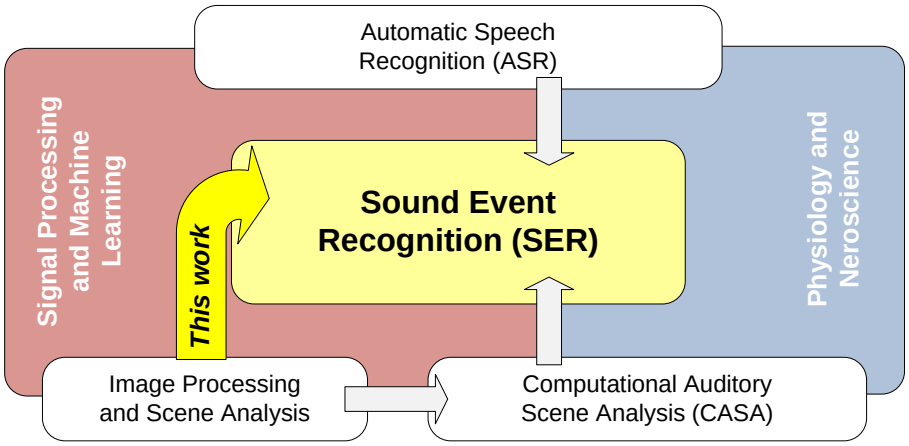


Figure 2.2: Overview of how sound event recognition (SER) draws inspiration from the wider field. In particular, while the field of CASA is inspired by ideas from image processing, in this thesis a direct link is made between the two by representing sound events as two-dimensional spectrogram images.

event can include a phone ringing, the clash of cymbals, footsteps or a bell ringing, among many others. The rich variety of these sound events that continually occur all around us carry important information and cues, and can be processed to provide knowledge about the environment. For example, valuable context can be gained from sound events occurring in meeting rooms, such as people laughing or entering the room.

Unlike speech or music, it is difficult to summarise the general characteristics of a sound event. This is firstly due to the wide variety of environments in which sounds occur, and secondly that sound events can be generated by many different types of interactions. Speech, on the other hand, is confined to the sounds that are produced by the human vocal tract and tongue. While this has many variations due to different speakers and their emotional state [29], it is still relatively well defined compared to general sound events. A comparison of the acoustical characteristics of speech, music and sound events is shown in Table 2.1. It can be seen that whereas speech and music have clear definitions for each acoustical characteristic, sound events generally are either undefined or can cover the full range of characteristics. For example, the bandwidth of speech is narrow, with much of the energy concentrated in the lower frequencies [30], while music is much broader, containing energy up to 20kHz and beyond. However, the bandwidth of sound events can vary considerably between the two, depending on the source and mode of excitation of the sound. Overall, this

Acoustical Characteristics	**Speech**	**Music**	**Sound Events**
No. of Classes	No. of Phonemes	No. of Tones	Undefined
Length of Window	Short (fixed)	Long (fixed)	Undefined
Bandwidth	Narrow	Broad	Broad Narrow
Harmonics	Clear	Clear	Clear Unclear
Repetitive Structure	Weak	Weak	Strong, Weak

Table 2.1: Comparison of the acoustical characteristics of speech, music and sound events, adapted from [28].

explains why it is common to define a narrow scope for the problem of SER, such as choosing a specific type of sounds, so that at least some of the characteristics can be defined.

To bring structure to the domain, it is necessary to develop a sound taxonomy, which separates sounds into several groups and sub-groups. This enables other researchers to understand the data domain. An example is shown in Fig. 2.3, where the class of hearable sounds is split into five categories [31]. Here, a non-speech sound event would fall under the natural or artificial sound classes, depending on the source of the sound. The classes are chosen to follow how humans would naturally classify sounds, and examples are given under each to describe the class. However, this human-like classification can lead to ambiguity and it is also notable that while both speech and music are well structured, natural and artificial sounds only act as general groupings, without much internal structure to the class. Other attempts have been made to develop sound event taxonomies that have similarities with the phoneme structure of speech [32], however there is still ambiguity that depends on individual human perception, hence this topic is open to further research.

One special group of sounds in Fig. 2.3 belong to the noise category. In this thesis, noise is considered differently from sound events, as it has unique properties such as a long duration and typically a slowly varying spectral content. A simple example of noise would be background office noise, created by computer fans and air-conditioning

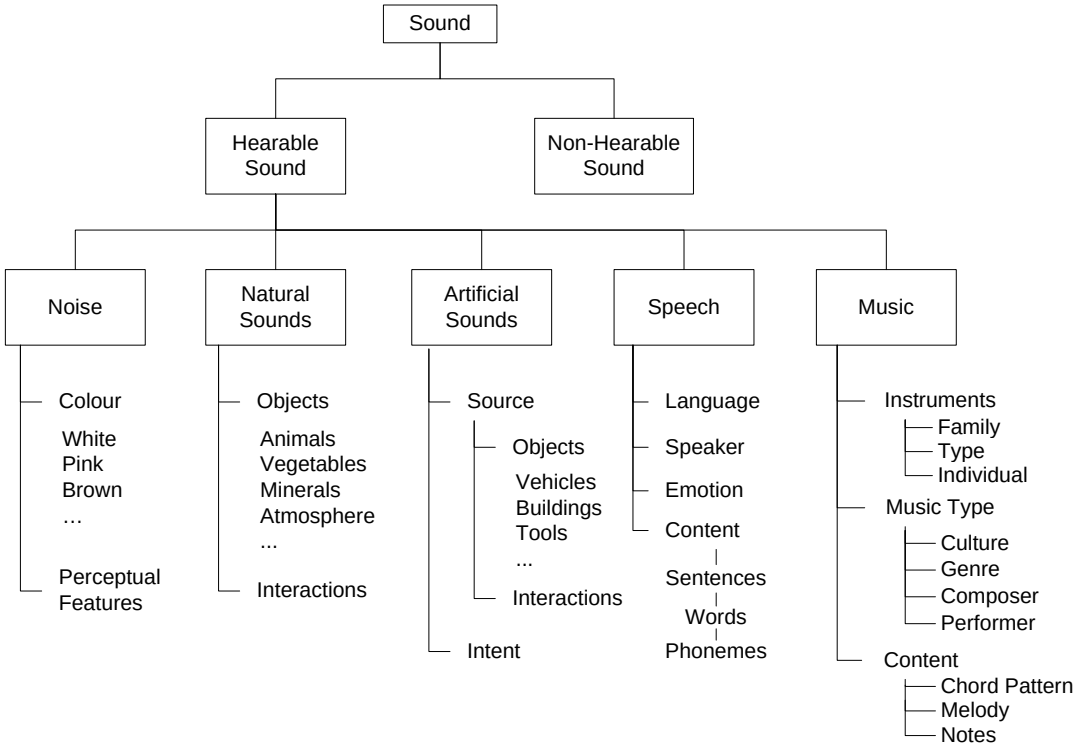


Figure 2.3: Taxonomy for Sound Classification, adapted from [31]

units. Another example could be rain, which consists of many impulsive rain drop sounds that combine together to produce a spectral content close to white noise. In general, noise is quite subjective and depends on our individual perception of what makes a sound undesirable to listen to [31]. Therefore, in this thesis, noise is simply considered to have a long duration with slowly-varying spectral profile, while any sound that has a well defined onset and offset is considered as a sound event that may carry useful information. This includes sounds such as keyboard clicking, footsteps or doors closing, which are often considered as impulsive noise in the field of ASR. However, here they are considered as meaningful sound events that can be used in a range of novel applications, which are now discussed in the next section.

2.1.2 Applications

The potential applications for sound event recognition (SER) are diverse, partly due to the very general interpretation of a sound event, as introduced above. Because of this,

there are few standardised datasets for comparison of different methodologies, and instead many authors focus on a specific topic of interest to a particular application. This is in contrast to the comparatively closed field of ASR, where there are international benchmarks for comparing system performance, such as the Aurora-2 [14] and Switchboard [15] corpuses. Therefore, it is necessary to compare and analyse the merits of each system for their given task. To provide an overview of the typical applications for SER systems, the approaches are grouped into three categories: environmental sound event recognition, acoustic surveillance, and environment classification. These are now introduced below.

Environmental Sound Event Recognition This is understood here as a chosen subset of sounds that might be found in a given environment. The environment of interest then determines the scope of the recognition problem. It is important to note that although speech may be included as a sound event category, the content of the speech would not be directly interpreted. Instead the detected speech content may be passed to an ASR system for recognition.

One application environment that attracts research is in the area of healthcare [12, 33, 34]. For example, the objective in [33] is to recognise and count coughing sound events, to enable the automatic assessment of cough frequency over a long period of time. To do this, an audio recording device is placed on the patient, and the detected sound events are classified simply as cough or non-cough. In a more recent work [34], the scope is expanded to include a wider set of healthcare audio events, including falling-down-stairs, screaming and collapsing.

Another area that has attracted research is focussed on the detection and classification of sounds in a meeting room environment [35–39]. It is also one of the only areas of SER research that has a standardised database for comparing the performance of competing systems. This database is from the Classification of Events, Activities and Relationships (CLEAR) 2006/7 workshop evaluations [13, 40], which is part of the Computers in the Human Interaction Loop (CHIL) project [41]. Here, the acoustic events of interest include steps, keyboard typing, applause, coughing, and laughter, among others, while speech during the meetings is ignored. It was found that achieving an accurate detection and segmentation of the sound events from the continuous audio was the most chal-

lenging task [40]. In particular, several systems have difficulties dealing with overlapping sound events, which occur frequently in the meeting room environment [38].

Other works are focussed on detecting sound events in real-world recordings, for example to tag a holiday recording as being at the "beach" playing with the "children" and the "dog". Several works in this area have directly utilised conventional techniques from ASR [42, 43], using HMM to detect and segment the sound events from the continuous audio stream. A second layer can also be added to such approaches to improve the recognition performance. For example, in [44] the first layer output, from a conventional HMM system, is rescored using noise-adaptive SVM kernel classification to give an improved recognition result. More recent works have also looked towards feature selection to improve the performance of real-world SER systems [45], or incorporated semi-supervised learning and novel classification methods [46].

Acoustic Surveillance While visual clues are standard for many surveillance and monitoring scenarios, in some situations it may be easier to use only the audio information for detection. In other situations it may be possible to use audio as a complementary source of information. Therefore, this application has received a considerable amount of research, and can be divided according to human or animal surveillance.

Acoustic surveillance of human environments is the task of automatically detecting abnormal situations based on the audio recorded from the environment. Examples include monitoring of the office environment [47], or detection of aggressive sound events, such as screams, explosions and gunshots [48–50]. Typically, acoustic surveillance systems use low-level audio features, such as those found in the MPEG-7 standard [25]. However, a key problem is the robustness of the system to noise. One approach to improve the robustness is to model the steady-state environment, and use the principle of "novelty detection" to detect sound events that differ from this normal [48]. Such systems can then be further enhanced by using feature selection [50], or a hierarchical classification structure [49].

Acoustic surveillance is also used in the monitoring of biological ecosystems,

often with a particular focus on detecting and counting bird vocalisations [11,51, 52]. The advantage of such systems is that they allow for long term monitoring of sensitive ecosystems without requiring the presence of an observer. Other applications include the recognition of animal sounds [53] using a range of audio pattern recognition approaches. While environmental noise again provides the biggest challenges for these systems [11], in some cases the nature of particular vocalisations can be utilised to improve the robustness. For example, certain bird sounds can be detected by extracting a specific frequency content or searching for a defined repetition rate [52].

Environment Classification This is the task of recognising the surrounding environment from the audio signal, for example an office, street, or railway station [54]. It is sometimes referred to as scene or context recognition in the literature [55,56]. The idea of this application is that information can be gained from the environment to be used for the next generation of context sensitive devices. By recognising the environments, the device can gain valuable information regarding the user's current location and activity and adjust its settings accordingly [56].

Environment classification can also be used in hearing aid systems, which can tune parameters such as the audio compression and the use of directional microphones to yield the best listening experience for the user for the current environment [57–59]. A common approach for such systems is to directly utilise approaches found in ASR [60,61], or to boost the performance by extracting additional features [62]. However, performance comparison between these systems is difficult, as there is no standardised database for evaluation of the different methodologies.

2.1.3 Challenges

In the previous section, it can be seen that ASR techniques have sometimes been used to address the challenges faced in SER applications. This is possible since they share a similar system structure, and are both based on similar signal processing concepts. However, the nature of the challenges faced in each domain are different, and SER

Problem Faced	**Speech (ASR)**	**Sound Events (SER)**
Scope	Language, emotion, speaker, accent	Localisation, segmentation, categorisation
Recording Environment	More controlled, often close-talk	Uncontrolled, variable distance, low SNR
Detection	Speech/non-speech, long segments	Sound event/noise, short segments, potential overlap
Feature Extraction	Frame-based, captures vocal information	Segment-based, wide range of signal information
Pattern Recognition	HMM, connected phonemes	HMM/SVM/ANN, unconnected events

Table 2.2: Comparison of the challenges faced in conventional ASR and SER systems.

systems should be specifically designed to address the problems at hand. A summary of the challenges is given in Table 2.2, which also compares them to the equivalent problem faced in ASR. These are grouped according to the key aspects involved in designing a typical SER system, including the scope of the problem, the recording environment, detection, feature extraction and pattern recognition. A discussion on each of these challenges is now given, with comparisons made between the equivalent problems faced in the field of ASR:

Scope of the Problem The scope of the task at hand is very different in ASR compared to SER. This is because the underlying sound information in ASR has a structure according to the language of the speech. Therefore, language modelling is an important topic of research that can greatly improve the performance of a system by using contextual knowledge about the words being spoken [63]. It also gives rise to the sub-topic of language identification, which in turn gives rise to machine translation [64]. There are also other aspects that contribute to the variability of speech that must be considered. These include the speaking rate, the emotion and accent of the speaker [29] and also code-switching of the spoken language [65]. Further to this, there are also sub-topics such as speaker verification for biometrics [66] and speaker diarisation, which is the task of identifying "who spoke when" [67].

For SER, the underlying sound event information is even less structured, partic-

ularly as no sub-word dictionary exists in the same way as for languages. Despite this, it is possible in certain applications to use a prediction-driven approach to infer a limited amount of context information from the sound event sequence [21]. However, research is typically focussed on other aspects such as localisation of sound sources in the environment [68–70], and segmentation of sound events from the continuous audio [71, 72]. In addition, the similarity and categorisation of sounds is another topic of interest that aims to group sound sources that have a similar behaviour and characteristics to bring about some ordering to the search space [73, 74].

Recording Environment Despite recent advances in distant speech processing [75, 76], ASR traditionally focuses on speech recorded from close-talking microphones. This increases the signal-to-noise ratio (SNR) and reduces the effect of the surrounding environment [77]. However, most applications in SER are in uncontrolled environments, where the distance of the source, the degree of background noise, and the nature of any interfering sources are all unknown. Therefore, noise reduction or compensation is important, and SER systems are often evaluated in mismatched training and testing conditions.

Detection In ASR, this module consists of a speech/non-speech classification system, where any non-speech segments are often considered to be noise and rejected [78]. The detected speech segments may then have a relatively long duration, of the order of several seconds, and contain several interconnected words. These can be further decomposed into a sequence of phonemes, which cannot easily be separated by any stand-alone detection mechanism.

In contrast, sound events are more commonly disconnected from one another, and are less likely to have a strongly interconnected temporal structure in the same way as phonemes in speech. This enables the use of a wider range of detection modules to perform segmentation of the continuous audio. For example, one approach is to use a novelty-detection system that considers any rapid change against the long-term background noise to be a sound event [48]. An alternative is to use a sliding window detector that performs classification on each fixed-length segment in turn [79].

Feature extraction Due to the physical nature of speech production, it is common in ASR to extract frame-based acoustic features which capture the essential information about the vocal tract shape [80, 81]. While SER can be based on such features [60, 82], sound events contain a wider range of characteristics and non-stationary effects, which may not be captured in such frame-based features [3]. Hence, it is common to incorporate additional features to better capture the audio signal [2]. In addition, feature selection can also be used to derive an appropriate set to improve the recognition of a particular sound class [83].

Pattern Recognition As the phonemes that occur in ASR are not isolated acoustic events, it is common to use Hidden Markov Models (HMMs) to find the most likely sequence of phonemes given a set of input features [84]. While such an approach also works for SER, it is also possible to use segment-based features that capture both spectral and temporal information. Such features capture the information from a sound event that is contained within a segmented sound clip. This allows a wider range of methods to be applied for classification, such as Support Vector Machines (SVM) [38, 85], or Artificial Neural Networks (ANN) [86].

Given these important aspects, it is clear that the design of an SER system should be tailored to address the specific problems at hand. However, before reviewing these state-of-the-art SER techniques, it is important to understand the design of a conventional sound event recognition system. This is introduced below.

2.1.4 Typical Sound Event Recognition System

A typical SER system is composed of the following key modules: detection, feature extraction and classification, as shown in Fig. 2.4. The idea is that the detection module first segments sound events from the continuous audio signal, before feature extraction is performed to characterise the acoustic information for classification. Finally, classification matches the unknown features with an acoustic model, learnt during a training phase, to output a label for the segmented sound event. Each module forms their own distinct area of research, and face their own set of basic challenges. Therefore, this section provides a review of the typical approaches for each module in the literature.

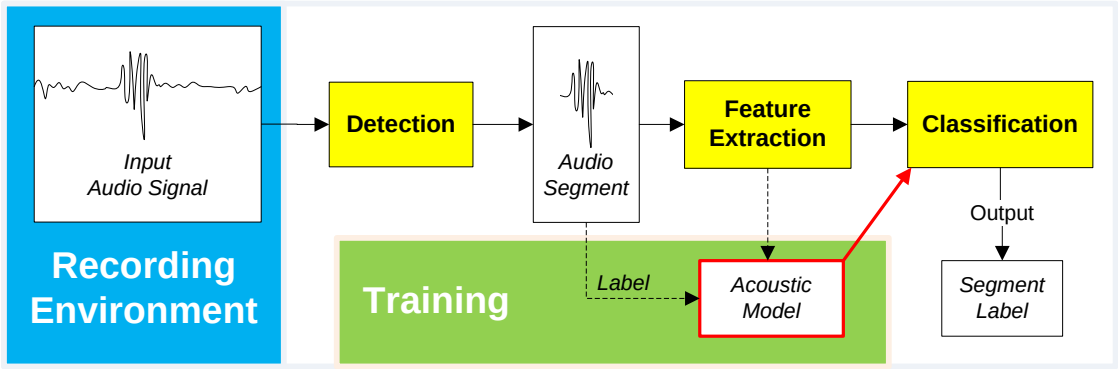


Figure 2.4: The structure of a typical sound event recognition (SER) system

Additionally, the problem of noise robustness is important in the design of a typical SER system, hence a range of solutions to this problem are also discussed.

Detection

The detection module is concerned with finding the start and end points of each sound event, and segmenting it from the continuous audio stream. In general, approaches can be assigned as belonging to one of two categories: detection-and-classification, or detection-by-classification [79], where the latter combines detection and classification into a single pattern recognition problem.

The detection-and-classification approach consists of a separate detection module that extracts a sound event segment, of a variable length, from the continuous audio signal. Low-level audio features are commonly used for this task, such as zero-crossing rate, higher-order statistics, pitch estimation, or spectral divergence [78]. The resulting segmentation does not try to interpret the data, but is a form of "novelty detection" that aims to detect segments that are different from the underlying background noise. This is achieved by comparing the feature values to a threshold that is learned from the recent noise profile [79, 87]. The advantage of this approach is that a fixed length segment does not need to be chosen in advance, which is beneficial in SER tasks where the duration of different sound events may vary considerably. However, the disadvantage is the difficulty in choosing a suitable threshold, as this is crucial and may vary over time in non-stationary noise.

Alternatively, the detection-by-classification approach performs classification of se-

quential segments extracted from the audio, where the detection window shifts forwards over time [38]. The output at each time step is then a decision between noise, or one of the trained sound events [44, 88]. The advantage of this approach is that only one set of features needs to be extracted from the audio as the detection and classification modules are combined. The disadvantage is in choosing an appropriate window size and classification method that can work well across a range of experimental conditions.

Feature Extraction

The purpose of feature extraction is to compress the audio signal into a form that characterises the important sound event information. A good feature should be able to discriminate easily between different classes of sounds, while keeping the variation within a given sound class small. It should also be insensitive to external influences, such as noise or the environment.

The most popular approach is to extract a feature from sequential short-time windowed frames, each around 25-60 ms in length [89]. These are known as frame-based features, since each frame of the signal is represented by a single vector. The most popular frame-based features are Mel-Frequency Cepstral Coefficients (MFCCs) [90], which represent the discrete cosine transform (DCT) of the log-spectral power of the signal mapped onto the non-linear Mel frequency scale. The zeroth coefficient, representing the mean energy, and the higher-order coefficients are often discarded, leaving a compact representation of the spectral envelope. In addition to MFCCs, there are a wide variety of other features that have been developed to capture the information contained in the signal. These include features that can capture the temporal evolution of the signal, the harmonic or perceptual information, or the sound information across time and frequency [91].

As certain features may be more suitable for recognising a particular sound event class than others, it is often necessary to choose a feature set for a particular task. While it is possible to use prior knowledge about the signal and the performance of individual features to choose a feature set, it is more common to perform this selection automatically. There are a variety of different feature selection algorithms to choose from [92], but in general the idea is to select the most discriminative features for a particular classification task [44, 45].

Classification

The purpose of this module is to classify the extracted features to produce a label assigning an audio segment to one of classes presented during training. A basic approach is to simply store the training features in a database, and use a distance measure to compute the similarity between the database and the features observed during testing. This is the basis for techniques such as k-Nearest Neighbours (kNN) and Dynamic Time Warping (DTW). However, such techniques are sometimes less popular as they require more computational cost than the equivalent model-based methods [63]. The most popular techniques therefore create a model of the feature vector space during training, and then measure the distance between the observed features and the model during testing. This idea is found in each of the following techniques: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) [93]. As these are fundamental to a wide variety of state-of-the-art methods SER and ASR systems, they are now briefly described below.

Gaussian Mixture Models – This method models the feature space as a mixture of Gaussian density components, where the observed features for audio processing tasks are typically real-valued vectors, $y \in \mathbb{R}^L$. The distance between an observed feature and the model, $f(y)$, is then as follows:

$$f(y) = \sum_{m=1}^{M} P(m)\mathcal{N}(y; \mu_m, \Sigma_m), \tag{2.1}$$

where M is the number of Gaussian mixtures, $P(m)$ is the prior weight of each mixture component, and $\mathcal{N}(y; \mu, \Sigma)$ is the Gaussian density with mean, μ, and covariance, Σ, evaluated as follows:

$$\mathcal{N}(y; \mu, \Sigma) = \frac{1}{(2\pi)^{L/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right). \tag{2.2}$$

For training, a set of features vectors, $Y = \{y_1, y_t, \ldots, y_T\}$, is used to determine the above parameters using the Expectation-Maximisation (EM) algorithm. This starts with an initial guess of the model parameters, and subsequently repeats the E-step, to evaluate the current model, followed by the M-step to maximise the log-likelihood of

the model.

Hidden Markov Models – The HMM extends the modelling ability of the GMM, by modelling the temporal information of the features. This is achieved using a set of interconnected hidden states, where the output probability distribution of each state is modelled by a GMM, and the transitions between states are determined by a set of probabilities [84]. Given a sequence of input feature observations, the aim is to calculate the most likely sequence of states that could account for the observations [94]. Mathematically, y_t is defined as the observation at a given time instance, t, and $q_t \in \{1, \ldots, K\}$ as the hidden state, where there are K possible states in the model. The HMM parameters, θ, that need to be estimated are therefore 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 $P(y_t|q_t)$. This is performed using the Baum-Welch algorithm to maximise the likelihood of the training data, Y:

$$\theta^{k+1} = \arg\max_{\theta} P(Y|\theta^k) \tag{2.3}$$

where the Baum-Welch can be seen as an implementation of an EM algorithm [84].

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, q_{best}, can be written as follows:

$$q_{best} = \arg\max_{q} P(Y, q|\theta) \tag{2.4}$$

$$= \arg\max_{q} P(Y|q, \theta).P(q|\theta). \tag{2.5}$$

The HMM-based approach is popular in many audio processing application, since it performs a kind of late-integration of the temporal information, hence can model the time evolution of the frame-based feature vectors [92].

Support Vector Machines – This is a binary classifier that calculates the separating hyperplane between two clusters of points in a high-dimensional space [95, 96]. Mathematically, each feature vector, $y \in \mathbb{R}^L$, is treated as a point in an L dimensional space. The set of training feature vectors $Y = \{y_1, y_t, \ldots, y_T\}$, is then represented as a $T \times L$ data matrix, plus a label matrix d, where $d_t \in \{-1, +1\}$ indicating the class label. Assuming the data is linearly separable, the separating hyperplane can be

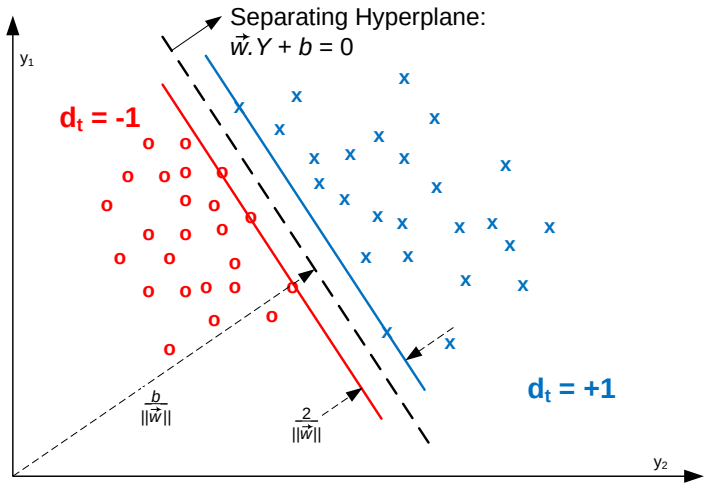


Figure 2.5: Diagram to show the problem formulation for linear SVM.

described by:

$$\vec{w} \cdot Y + b = 0 \tag{2.6}$$

where $\vec{w}$ is the normal to the hyperplane and $\frac{b}{\|\vec{w}\|}$ is the perpendicular distance from the hyperplane to the origin, as shown in Fig. 2.5. The goal is to find parameters $\vec{w}, b$ such that the hyper-plane that best separates the two clusters is found, which provides the maximum margin between the closest points of each class. The margin is found using geometry to be $\frac{2}{\|\vec{w}\|}$. Hence, the task is to minimise $\|\vec{w}\|$, subject to the constraint that points should not fall into the margin. This problem can be solved by the following quadratic program:

$$\min \tfrac{1}{2}\|\vec{w}\|^2 \qquad s.t. \quad d_t(y_t \cdot W + b) - 1 \geq 0 \quad \forall_i. \tag{2.7}$$

While this SVM formulation considers only the linear separation of two classes, modifications can be adopted to add support for overlapping data, non-linear kernel mappings, and solutions for multi-class problems [95].

Noise Robustness

Many speech and sound event recognition systems are trained with data recorded in environments with a high signal-to-noise ratio (SNR). However, previous studies have clearly shown that increasing the difference in SNR between training and testing

rapidly decreases the performance [97]. This decrease in performance is due to the distortion of the spectral information by the noise, and can be summarised by the MixMax principle [98]. This states that the interaction of two signals, s_1, s_2, in the log-spectral domain can be written as follows:

$$\log\left(|s_1| + |s_2|\right) \approx max\left(\log|s_1|, \log|s_2|\right). \tag{2.8}$$

The result is that certain spectral components become masked by the noise or changed significantly enough to affect the feature vector for classification.

A simple solution to this problem is called "multi-conditional" training, which provides the recognition system with data from a variety of different noise conditions and SNRs during training. This allows the acoustic model to capture information about how the sound events might be received in an unseen noise environment. The drawback of this approach is that it requires a large amount of training data, which often is simply simulated and may not match the real-life conditions of the testing environment. In addition, multi-conditional training often reduces the recognition accuracy under high SNR conditions, due to the reduced discrimination of the acoustic models.

Other conventional techniques for noise robustness typically fall into three categories: signal enhancement, feature compensation, and model adaptation. Each address a different module of the typical SER system, as shown previously in Fig. 2.4. Firstly, signal enhancement aims to reduce the amount of noise in the audio signal captured in the recording environment. This enables the extracted features to be closer to the clean training condition. Examples of signal enhancement approaches include Wiener filtering, spectral subtraction [99] and the ETSI Advanced Front End toolkit [100]. Secondly, feature compensation can carried out in the feature domain, 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) [101], which are simple but effective methods for improving the performance in mismatched conditions. Finally, model adaptation can be applied to adapt the acoustic models, trained on clean data, to provide a better representation of the noisy features extracted from the signal during testing. Example of this include parallel model combination (PMC) and maximum likelihood linear regression (MLLR) [102].

An alternative approach for noise robustness is missing feature theory. This is different from the above methods as it treats noise-corrupted regions of the feature as unreliable, and therefore missing [103,104]. This effectively splits the feature vector, y, into reliable, y_r, and unreliable, y_u, components to form a missing feature "mask" [105]. The challenge then is to accurately estimate this mask, with extensive research carried out on this topic in the field of Computational Auditory Scene Analysis (CASA) [24]. Here, principles such as sequential organisation and grouping are used to segment the spectrogram [106–108], with motivation provided by a study of human hearing mechanisms [20,23]. Missing feature classification is then performed either by restoring the missing feature elements, or modifying the classifier to accept the missing feature elements [109]. These approaches are called imputation and marginalisation respectively, and are summarised below.

Imputation – This proceeds by finding suitable replacements for y_u by using values drawn from the conditional distribution of the observation probabilities for a given state, q_i, given the reliable components. It is shown in [103] that this simply equates to the weighted mean of the mixture components for the unreliable data:

$$\hat{y_{u,i}} = \sum_{m=1}^{M} P(m|y_r, q_i)\mu_{u|m,q_i} \tag{2.9}$$

where $P(m|y_r, q_i)$ are the "responsibility factors" for each mixture component, as follows:

$$P(m|y_r, q_i) = \frac{P(m|q_i)\mathcal{N}(y_r; \mu_m, \Sigma_m|q_i)}{\sum_{m=1}^{M} P(m|q_i)\mathcal{N}(y_r; \mu_m, \Sigma_m|q_i)} \tag{2.10}$$

where $\mathcal{N}(y; \mu, \Sigma|q_i)$ is the Gaussian density with mean μ and diagonal covariance Σ as in (2.2), evaluated for state, q_i.

Marginalisation – This aims to compute the output probabilities of each state by integrating out the distribution of the unreliable components within the bounds of the observed spectral information [103]. By utilising the independence of the GMM mixture components, the reliable and unreliable parts of the feature vector can be treated separately, allowing the unreliable components, y_u, to be integrated out as

follows:

$$f(y_r|q_i) = \sum_{m=1}^{M} P(m|q_i)\mathcal{N}(y_r; \mu_m, \Sigma_m|q_i) \int \mathcal{N}(y_u; \mu_m, \Sigma_m|q_i)\ \mathrm{d}y_u \tag{2.11}$$

Marginalisation can also make use of soft masks, where each element is assigned a probability that it is reliable rather than making a hard decision [104].

Both imputation and marginalisation can also utilise the observed noise energy as an upper bound in the computation to improve the performance. The choice depends on the application, since although marginalisation typically performs better [103], imputation has the advantage in that it reconstructs an estimate of the clean feature vectors. These can then be used in subsequent processing by a conventional recognition system, such as extracting MFCCs from the restored data.

2.2 State-of-the-Art Approaches

In the previous section, an overview of sound event recognition (SER) was provided, with a discussion on the potential applications, the challenges faced in SER, and the design of a typical SER system. Next, a review of the state-of-the-art techniques for SER and their limitations is provided. The aim is to give a deeper insight into the range of recent techniques that have been developed specifically for SER, which here are broken down into two categories. The first category focuses on extracting novel features that can better capture the information in the sound events compared to traditional ASR features, such as MFCCs. The second category focuses on novel approaches to modelling the acoustic signal, which often draw on inspiration from the human auditory system. The approaches in these categories are summarised in Fig. 2.6, and are discussed in detail below.

2.2.1 Audio Features

This group of state-of-the-art approaches is concerned with finding novel representations of the sound event information to produce discriminant features for classification. The group can be further broken down into three sub-categories: low-level audio features, temporal features and spectro-temporal features. Each of these are now reviewed

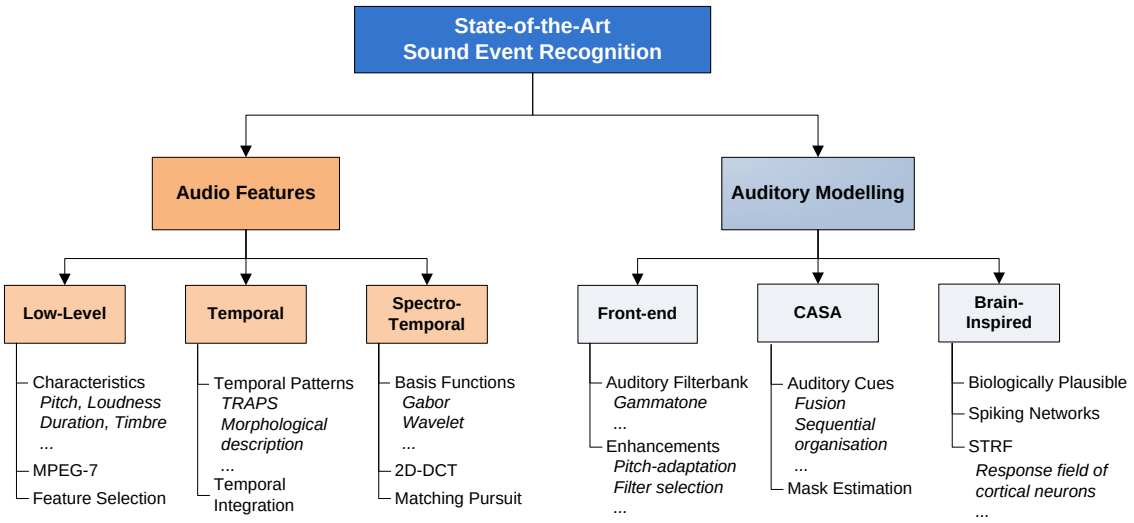


Figure 2.6: Overview of a range of state-of-the-art approaches for sound event recognition.

in detail.

Low-level Audio Features Sound events can be described through a common set of characteristics, such as pitch, loudness, duration and timbre [110]. Although common ASR features can capture these to an extent, they often do so implicitly, rather than designing a feature to capture a specific characteristic directly. For example, MFCCs capture loudness through the zeroth coefficient, and pitch and timbre are represented in the remaining coefficients. Hence, they may not provide the best representation for the full range of sound characteristics, despite providing a good SER baseline performance [111].

This leaves scope to develop novel ways of representing the sound event information. In particular, timbre is important as it represents a range of distinctive characteristics about the sound [110]. Therefore, several works focus on extracting novel features that characterise aspects of the sound event timbre. In particular, these look to capture elements such as the spectral brightness, roll-off, bandwidth and harmonicity [110, 112]. Brightness is defined as the centroid of power spectrum, while spectral roll-off measures the frequency below which a certain percentage of the power resides. Both are a measure of the high frequency content of the signal. Bandwidth captures the spread of the spectral information

around the centroid, while harmonicity is a measure of the deviation of the sound event from the perfect harmonic spectrum [110]. Many of these novel sound features are standardised in the MPEG-7 framework to provide a unified interface for modelling audio information [25]. Other descriptors are also added, such as spectral flatness, kurtosis, sharpness, slope, audio power, and fundamental frequency [89]. MPEG-7 also includes an audio spectrum projection (ASP) feature, which can be seen as a generalisation of the traditional MFCC approach, with a class-specific basis projection used in place of the DCT [25]. However, it was shown in [26] that MFCC features can still outperform MPEG-7 ASP features on a simple sound recognition task.

Given such a large set of audio features to choose from, other works have focussed on feature selection. These approaches aim to automate the process of selecting a suitable feature set for a given sound class that can discriminate well against other sounds. For example in [88], 138 low-level audio features are extracted, and decision tree classifiers are used to select suitable features for modelling each sound object. Another example, in [113], uses the correlation-based feature selection (CFS) method on a base set of 79 features, using the implementation in the WEKA toolkit [114]. While such feature selection approaches try to determine the best subset of features from a predefined subset, an alternative approach is to generate a novel feature set by combining a library of elementary operators. This is the approach taken in [83], where their extractor discovery system (EDS) can explore a feature space of containing billions of possibilities. This is achieved by combining up to 10 operators from a set of 76 basic operations, such as Fourier transforms, filters and spectral measures including the centroid. Experiments show that such feature sets generally achieve better performance than base features such as MFCCs, or can achieve an equivalent performance but with fewer features [83].

Temporal Features The varied nature of the acoustic signals means that representing their frequency content alone may not be sufficient for classification. A simple approach is to combine frame-based spectral features, such as MFCCs, with their delta and delta-delta coefficients to capture their local temporal transitions [115]. However, other features aim specifically to capture the important information in

the temporal domain [116]. Fig. 2.7 demonstrates that this temporal information can be extracted across a range of different time and frequency scales, including capturing both spectral and temporal information in the feature. However, the following methods focus solely on extracting the temporal information from the signal as in Fig. 2.7b, while spectro-temporal features are discussed later in this section.

An early approach for temporal feature extraction for ASR is called "temporal patterns" (TRAPS), which extracts features over a long temporal window from each frequency subband [117]. More recently, a number of related approaches have been proposed for SER. One example can be found in [118], where the aim is to characterise sound events through a "morphological" description of the temporal information in the signal. This includes properties such as the *dynamic profile*, e.g. whether the sound has an ascending or descending energy profile, the *melodic profile*, describing the change in pitch, and the *complex-iterative* nature of sound repetitions. The advantage of this approach is that it naturally describes the sound events in a form that is similar to human description, making the technique useful for indexing sounds for an audio search engine [118]. A different approach is proposed in [85], where the aim is to characterise sound events through a parametric representation of their subband temporal envelope. This captures the distinctive spectro-temporal signature of the sound events, and allows a comparison of sounds using a distance measure based on the parametrised representation.

A different approach for capturing the temporal information in the signal is to use temporal feature integration to transform a set of frame-level features into a segment-level feature vector for classification [38, 110]. This is more common in SER compared to ASR, as sound events more commonly occur in isolation compared to the connected phonemes in speech. Typically, a statistical model of the temporal information is used, where parameters such as the mean, variance and higher-order statistics are captured [92]. However, these simple statistics ignore the temporal dynamics among successive feature vectors. One solution is to model the temporal information by fitting an autoregressive (AR) model to the sequence [119]. Alternatively, a classifier such as HMM can be used that

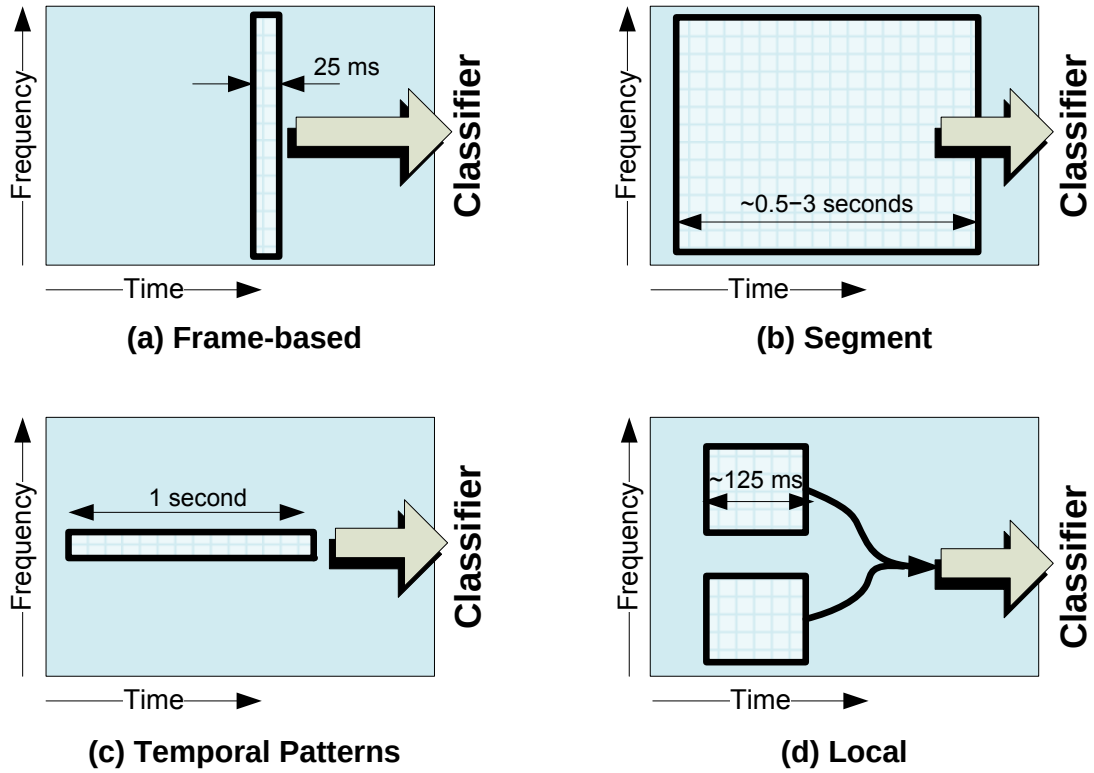


Figure 2.7: Schematic of frame-based, temporal, segment and local feature paradigms for signal processing on the spectrogram.

performs a kind of "late" feature integration [120] by fitting a generative model to the temporal evolution.

Spectro-Temporal Features A natural extension to both spectral and temporal feature extraction is to consider features that jointly model the spectro-temporal information. For ASR, this approach has been used to better capture certain features of speech, such as formants and formant transitions [121]. The most common approach is to use the correlation between a set of wavelet functions and the time-frequency base representation to extract a conventional feature for classification. The most popular wavelet representation is based on complex Gabor functions [122], which are two-dimensional sine-modulated Gaussian functions that can be tuned to model a range of spectro-temporal patterns.

Recently, there has been interest in extracting information from local time-

frequency regions in the spectrogram, as shown diagrammatically in Fig. 2.7d. One approach is to perform the two-dimensional Discrete Cosine Transform (DCT) of each local time-frequency patch on a regular grid, which is equivalent to performing correlation with a set of 2D-DCT bases [123]. The result can then be concatenated together to form a frame-based feature, which can be improved further by removing the higher-order components to provide both smoothing and dimensionality reduction [124]. While such approaches often used a fixed set of basis functions, a recent proposal aims to learn the spectro-temporal modulation functions from the data [125]. Independent Component Analysis (ICA) is used for this purpose, and it is shown that this approach can learn functions that give an improved performance.

A related approach is based on decomposition of the time domain signal using Matching Pursuit (MP) [126]. This provides an efficient way of selecting a small basis set that represents the signal with only a small residual error. As before, a Gabor wavelet dictionary is commonly used [127], as it can better capture the non-stationary time-frequency characteristics in the signal compared to the one-dimensional Haar or Fourier bases [28]. It has also been noted that the Gabor bases are more effective at reconstructing a signal from only a small number of bases [2]. Another advantage of using MP is that the decomposition has a denoising effect on the signal [3]. This is because the residual from the decomposition is discarded, which can be assumed to contain incoherent elements of the signal that are not modelled by the basis functions.

2.2.2 Auditory Modelling

This group of state-of-the-art approaches is concerned with finding novel ways of modelling the audio signal, often by drawing on inspiration from the human auditory system. The group can be further broken down into three sub-categories: modelling of the auditory front-end, computational auditory scene analysis (CASA), and brain-inspired neural approaches. Each of these are now reviewed in detail.

Auditory Front-End The traditional method for acoustic analysis is to use the short-time Fourier transform (STFT), combined with perceptual filtering through

a set of triangular-shaped Mel filters [128]. This forms the basis for traditional MFCC features. However, the STFT has several drawbacks that may limit its effectiveness, particularly when applied to sound events. Firstly, the method assumes the signal to be stationary during the short-time analysis window, which may not be true for highly non-stationary sounds. Secondly, there is a trade-off between the time and frequency resolution of the STFT representation, as it is closely linked to the size of the analysis window. Therefore, a recent trend has been to find alternative ways of performing the time-frequency decomposition.

One popular method is inspired by studies into the auditory modelling in the human ear, in particular the signal analysis performed by the human cochlear [129, 130]. This is called as the gammatone filterbank decomposition [131], as the filter function can be characterised as a sine-modulated gamma distribution function. As an approach for time-frequency analysis, gammatone filtering overcomes the drawbacks associated with the STFT, since there is no longer a trade-off between time and frequency resolution. In addition, it has the advantage that the gammatone function has been shown to be highly correlated with natural sounds [132], hence should provide an efficient representation.

A typical system, based on such auditory modelling, uses the gammatone time-frequency analysis to enhance the traditional pipeline for audio signal processing [36, 133]. The resulting features are called the Gammatone Cepstral Coefficients (GTCCs), since they simply replace the STFT and Mel filtering modules used in MFCC extraction with Gammatone filtering [134]. For non-speech audio classification, it has been shown that GTCCs can outperform both MFCCs and MPEG-7 using both kNN and SVM classifiers [86]. Further to this, recent works have also demonstrated improved performance through pitch-adaptivity [135] or selection [136] of the gammatone filterbanks. For example, in [136], filterbank channel selection is performed to adapt an SER system to changing environmental conditions. This was shown to outperform both MFCC and selective Mel-filterbank features.

An alternative to this is to use a more complete auditory model based on the next stage of auditory processing in the cochlea. This is formed by 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 [137, 138]. An example of a recent work based on this is the auditory image model (AIM) [139]. This model generates a stabilised auditory image (SAI), and the feature extracted from this is shown to outperform MFCCs on a task involving different speaker characteristics.

Computational Auditory Scene Analysis This is a field that originated in the 1990's based on Bregman's work on human perception of sound though Auditory Scene Analysis (ASA) [20]. 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 [140]. Bregman's interpretation is that human sound perception is like building up a picture of the "auditory scene". This idea is exemplified by the "cocktail party" scenario, where a human listener is easily able to follow a conversation with a friend in a room with many competing conversations and acoustic distractions. Through this work on ASA, cues were discovered that the auditory system uses to understand sounds. The main two effects present are "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 [21]. It was found that the main cue for fusion was having a common onset, within a few milliseconds, while learned schema appeared to be the main cue for assigning disconnected frequency elements to the same sound source.

This understanding has been utilised in audio processing in several ways. The most notable of these has been for the process of mask estimation for missing data classification [24], which was previously introduced in Section 2.1.4. CASA-based mask estimation includes works based on onset-offset analysis [106, 141], grouping cues [23], pitch-based grouping [107], or top-down segmentation [142]. Much of the work is focussed on speech perception in noise [134, 142] or speech separation from multi-talker mixtures [23, 107]. However, the performance of missing data classification systems is highly dependant on the accuracy of the mask [103]. Therefore, more recent decoding methods have been developed to compensate for the deficiencies inherent in the mask estimation. One approach is based on speech fragment decoding [142, 143], whereby the mask is separated into a set

of local fragments, where each fragment should belong to a single source. The system then proceeds by searching for the best combination of fragments that represent the target source. This can be further enhanced by utilising knowledge about the noise present in the signal [144]. An alternative approach in [108] uses an uncertainty decoder to allow for errors made in the mask estimation process. Here, the unreliable elements are first reconstructed using a prior speech model, and then transformed into cepstral features [134]. However, the uncertainties in the reconstruction are also transformed into the cepstral domain, and are then used to adjust the variance of individual Gaussian model components, which gives improvement over the baseline system.

Brain-Inspired Approaches These approaches are based on biologically plausible pattern recognition, such as emulating effects that have been found in the human brain. The aim to understand the human physiology involved in the auditory system [145–147], and build systems that replicate the processing mechanisms that work very effectively for humans.

One approach is based on the simulated output of the inner-ear processing, in the form of spike trains on the auditory nerve [137, 148]. These can then be used as the input for a spiking neural network for brain-inspired pattern recognition. The approach has been used for both sound [149] and speech [150, 151] recognition, as well as to develop an efficient audio encoding scheme [152]. For example, [150] uses a simple feature based on the sound onset, peak and offset times, and then combines this with a spiking neural network for recognition. The system is trained to produce a set of neurons that are all spiking at the same rate at a particular time after the onset of the sound, 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.

Another approach is based on the measured Spectro-Temporal Response Field (STRF) from the auditory cortex region of the brain [153, 154]. The STRF represents the characteristic response function of a particular neuron in the cortex to a range of frequencies over time, and is found to bear some similarities with Gabor functions [155]. It can be used as a quantitative descriptor for complex

sounds, where it can be used to reconstruct sounds and act a measure of speech intelligibility [155]. The STRF has also been used to discriminate speech from non-speech in the presence of high levels of noise and reverberation, by transforming the auditory STRF representation into a feature that can be used in conventional pattern classification using SVM [156]. The reported results show that the STRF approach compared well to two conventional systems on the same task.

2.2.3 Limitations

The previous section introduced a range of state-of-the-art approaches for SER. However, while such techniques may perform well in matched experimental conditions, there are limitations that may reduce their effectiveness in challenging unstructured environments, where noise, multiple sources and distortion are present. Three common limitations are identified: the frame-based features used, insufficient temporal modelling and the problems of missing feature mask estimation. These motivate the research in this thesis to find an alternative approach to address the problems faced in SER, hence are discussed below.

Frame-based Features These represent each short-time window from the continuous audio signal with a vector of feature values, as shown diagrammatically in Fig. 2.7a. It is the most commonly used approach in ASR, where MFCC features are ubiquitous across many different systems. The limitation of frame-based features is most evident when noise or multiple sources interact within the same short-time window. In such cases, the noisy feature vectors will produce a low score against the clean model, as the feature dimensions do not match with the distribution found in the model trained on clean data.

This problem affects many of the state-of-the-art low-level audio features, such as MPEG-7 [25, 50, 62], and spectro-temporal feature extraction methods [2, 3]. This is because such features are designed to extract more information from the audio to achieve an improved performance, but are not inherently robust to mismatched conditions. Other auditory-inspired techniques also tend to suffer the same problem, as they are commonly combined within the conventional frame-based paradigm. Examples include auditory inspired the Gammatone Cepstral

Coefficients (GTCCs) [86, 134], which simply replace the front-end processing of traditional MFCCs.

Temporal Modelling The modelling of spectral and temporal information in sound events is important, and traditional features such as MFCCs do not capture the temporal information sufficiently. Even when combined with existing temporal feature integration methods, such as taking the mean and variance of the feature vectors across time [92], this still does not fully capture the temporal ordering of the sound. Also, while the temporal modelling of the frame-based features in HMMs is an improvement, the feature vector transitions still follow an exponential distribution [63], which is a poor model of the sound information. Speech is less affected by the deficiencies of the HMM modelling as the duration of the average phoneme duration is short. On the other hand, some sound events can have a quite stationary temporal distribution over longer time periods, which will not be well modelled. Also, while brain-based techniques provide biologically inspired methods for SER, little is actually known about how the auditory cortex models the temporal information [157]. Therefore, such systems often resort to using traditional pattern classification techniques [156], or find other ways to capture the temporal information [150].

Missing Feature Mask Estimation The problem of robust recognition with missing features has been studied extensively [103, 104]. However, it has been shown that the performance of these methods depends heavily on the accuracy of the estimated mask [158]. This poses a significant problem for SER, where the non-stationary nature of the noise across time, frequency and dynamic range makes developing a reliable mask particularly challenging. This is compounded by the difficulty in using feature-based mask estimation techniques [105], as it is difficult to design a feature to reliably capture the wide variety of sound event characteristics. Any noise elements that are incorrectly marked as reliable, such as those with similar sharp peaks to the signal, will severely affect the classification performance in real-life SER applications.

2.3 Baseline Experiments

In the previous section, a range of recent state-of-the-art techniques was introduced, including those based on novel audio features and auditory modelling. Due to the wide variety of techniques and datasets, it is difficult to establish a solid baseline performance amongst these methods from the literature. Experiments are therefore conducted in this section to establish a baseline SER performance amongst both conventional audio processing and more recent state-of-the-art techniques. This also includes methods based on noise reduction, multi-conditional training, and missing feature techniques, which are popular techniques for dealing with real-world environmental conditions. Training is carried out using only clean samples, with the methods tested on a standard database of environmental sounds in both clean and mismatched noise conditions. These noise conditions, including speech babble and factory floor noise, are chosen to simulate a more realistic testing environment.

The rest of this section first details the experimental setup and baseline methods that are implemented, before discussing the results that are obtained.

2.3.1 Experimental Setup

Database

A total of 50 sound classes are selected from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustical Environments [159], giving a selection of collision, action and characteristics sounds. The isolated sound event samples have a high signal-to-noise ratio (SNR), and are balanced to give some silence either side of the sound. The selected categories cover a wide range of sound event types, including wooden, metal and china impacts, friction sounds, and others such as bells, phones ringing, and whistles. Many of the sound events have a sparse time-frequency spectrogram representation, with most of the power contained in a particular frequency band, while several others are more diffuse, such as the buzzer or sandpaper sounds.

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

For each experiment, except for the multi-conditional method, the classification accuracy is investigated in mismatched conditions, using only clean samples for training. The average performance for each method is then reported in clean and at 20, 10 and 0 dB SNR for the following four noise environments: "Speech Babble", "Destroyer Control Room", "Factory Floor 1" and "Jet Cockpit 1", obtained from the NOISEX'92 database [160]. The noise segments are randomly selected from the above samples, then scaled to the correct SNR and artificially added in the time domain. All four noises have their energy concentrated in the lower frequencies, and represent realistic non-stationary noise conditions. The average performance across all four noise environments is reported at each SNR.

Conventional Methods

To establish a baseline performance for SER, a broad range of methods for sound event recognition are investigated. The following conventional approaches are implemented:

1. Base Methods:
   (a) MFCC-HMM, using 36-dimension frame-by-frame MFCCs, with 12 cepstral coefficients, without the zeroth component, plus their deltas and accelerations.
   (b) MFCC-SVM, with temporal integration performed by extracting the mean and variance of the MFCCs over the clip.
2. Noise Reduction:
   This uses a system based on the MFCC-HMM method above, but noise is removed for both training and testing using the following algorithms:
   (a) Spectral Subtraction [99]
   (b) Advanced Front End (AFE) [100].
3. Missing Features [103]:
   Here, an HMM system is trained using 36-dimension Mel-frequency spectral coefficient (MFSC) features without deltas. A missing feature mask is then gener-

ated using a noise estimate based on the first 10 frames. The following missing feature methods are then implemented:

   (a) Bounded Imputation.

   (b) Bounded Marginalisation.

4. Multi-Conditional training:
   Here, the baseline MFCC-HMM system is trained with features generated in both clean and 10dB SNR noise, under 3 out of the 4 noise environments. Testing is then carried out on the remaining noise environment.

Each of the HMM methods above uses 5 states and 6 Gaussian mixtures, as this was found to provide a good trade-off between performance and computational complexity in preliminary experiments. Training and testing are both carried out using the popular hidden Markov model Toolkit (HTK) [161]. The only exception are the missing feature methods, where testing is carried out with a local Matlab HMM decoder, which is modified to perform the missing feature imputation and marginalisation.

State-of-the-Art Methods

In addition to the baseline methods above, a range of more recent methods are evaluated to represent a cross-section of the state-of-the-art techniques introduced in Section 2.2. The following methods are implemented:

1. MPEG7-HMM [25]:
   This method extracts 57 features from each frame including the audio power, fundamental frequency, zero crossing rate, and the short-time spectrum envelope, centroid, roll-off, spread and flatness. The dimension is reduced to 12 using PCA, and combined with their deltas and accelerations to provide a 36 dimension feature comparable to MFCCs.

2. Gabor-HMM [122]:
   The best performing 36 Gabor features are selected using a feature finding neural network (FFNN) [121]. This consists of a linear single-layer perceptron in conjunction with secondary feature extraction and an optimization rule for the feature set.

3. GTCC-HMM [134]:
   Here, gammatone cepstral coefficients are extracted using a total of 36 gammatone filters. The dimension is subsequently reduced to 12 using the DCT and combined with deltas and accelerations to give a 36 dimension feature.

4. MP+MFCC-HMM [2]:
   Matching-pursuit is used to decompose each signal window to find the top 5 Gabor bases. Four features are then derived, by finding the mean and variance of both the frequency and scale parameters of the Gabor bases. These are concatenated together with conventional MFCC features, and combined with their deltas and accelerations to give a $(12 + 4) \times 3 = 48$ dimension feature vector.

As with the previous baseline methods, each system uses a 5 state HMM with 6 Gaussian mixtures, with both training and testing performed using HTK.

2.3.2 Results and Discussion

The experimental results for each of the methods is now presented, with the aim to establish a solid baseline to provide a comparison with the work in this thesis. The performance of the conventional methods is now analysed, followed by a further comparison against the results achieved by the state-of-the-art methods.

Results: Conventional Methods

The performance of the conventional methods is reported in Table 2.3. Firstly, comparing the two base systems in the top segment of the table, it can be seen that MFCC-HMM produces the best performance, with an average classification accuracy of 57.3%. Compared to this, the MFCC-SVM system performs much worse than MFCC-HMM, despite using the same base features. This is explained by the different kinds of temporal integration performed in each method. In particular, MFCC-SVM uses the mean and variance of the features over the whole clip, which captures less of the temporal information compared to the late-integration performed by MFCC-HMM.

Next, the performance gain achieved by using the two noise reduction techniques is evaluated. The results show that the ETSI Advanced Front End (AFE) achieves an average accuracy of 73.9%, which is marginally better than 72.5% achieved by

Group	Method	Clean	20dB	10dB	0dB	Avg.
Base	MFCC-HMM	$\mathbf{99.4 \pm 0.1}$	71.9 ± 8.5	42.3 ± 8.7	15.7 ± 4.2	57.3
	MFCC-SVM	98.5 ± 0.2	28.1 ± 5.0	7.0 ± 2.4	2.7 ± 0.6	34.1
Noise Reduc.	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
Missing Feature	Imputation	94.3 ± 0.5	90.3 ± 1.5	80.4 ± 4.7	60.5 ± 9.2	81.4
	Marginalisation	93.6 ± 0.4	85.6 ± 2.9	74.7 ± 3.0	50.1 ± 7.4	76.0
Multi-Conditional		97.5 ± 0.1	$\mathbf{95.4 \pm 1.3}$	$\mathbf{91.9 \pm 2.7}$	$\mathbf{67.2 \pm 7.3}$	**88.0**

Table 2.3: Classification accuracy results for experiments on the conventional audio processing methods. The standard deviation is also reported ($\pm$) across five runs of the experiment and the four different noise conditions.

spectral subtraction. This is expected since the techniques used in the AFE, such as double Wiener filtering, are much more sophisticated than simple noise reduction using spectral subtraction. The performance is also a considerable improvement over the equivalent MFCC-HMM system without noise reduction, despite a small drop in performance in clean conditions.

The third segment of Table 2.3 shows the results for the two missing feature methods. It can be seen that both imputation and marginalisation approaches can outperform the noise reduction results, with a classification accuracy of 81.4% and 76.0% respectively. However, it was found that the performance in clean conditions was less than the previous methods. One of the factors contributing to this result may be that the missing feature approaches are limited to the spectral domain, where the sparse time-frequency spectrum of the sound events can cause particular mixture variances to become very large, for example when a decaying time envelope was present. In such cases, the likelihood of these mixtures could be high across a number of different sound classes, causing confusion between similar sounds. Further to this, when these mixtures were marked as missing in the marginalisation approach, their likelihood would be very low due to the narrow range of the integration compared to the variance. The result is the poorer overall performance of marginalisation compared to the imputation approach.

The results for the multi-conditional MFCC-HMM method are found in the final segment of Table 2.3. Here it can be seen that the system achieves an average accuracy

of 88.0%, which is a significant improvement compared to both the noise reduction and missing feature methods. The computational cost is also much less than the missing feature approaches, as it does not require any modification to the HMM decoding process. However, it does not necessarily provide a fair comparison with the other methods, as they performed testing in mismatched conditions, with only clean data for training. In addition, the performance of the multi-conditional training method may vary in practical applications, depending on how close the testing noise environment is to those found in training. Despite this, it still achieves a state-of-the-art performance, and provides a well-performing baseline that will be compared to future methods.

Results: State-of-the-Art

The performance of the four state-of-the-art SER approaches is reported in Table 2.4. Here it can be seen that MPEG-7 gives the lowest overall performance, with an average accuracy of just 33.6%. This can be compared to the best performing MP+MFCC method that achieves 58.4% on average. The poor performance of MPEG-7 can be explained by some of the features included, which may not be robust or well suited to environmental sound events. For example, the audio power feature will vary according to the SNR level, and the fundamental frequency feature may be difficult to estimate accurately across the wide range of sound classes.

The results also show that both the Gabor and GTCC methods perform marginally better than the basic MFCC-HMM system in clean conditions. In particular, the Gabor method achieves the best performance with 99.8% of samples correctly classified. However, both methods perform less well in mismatched conditions. For the Gabor method, this can be explained by the feature selection process, which becomes strongly tuned to the training conditions. This means the approach does not generalise as well to the mismatched noise conditions, compared to the simple MFCCs .

Finally, the best overall performance is obtained by the MP+MFCC approach, which uses matching pursuit to complement the MFCC features. As expected, the MP method performs marginally better than the original MFCC-HMM, but with an average improvement of just 1.1%. This is possible because each window of the signal is decomposed into a small number of Gabor bases. These are expected to represent the most prominent information in the signal, which will naturally be retained even under

Method	Clean	20dB	10dB	0dB	Avg.
MPEG-7	97.9 ± 0.3	25.4 ± 2.1	8.5 ± 0.6	2.8 ± 0.7	33.6
Gabor	$\mathbf{99.8 \pm 0.1}$	41.9 ± 6.8	10.8 ± 2.8	3.5 ± 1.2	39.0
GTCC	99.5 ± 0.2	46.6 ± 10.0	13.4 ± 2.5	3.8 ± 1.2	40.8
MP+MFCC	99.4 ± 0.2	$\mathbf{78.4 \pm 5.7}$	$\mathbf{45.4 \pm 6.3}$	$\mathbf{10.5 \pm 2.3}$	**58.4**

Table 2.4: Experimental results for four state-of-the-art methods for sound event recognition.

noisy conditions. The disadvantage is the additional processing required to perform the decomposition, which could be high enough to offset the small gain in performance.

Conclusion

The experimental evaluation can now be used to establish a solid baseline for future comparison. The results show that the best overall performance is achieved by the multi-conditional MFCC-HMM method, with an average accuracy of 88.0%. However, the multi-conditional method requires both clean and noisy samples for training, hence does not provide a direct comparison between other methods that require only clean samples for training. Therefore, another baseline method will be selected based only on clean training data. For this, the ETSI-AFE is chosen, as it is a well known approach for noise robust feature extraction and achieves a good performance of 73.9% on this task.

It should be noted that the missing feature methods also require only clean data, and achieve slightly better results than the ETSI-AFE. However, they are not selected for the following reasons. Firstly, the performance in clean conditions was poor compared to the other baselines, due to problems in modelling the sound events in the spectral domain. Secondly, they require a significant amount of additional processing, due to the additional calculations required during the HMM decoding process. Together, this limits the effectiveness of the missing feature methods as a practical system, meaning they do not provide a good baseline for comparison.

2.4 Summary

This chapter has given an introduction to sound event recognition (SER) and the current state-of-the-art. This was provided by first giving an overview of SER, including the wide array of potential applications, and the challenges faced in building a successful system. Then, an overview of a typical SER system was given, with a review of the traditional approaches for detection, feature extraction, classification and noise robustness. Next, a range of state-of-the-art methods and their limitations for SER was reviewed, with the techniques grouped into two categories. The first category is novel audio features, which covers a range of techniques for producing discriminant features for classification. The second category are auditory modelling approaches, which draw inspiration from the human auditory system to introduce biologically plausible processing methods to complement traditional systems. Finally, a set of experiments were carried out to establish the baseline performance among both conventional and state-of-the-art methods for SER. Together, this chapter motivates the work in this thesis to find novel ways of capturing the sound information. The next chapter introduces the proposed approach, where the idea is to extract features and perform classification based on spectrogram image processing.

Chapter 3

Spectrogram Image Processing

In the previous chapter, an introduction to sound event recognition was given, including a review of the current state-of-the-art techniques and their limitations in mismatched conditions. Many of these techniques are based on a short-time frequency analysis of the continuous audio signal, which can be visually represented as a two-dimensional time-frequency image called the spectrogram. In this chapter, the idea of performing SER by combining image processing approaches with the spectrogram representation of the sound is introduced. This is referred to here as spectrogram image processing, where features are extracted from the two-dimensional spectrogram image to jointly characterise the time-frequency sound information.

The chapter is organised as follows. Section 3.1 first provides motivation for the idea of spectrogram image processing, and discusses the differences between spectrograms and conventional images. Section 3.2 then reviews relevant techniques in image processing that are used in state-of-the-art image classification and object detection systems. Based on this, the Spectrogram Image Feature (SIF) approach is then proposed in Section 3.3 for robust sound event classification in mismatched noise conditions. Experiments are then carried out in Section 3.4 to compare the SIF against a range of baseline techniques.

3.1 Motivation

The time-frequency representation of an audio signal is commonly referred to as the spectrogram, and it can be visually analysed by a trained researcher to recognise many underlying sound events in a process called "spectrogram reading" [4]. However, this has not become a popular approach for automatic classification, as the field is driven by speech research where frame-based features are popular. As opposed to speech, sound events typically have a more distinctive time-frequency representation, each containing a unique mixture of harmonic, impulsive and diffuse spectral structures that can be used to distinguish between each sound. In addition, sound events are more commonly disconnected from one another, unlike the strongly interconnected temporal structure of speech or music. Together, this makes sound events more suitable than speech or music for classification based on their visual signature – an approach referred to in this thesis as "spectrogram image processing".

This rest of this section provides the motivation for using image processing methods applied to the spectrogram as a basis for sound event classification. First, an overview of the idea is given, before introducing the advantages and disadvantages of the most common spectrogram representations. Then, an analysis of the important differences between the spectrogram and conventional images is provided, before reviewing the small number of previous works that have utilised the idea of spectrogram image processing.

3.1.1 Overview

Through visual inspection of the spectrograms of typical sound events, it is clear that a large amount of information is contained in the joint time-frequency representation. With careful analysis, it is possible to recognise similar sound events based on this visual image information. An example of this is given in Fig. 3.1, which shows the spectrogram images of a bottle being tapped in both clean and 0dB noise conditions. Here, the spectral information belonging to the sound event is easily distinguished from that of the background noise, as demonstrated by the highlighted areas in the figure. This is due to the consistent appearance of the bottle sound, which contains characteristic peaks and lines that are connected through a common onset corresponding to

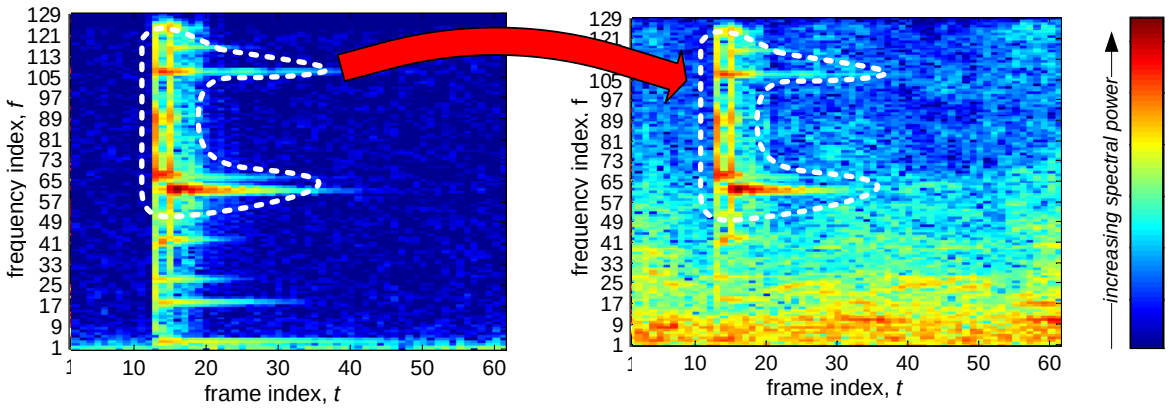


Figure 3.1: Examples spectrograms of a bottle sound in clean (left) and 0dB babble noise (right). The highlighted area demonstrates how the sound information is still represented clearly in severe noise conditions.

the moment of impact as the bottle was tapped. On the other hand, the noise forms the background of the images and can easily be ignored.

In the early days of speech research, visual information in the spectrogram was often studied by speech researchers, for example to analyse the phoneme structure of speech [4]. However, "spectrogram reading" has not become an automatic classification method in speech technology, due to the complicated lexicon structures of speech. Unlike speech, with its connected phoneme structures, sound events often have shorter durations but with more distinctive information contained in the time-frequency image representation. It should therefore be possible to extract this information to provide a discriminative feature for classification of the sound event. Such an approach, based on spectrogram image processing, would also represent a significant departure from conventional audio processing. Here, frame-based features such as MFCCs have historically been dominant, but capture only the frequency information within a short time window. Therefore, recently there has been growing interest in capturing joint time-frequency information from the audio signal [2,3]. Such works have demonstrated the potential advantages of operating in the spectrogram image domain, where the two-dimensional sound event information is naturally represented.

A related field that is concerned with the extraction of two-dimensional information is image processing. Here, two of the most important problems are to detect objects in an image, or classify an image into a predefined category. These problems share

many similarities with those faced in sound event classification, particularly when considering classification based on spectrogram image processing. For example, the whole spectrogram image could be classified by extracting low-level pixel information, or alternatively a "sound object" could be detected by finding local correspondences in the spectrogram between training and testing. This therefore opens up the wide range of techniques that have been developed in image processing, which can provide both the inspiration, and a solid basis for developing novel approaches for SER. However, there are a wide range of spectrogram image representations that can be generated, and each may have different advantages depending on the application. Therefore, these are introduced in the next section.

3.1.2 Common Spectrogram Representations

Throughout this thesis, the spectrogram is referred to generally as a time-frequency representation of the audio signal. However, there are a variety of different methods for generating such representations, beyond the traditional short-time Fourier transform (STFT). The most common techniques are introduced here, and the advantages and disadvantages of each are discussed.

Short-Time Fourier Transform This is calculated using the discrete Fourier transform (DFT) of a windowed frame from the continuous signal, $x_t[n]$, as follows:

$$S_{lin}(f,t) = \left| \sum_{n=0}^{N-1} x_t[n] w[n] \, e^{-i2\pi \frac{f}{f_s} n} \right| \tag{3.1}$$

where N is the number of samples per frame, $f = kf_s/N$ is the frequency bin for $k = 1, \ldots, N/2+1$, w is a window function such as the Hamming window, and t is the time frame index. It is more common to compress the dynamic range of the linear power spectrogram using the log function to give the conventional log power spectrogram. This is referred to simply as $S(f,t)$ and calculated as follows:

$$\begin{aligned} S'_{log}(f,t) &= \log\left[S_{lin}(f,t)\right] \\ S(f,t) = S_{log}(f,t) &= \max\left[S'_{log}(f,t), \max_{f,t}\left(S'_{log}(f,t)\right) - 80dB\right], \end{aligned} \tag{3.2}$$

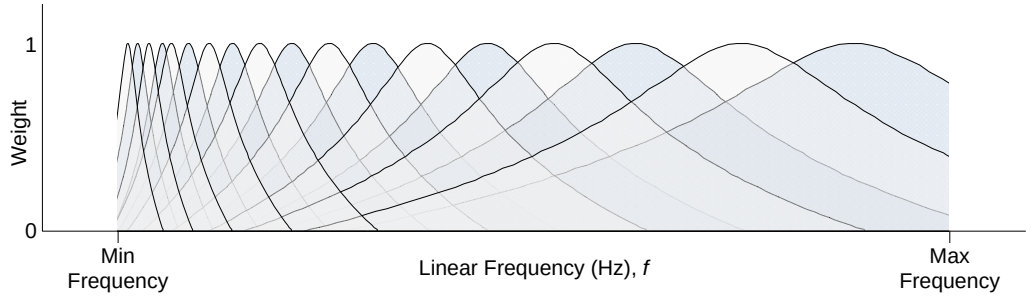


Figure 3.2: Illustration of the Gammatone filterbank.

where a simple thresholding is employed to ensure a consistent log-scaled representation. This is important as the logarithm can become a highly variable quantity as the spectral values tend towards zero.

The advantage of the STFT is that it is simple and fast to compute. However, it makes the assumption that the signal is stationary within each window, which may not be true for many sound events that have sharp discontinuities. In addition, there is a trade-off between frequency and time resolution, as choosing a longer window gives better frequency resolution at the cost of reducing the temporal resolution.

Gammatone This approach is derived from the cochlear filtering in the inner ear, hence can also be referred to as the cochleagram [131]. The impulse response of the filter is the product of a gamma distribution function and a sinusoidal tone centred about a particular frequency, f, as follows:

$$g(t) = t^{(N-1)} e^{-2\pi bt} cos(2\pi ft + \phi) \tag{3.3}$$

where t is the time, N represents the order of the function, ϕ is the phase shift, and b is related to the bandwidth of the gammatone filter. Typically the parameters $N = 4$ and $\phi = 0$ are used, while the values for b can be calculated using the equivalent rectangular bandwidth (ERB) scale as follows:

$$f_{erb} = 24.7 \times (\frac{4.37 \times f}{1000} + 1) \tag{3.4}$$

$$b = 1.019 \times f_{erb} \tag{3.5}$$

where f_{erb} is the ERB frequency for the linear frequency f, using the parameters as recommended in [162]. An illustration of this gammatone filterbank is given in Fig. 3.2. To derive a digital filter for efficient audio processing, the Laplace transform of the gammatone function is taken. For a fourth order filter this gives:

$$G(s) = \frac{6(-b^4 - 4b^3 s - 6b^2 s^2 - 4bs^3 - s^4 + 6b^2\omega^2 + 12bs\omega^2 + 6s^2\omega^2 - \omega^4)}{(b^2 + 2bs + s^2 + \omega^2)^4} \tag{3.6}$$

where $\omega = 2\pi f$. This can then be implemented using an eighth-order digital filter, where the output for each frequency channel f is as follows:

$$y_f[n] = -\sum_{k=1}^{K_a} a_{f,k} y_f[n-k] + \sum_{k=0}^{K_b} b_{f,k} x[n-k] \tag{3.7}$$

where $a_{f,k}$, $b_{f,k}$ are the filter coefficients for a given frequency and filter order. Finally, the gammatone spectrogram, $S_g(f,n)$, is produced by first taking the magnitude of the filter output, followed by compression of the dynamic range by taking the logarithm. This can be written as:

$$S_g(f,n) = \log \left| y_f[n] \right| \tag{3.8}$$

where f represents the centre frequencies of the filters on the ERB scale and n is the sample index of the input audio sequence.

The gammatone spectrogram has the advantage over the STFT that there is no trade-off between time and frequency resolution. In addition it has been shown that gammatone filters are highly correlated with natural sound signals [132], which should produce a sparse, high resolution spectrogram of the sound event. The disadvantage is the increased computational cost, and that the common ERB scale has less resolution at higher frequencies, where a wider frequency response is used to match that found in the basilar membrane.

Mel-Frequency Spectral Coefficients The Mel filterbank provides an approximation to the non-linear characteristics of the human ear, in a similar way to the gammatone filterbank. It is also one of the processing steps used in the calculation of the popular MFCC feature [90]. Compared to the gammatone spectro-

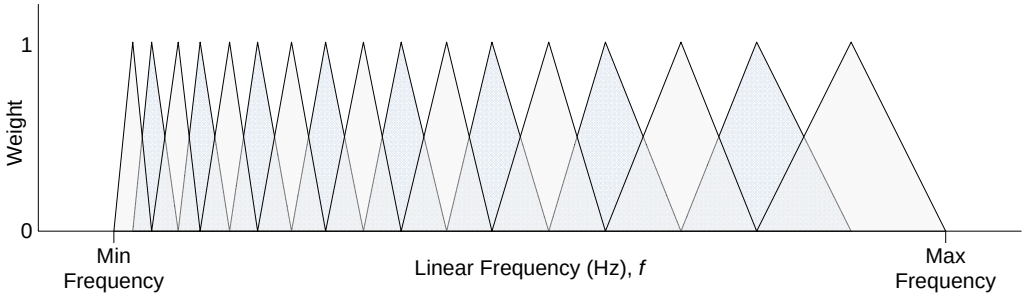


Figure 3.3: Illustration of the triangular Mel filterbank.

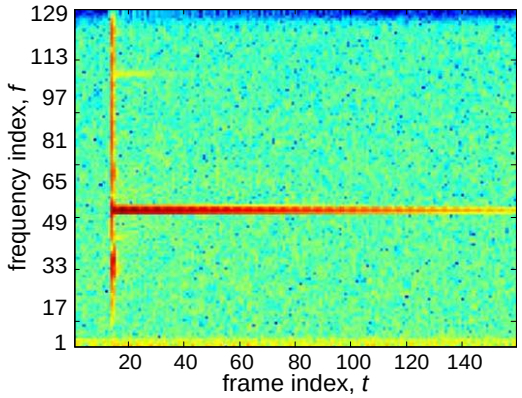
gram, more emphasis is placed on efficient signal processing and less on biological plausibility. Hence a set of triangular-shaped filters is typically used to filter the signal in the spectral domain. Filters are spaced evenly on the Mel frequency scale [128], which is calculated as follows:

$$f_{mel} = 2595 * \log\left(1 + f/700\right) \tag{3.9}$$

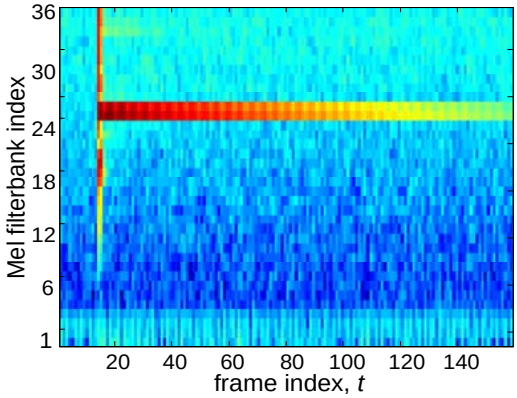
where f_{mel} is the Mel frequency for the linear frequency f. The width of each filter extends to the centre frequency of the neighbouring filters, as shown in Fig. 3.3, hence the filters become wider at higher frequencies. Log is also commonly taken to give an equivalent spectrogram representation, $S_{mel}(f, t)$, where f represents the centre frequencies of the Mel filters and t is the time frame of the STFT.

The Mel-frequency spectral coefficients (MFSCs) can also form the basis for an MFCC image representation of the sound event. This is formed by taking the DCT of the spectral coefficients across each time frame. The MFCC image isn't strictly a time-frequency spectrogram representation, since one axis represents the DCT coefficients rather than frequency. However, it is included here as it can still be considered as a sound event image.

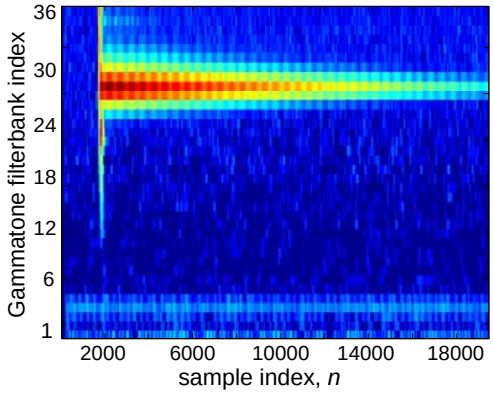
Wavelet Scalogram This approach is often considered an alternative to the STFT for describing the spectral information in an audio signal over time. It it performed by filtering the signal with a wavelet function that is based on a predefined "mother wavelet". This is then scaled and shifted over time to provide an analysis of the underlying audio signal. The wavelet transform can be written



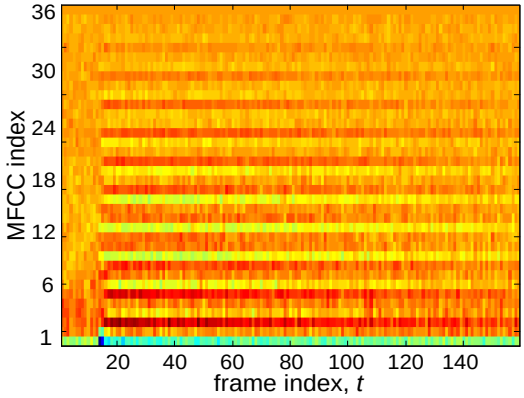
(a) Log power STFT



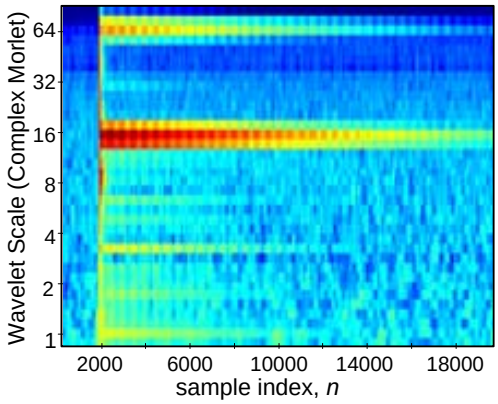
(b) Log power MFSC



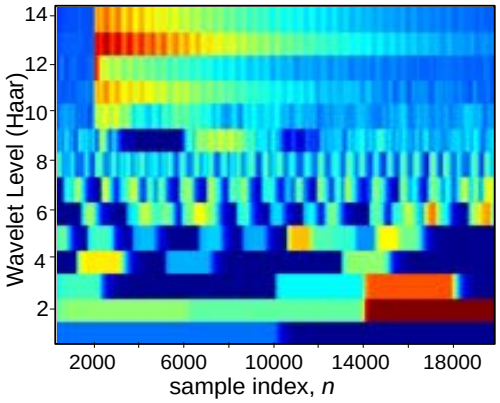
(c) Log power Gammatone



(d) MFCC Image



(e) Continuous Wavelet Scalogram



(f) Discrete Wavelet Scalogram

Figure 3.4: Example showing a bell sound with six different sound event image representations.

as follows [163]:

$$X(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \overline{\Psi\left(\frac{t-b}{a}\right)} x(t)\, dt \tag{3.10}$$

where Ψ represents the mother wavelet, a is the scaling factor and b represents the time shift factor. This is comparable to the STFT, since different values of a represents different frequencies in the signal, while b represents the time at which they occurred.

There are two methods of calculating the scalogram, by using either the continuous or discrete wavelet transform. The continuous wavelet transform (CWT) allows any valid combination of variables a, b to be used in (3.10). This requires a large number of convolution operations to be performed, hence the CWT has the disadvantage that it requires considerably more computation compared to the STFT. On the other hand, the discrete wavelet transform (DWT) only enables certain combinations of a, b to be used. This enables a more efficient implementation of the wavelet transform to be used, which repeatedly down-samples the signal by a factor of two at each level. The drawback is that it results in a trade-off between time and frequency resolution, where low frequencies have good frequency resolution but poor time resolution and vice versa. Therefore, the DWT is arguably better suited for use in signal reconstruction as opposed to spectrogram visualisation.

A comparison of the STFT, Gammatone, MFSC, MFCC, CWT and DWT sound event image representations is shown in Fig. 3.4. Although the visual signature of the bell sound can be seen in each image, the nature of each image varies significantly. In terms of the representation of the frequency information, the STFT in Fig. 3.4a provides the highest frequency resolution. On the other hand, the MFSC, gammatone and wavelet scalogram images all have similar non-linear frequency axes that emphasise the lower frequencies and compress the higher frequency information. However, the representation of the bell's harmonic is quite different in each of the representations. In particular for the gammatone and CWT scalogram, the shape is determined by the correlation between the gammatone or wavelet function respectively. This produces different results compared to the sinusoidal decomposition used in the STFT. Also, it can be seen that the MFCC image in Fig. 3.4d does not display the frequency structure

of the bell clearly in the image. This is due to the DCT, which mixes up the frequency components producing a number of horizontal lines that all correspond to the bell's harmonic. In addition, it can be seen that the temporal resolution varies, with three images based on the windowed STFT, and the other three are based on filtering in the time domain. In particular, the gammatone and CWT scalogram images provide the highest time resolution. Also, it is notable that the temporal resolution of the DWT in Fig. 3.4f is linked to the frequency resolution, making it more difficult to visualise the sound event clearly.

3.1.3 Spectrograms vs. Conventional Images

While the spectrogram is similar in some ways to a conventional image, it has some important characteristics that make it unique. These differences limit the direct application of the wide range of image processing techniques, since they may not be suitable for capturing the information in the spectrogram. Analysis of these differences provides the inspiration for adapting these existing techniques to design a novel system for SER. The comparisons are grouped into three categories: the image pixel information, the sound event geometry, and the challenges that must be overcome. These are discussed in detail below.

Pixel Information The word "pixels" is used here to refer to the elements of the image. In the case of the spectrogram image, the pixels represent the log power contained in each cell of the image, corresponding to the sound information occurring within a particular time and frequency bin. The stronger spectral elements belonging to a sound event thus have higher values than those with less energy. This means that the spectral power values represented by each pixel directly encode the relative importance of that pixel within the image. For example, a local maxima may be useful for characterising a sound event, while the gradient corresponds to changes in the perceived loudness of the sound.

For conventional images, each pixel corresponds to an intensity, such as a greyscale or colour value, where colour pixels are typically composed of separate red, green and blue (RGB) monochrome intensities [164]. The important difference is therefore that the pixel value does not directly relate to the relative importance of one

pixel over another. As the pixel intensity may vary according to different shade and lighting conditions, it is meaningless to try to base detection or segmentation of important regions on individual pixels in the image. Hence in image processing, the gradient of the pixel values is typically more important than the intensity itself, as gradients are more repeatable under different environmental conditions [165].

Sound Event Geometry The spectrogram of a sound event can simplistically be seen as a set of lines, modulation curves and diffuse patterns, with a background that is dependant on the particular environmental noise that is present. Examples include the distinctive lines that may represent harmonics, onsets or speech formants [106, 166], or the diffuse spectral pattern of an explosion or gunshot [167]. Due to the way sounds are generated, the patterns are of a highly stochastic nature, and therefore may vary significantly between repeated observation of the same sound event [168]. Due to this variability, the information in the extracted features is therefore typically captured through a statistical model, such as a GMM.

For conventional images, a number of similarities can be found with the geometry found in the spectrogram. For example, an image of a tree contains similar lines and diffuse patterns, such as those belonging to the trunk and leaves respectively, and has a background that varies depending on the location of the tree. However, while there may be small variations due to factors such as lighting, the geometry of individual objects in the image are physically fixed. In addition, many physical objects will have a convex geometry, meaning that they are enclosed within a relatively well defined boundary. This is unlike sound events in spectrograms, where the spectral information may be spread over a number of disconnected regions. Together, the nature of physical objects makes it possible to use deterministic structural models in image processing. For example, a face can be modelled as containing eyes, nose and a mouth, and the eye could either be present or absent depending on any occlusion [169].

Challenges Faced Although there are some similarities between the challenges faced in conventional and spectrogram image processing, there are also some notable differences. One example is the way in which noise affects spectrogram and con-

ventional images differently. For conventional image processing, the predominant sources are sensor noise and movement blurring, which causes variations in the pixel intensities. However, for audio processing, noise is present through a certain level of ambient background sound that is present in environment and mixed into the recorded sound signal. In the spectrogram, this is typically visualised as regions of diffuse noise patterns that form the background of the image. In addition, some high power noise may mask certain areas of the target sound event, as the mixing principle means that only the highest energy source is visible in each frequency bin [98]. This leads to the problem of finding areas of the spectrogram that belong to the noise rather than the signal, which is often termed "missing feature mask estimation" [103].

Another example of a challenge that must be considered differently is the detection of objects or sound events in the image. For conventional image processing, it must be possible to detect objects that have undergone substantial changes in position, rotation and scaling [170]. A range of techniques have been developed to overcome this, including the use of invariant feature representations [171,172]. For the spectrogram, the problem is simplified in some ways, since the frequency dimension is fixed. However, while this removes the problem of rotation and scaling, there still exists the significant problems of both time shifting and time warping [168,173]. In addition, the time warping may be non-linear, unlike image processing where physical constraints often limit the scaling of object dimensions to a linear change.

3.1.4 Previous Spectrogram Image-based Approaches

A small number of previous works have found inspiration in applying techniques from image processing to the spectrogram. However, only a few apply this to the task of SER, hence the scope is expanded to include applications in both music retrieval and ASR. The approaches commonly fall into the following three categories, depending on the scope of the extracted feature. The first is global features, which typically extract a single feature to represent the whole spectrogram image. The second is frame-based features, which are designed to be similar to conventional features, such that they can be combined with regular recognition systems. The final category is local features,

where each feature represents a localised time-frequency region of the spectrogram. These are now introduced below.

Global Features The following approaches extract a single feature to represent the information in the whole spectrogram image. The most common approach in this category is to extract low-level texture information from the spectrogram. An early example of this is presented in [174], which extracts a global feature from the spectrogram for music genre classification. The method uses a "texture-of-texture" approach, which recursively filters the spectrogram and sums over the whole image to measure the degree of correlation with each filter. An alternative approach to capture the texture information is to generate the grey-level co-occurrence matrix of the spectrogram image, as in [175]. This measures the relative frequency with which two nearby pixels, each having the same intensity value, appear within the image. Low-level features, such as energy, contrast, and homogeneity, are then extracted from the co-occurrence matrix to provide a compact representation of the spectrogram image. This approach is applied to an SER task to recognise both sports and gunshot sounds, with classification performed using a neural network.

An alternative approach is to extract higher-level information from the spectrogram, for example to represent the peaks and lines in the image. One approach, presented in [176], is based on using the straight-line Hough transform for word spotting [177]. The idea is to convert the lines in the spectrogram into a set of binary seams that capture the important speech information. These are extracted using an energy function, combined with a dynamic programming technique to maximise the energy along the seam. The subsequent binary seam pattern is transformed into the Hough space using the straight-line transform, and a subsection of the Hough space is then used for classification.

The disadvantage of these global approaches is that they cannot easily be combined with traditional frame-based recognition systems. Therefore, the next category of methods extract frame-based features, which can be used to replace or complement traditional audio features.

Frame-based Features These approaches extract a feature to represent each time frame of the spectrogram image. However, they often utilise the information over

a longer segment of the spectrogram, to incorporate some temporal information in the feature. Examples include the approach in [178] for music identification, which uses a set of Haar wavelet-like filters, originally developed by Viola and Jones for image processing [179]. Due to the large candidate filter set, a robust subset is selected using the Adaboost framework to provide a compact audio description. An improved approach is presented in [180], which uses a two-dimensional wavelet analysis on each segment, as opposed to selecting a set of pre-defined features. Due to the compressive properties of the wavelet transform, only the top wavelets are required to generate a compact, robust representation for each frame [181]. A different approach, in [182], uses a popular image processing feature for the task of ASR. Here, a histogram of oriented gradients (HOG) feature is extracted to represent each frame of the spectrogram. This is shown to give an improved performance in noisy conditions over conventional MFCCs.

An alternative frame-based approach is to use non-negative matrix factorisation (NMF) to generate features from the basis encoding. This is similar to the subspace decomposition approaches from image processing, such as the "eigenface" method for face identification. An example of the NMF approach is found in [183], which uses basis selection combined with a robust method for inferring the NMF encoding variables. A more recent approach is proposed in [3], which extracts seven novel features from the NMF decomposition. These features aim to capture characteristics such as sparsity and discontinuities, which represent important time-frequency structures. In combination with MFCCs, the proposed features are shown to produce a good performance on a large database of environmental sounds.

Local Features This group of methods extracts regions of localised time-frequency information from the spectrogram. This has the advantage, as in image processing, that local features may be less susceptible to noise and occlusion than the global feature methods. The methods typically fall into two categories: ordered and unordered, depending on whether they take into account the temporal information contained in the ordering of the features.

When temporal ordering is not enforced, the local feature methods are often termed "bag-of-visual words" (BOVW) methods. This is where a sparse vector

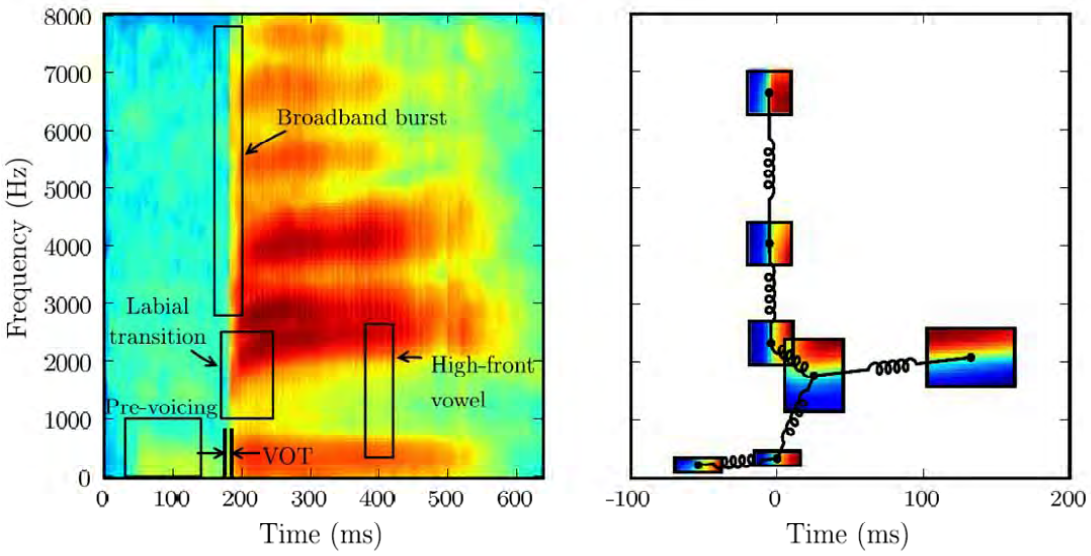


Figure 3.5: Overview of parts-based modelling for speech recognition, inspired by approaches developed for image processing (reproduced from [173]).

containing the counts of the local image features (visual words) is calculated for the image. For example, the popular HOG features have been used for the tasks of both music genre classification [184] and SER [185], with histograms of descriptors calculated and classified using SVM. Another BOVW approach, in [1], uses the passive-aggressive model for image retrieval (PAMIR) method with features extracted from a "stabilised auditory image" (SAI) which is generated using an underlying auditory model [139]. This outperformed conventional GMM and SVM classifiers, and showed a significant advantage over traditional MFCC features. Alternatively, [186] treats the spectrogram as a texture image, and uses a random sampling of the image to extract features from image blocks to capture the local time-frequency structures. The minimum matching energy of each extracted block, scanned over the whole spectrogram, is then used as the feature for classification.

When the temporal information in the local features is utilised, it captures information about the geometry of the underlying object structure. For example, in [187], Viola and Jones features are used for ASR, where both the subset of binary features, and their time-frequency location, are selected using the Adaboost

algorithm. A similar approach is used in [188], although here the features are randomly selected during training, and a "pooling" approach is used during classification to find the best activation of each feature within a local time-frequency window. Other approaches are more strongly linked with the geometry of the sound information. For example, pairs of adjacent keypoints are selected in [189] to form a geometrical "hash" that can be reliably repeated during testing. This is shown to work well for music identification, and also well-structured sound events such as alert sounds. However, the approach does not work well for organic sounds, where the spectral and temporal structure may vary between repeated occurrences [190]. Finally, a parts-based approach is presented in [173], which uses a deformable template of local pattern detectors to extract the fundamental aspects of speech. An example of this is shown in Fig. 3.5 for the spoken letter "B", where the hand-labelled speech cues can be represented as a flexible collection of local features.

Many of the above methods for extracting features from the spectrogram image have been inspired by previous work in the field of image processing. Therefore, the next section provides a detailed review of the most relevant methods that have been used for both image classification and object detection. This provides a broader insight into the available techniques, to help understand which methods may be best suited for the task of spectrogram image processing in this thesis.

3.2 Review of Image Processing Methods

The previous section provided the motivation for representing sound events through their spectrogram image, and introduced the idea of performing SER using spectrogram image processing techniques. In this section, a review of the relevant methods from the image processing domain is presented. The aim is to provide a better understanding of the types of methods that are available, and which may be best suited to the problem of SER. Together, this provides the foundation for the subsequent methods that are proposed in this thesis.

The available techniques in image processing are broken down into the following three categories: content-based image retrieval, feature-based methods and appearance-

based methods. This is shown in the diagram in Fig. 3.6, which additionally lists some of the most popular techniques in each area. Note that there are also previous works on geometry-based methods for object detection, typically using active shape models or geometric primitives such as boxes, spheres, etc [191, 192]. However, these are not included here, as they are largely considered obsolete in the face of recent state-of-the-art methods [193].

3.2.1 Content-based Image Retrieval

This is the problem of organising or categorising images in a large database, which can subsequently be searched using keywords or a query image to find similar matching images [194]. In general therefore, the task of content-based image retrieval (CBIR) is to determine the similarity between two images. The aim is to replace the need for manual annotation of images, since this becomes impossible on the ever-increasing volume of image data available.

The most common solution to the problem of CBIR is to extract low-level features that represent either the whole image or particular sub-regions of the image. The idea is to capture the fundamental information in the image, such as colour, texture or shape, which together can characterise the content of the image. For the purpose of image retrieval, the extracted features are first organised into a single feature vector to represent the query image. Then, the similarity between images stored in the database can be calculated by measuring the distance between their corresponding features [195]. The choice of distance measure depends on finding a method that best captures the similarity between the specific features. For example, colour histograms are commonly compared using histogram difference [196], histogram intersection [172], or earth-movers distance [197]. Note that many of these CBIR features are contained in the MPEG-7 standard [198], which contains descriptors for a variety of low-level visual features.

The most basic features that can be extracted from the image are descriptors for the colour content of the image. It is worth noting that colour information is generally preferred over grey-level intensities, due to the advantage of increased illumination invariance and discriminative power [199]. Common features for capturing colour include the colour moments [194], colour histogram [172], and colour layout features.

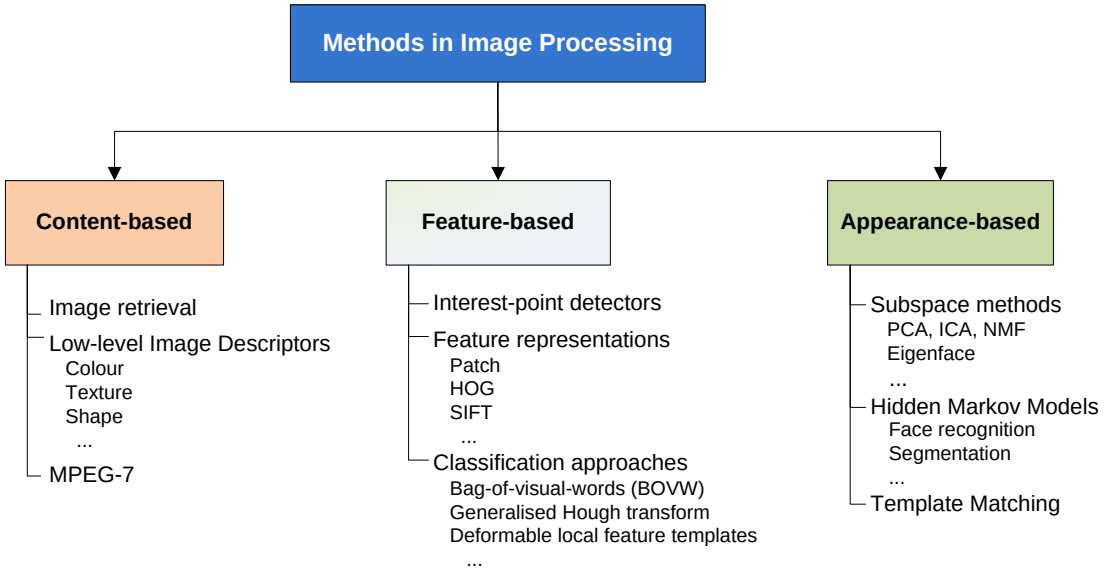


Figure 3.6: Overview of the different methods in image processing.

For example, the colour layout feature partitions the image into an 8×8 grid, and summarises the colour in each segment by taking the mean across each dimension of the colour space [198]. Another set of features aims to characterise the texture information, such as the coarseness, contrast, and directionality of the image pixel information [195, 200]. Texture information can also be extracted through a Fourier [201] or Wavelet [202] decomposition of the image, while time-varying texture from a sequence of images may be captured using dynamic textures [203]. The final set of features are shape descriptors, where the underlying shapes are typically characterised either through the shape of the bounding contour or the region contained within [204]. For example, shape regions can be characterised simply through their area, eccentricity, and orientation [205], or in a more complex way by projecting the pixel distribution onto a set of basis functions [204].

3.2.2 Feature-based Methods

This group of methods typically focuses on the problem of detecting and localising trained objects in real-world images of cluttered scenes [206]. This is different from CBIR in several ways. Firstly, only a particular sub-region of the image conveys information about the desired object, whereas the rest of the image may be clutter. And

secondly, information such as the number of objects, and both their size and position in the image, is unknown. Therefore, it is no longer the case of simply measuring the distance between the query image and the database, as a detection step is first required to localise the objects in the query image. As highlighted in Fig. 3.6, the solution is to extract local features from the image by first detecting interest-points, or "keypoints", on the object. These are intended to localise repeatable points on an object, and must also be robust to scale and rotation. Classification is then typically performed using either a sliding-window, Hough transform, or bag-of-visual-words (BOVW) approaches, as shown in Fig. 3.7 and discussed below.

The first and most important aspect of these methods is to extract reliable local features that represent the image information consistently. One factor is that the detected keypoints must be repeatable across a range of illumination, scales, and affine transformations. Therefore, considerable research has been carried out in developing reliable detectors [165], with one of the most popular methods known as difference-of-Gaussians (DoG). This method successively smooths the input image with a Gaussian kernel, with the difference between different levels of smoothing used to localise important image gradients [207]. Another factor is that the local features must be extracted to represent the image information in a robust way. Among the many different techniques, the histogram of oriented gradients (HOG) approach is often the most popular [208–211]. This feature counts the image gradient orientations in localised regions of the image, hence is relatively robust to changes in lighting and colour. The HOG also forms part of the popular scale-invariant feature transform (SIFT) object detection approach [171], which extracts a local HOG feature in the regions surrounding the keypoint. Many enhancements to the SIFT approach have been proposed [212], including principal component analysis (PCA-SIFT) and speeded-up robust features (SURF), which offer improvements in illumination invariance and speed respectively [213].

The SIFT and many other approaches, use the generalised Hough transform (GHT) for object detection. This is an extension of the original Hough transform, which had the limitation that it could only be used for parametrised shapes such as lines, circle, etc [177]. While the formulation of the GHT is somewhat different, it still uses a bottom-up voting scheme similar to that of the original transform. In this way, the extracted local features cast votes for possible object hypotheses in the Hough accumulator space [215], with local maxima corresponding to potential object hypotheses. Hypotheses that are

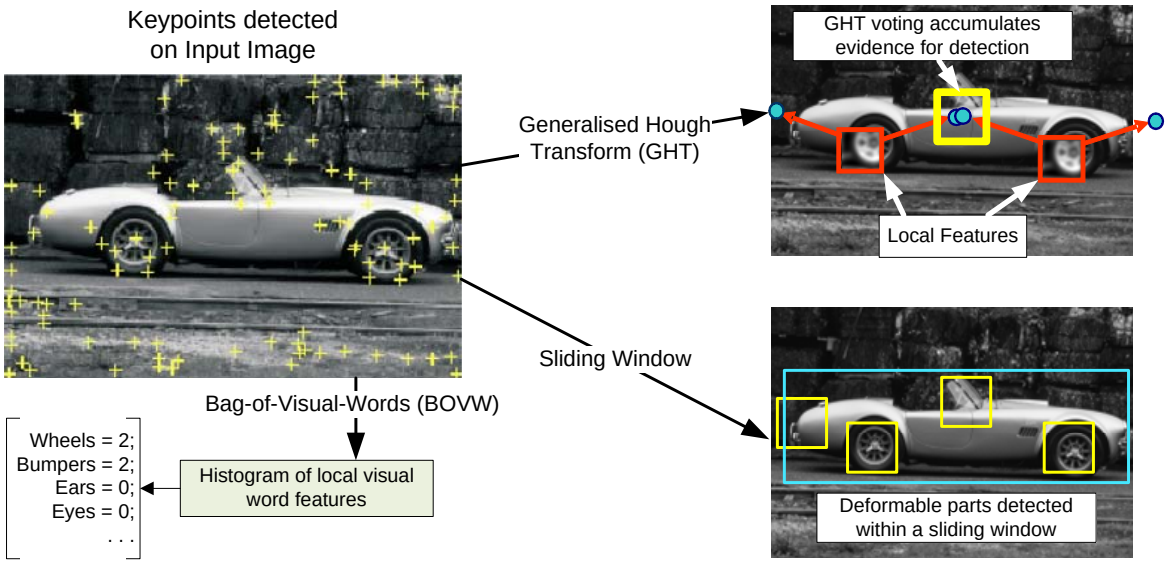


Figure 3.7: Examples of three different paradigms for feature-based object detection. The GHT approach is taken from [206], then sliding-window approach is taken from [214], while the BOVW approach is from [208].

consistent with an object detected in the image will receive many votes, hence the sum of votes defines the final object score. In the SIFT approach, the Hough accumulator space is discretised into broad bins, which are then searched for maxima, and an affine transformation is subsequently used to confirm object hypotheses [207]. In [216], contour fragments are used as local features in the GHT, giving the added advantage that an accurate shape boundary segmentation can be achieved [217]. A more recent approach, called the Implicit Shape Model (ISM) [206, 218], introduces probabilistic Hough voting, and also avoids discretising the search space by finding maxima using mean-shift mode clustering. An additional step can also be taken to place the Hough transform in a discriminative framework, where the aim is to learn weights for the probabilistic ISM voting that increase the separability of object detections [219].

An alternative paradigm for object detection uses a sliding-window to examine each windowed region of the image, with a binary classifier, such as SVM, is used to detect whether or not the target object exists in the window [220]. This approach forms the basis for a number of popular systems, including the AdaBoost system of Viola and Jones [179]. This method uses a large set of simple Haar-like difference features that can be rapidly calculated using their "integral image" approach. It has also

been shown that this approach can be enhanced by sharing common features between multiple object classifiers [221]. In [222], the sub-window is partitioned into small fixed-size regions and a one-dimensional HOG feature is extracted to characterise the whole window. An alternative approach is to model objects as a subset of parts in a deformable configuration within the window [223,224]. For example, in [214], template parts are modelled using HOG features, and a multi-scale search is performed using SVM for classification. However, compared to the GHT, the sliding-window paradigm is arguably a less natural approach, since humans do not appear to scan an image exhaustively to detect objects. In addition, it is less efficient to exhaustively search every sub-window of the image at different scales and rotation, and it may also lead to a large number of false detections.

While the above approaches have dealt with object detection, feature-based methods are also applied for image classification. This is referred to as the bag-of-visual-words (BOVW) approach, particularly when the spatial ordering of the local features is not rigidly enforced [209]. The most common BOVW feature is simply a histogram of the number of occurrences of particular local feature clusters within a given image [208]. Then for classification, SVM or naïve Bayes are often used, where SVM has the advantage that many different kernels can be developed, such as using the earth-movers distance [225] or pyramid histogram intersection kernel [226]. An alternative classification approach is the passive-aggressive model for image retrieval (PAMIR) [227], which has a fast and robust training procedure that uses a ranking-based cost function to minimise the loss related to the ranking performance of the model. This allows PAMIR to efficiently learn a linear mapping from a sparse BOVW feature space to a large query space. More recent BOVW approaches have suggested that incorporating some spatial information is beneficial to the overall image classification performance. For example, in [210, 211], the two-dimensional image space is partitioned into a sequence of increasingly coarser grids, and then a weighted sum over the number of matches that occur at each level of resolution is used as the feature. It was also recently shown that performing clustering of local feature patterns to form the visual-word codebook may be detrimental. For example, in [228], a naïve Bayes nearest neighbour classifier is able to achieve state-of-the-art image classification performance, while requiring no learning or quantisation.

3.2.3 Appearance-based Methods

This group of methods are commonly used for both object recognition and image classification, and utilise only the appearance of the pixels in the image, as opposed to shape or contour models. While this has similarities with the feature-based approaches, the methods presented here typically interact directly with the pixel information, rather than extracting features to characterise a region. Two different approaches are considered here. The first is based on modelling the pixel distribution and performing classification with hidden Markov models (HMMs), while the second are subspace methods, which decompose the image into a set of constituent bases.

Approaches that use the HMM technique have been previously used for face classification. Here, the two-dimensional image is transformed into a sequence of one-dimensional feature vectors, and the statistical properties of the sequence are captured in an HMM [229]. Analogous to similar systems in SER, the output of each state of the HMM is modelled using a multi-variate GMM. The features can then be either the raw image pixels [229], two-dimensional DCT coefficients [230], or wavelet decomposition coefficients [231]. The idea is that the face can be divided into a small number of distinct regions, such as eyes, nose, mouth and chin, each of which corresponds to one state in the HMM. Since the face is largely symmetrical, a horizontal strip can be used as the observation vector, with the HMM state sequence proceeding from the top to the bottom of the image. An extension to the above technique is to use a pseudo two-dimensional HMM. Here, the observation densities of the vertical sequence now become HMM super-states, and each row is now modelled by a one-dimensional HMM [232]. Although this is more complex than the simple one-dimensional HMM, it is more appropriate for two-dimensional data as it can capture more variation between images [233].

The other common approach is based on a subspace decomposition of the image pixel data. The idea is that the image data can be projected onto a small number of orthogonal basis images, to provide a compact representation of the image. Then, the original image can be reconstructed through a linear combination of the basis images [234]. Common methods used to perform the decomposition include principal component analysis (PCA), independent component analysis (ICA) or non-negative matrix factorisation (NMF) [170]. The subspace approach first became popular with

the introduction of the "eigenface" technique in [235], where the eigenvectors of the PCA decomposition are used to define the subspace of face images called the "face space". Faces could then be detected within an image by measuring the distance between each sliding window input and the reconstructed face space [236]. Non-face images would be poorly reconstructed using the face space basis images, hence could be easily rejected. However, the method is not limited to faces, and has more recently been applied to both pedestrian [237] and generic object [238, 239] detection tasks, among others. The drawbacks of the subspace approach are its limited robustness to both noise, occlusion and cluttered background [240]. This is because the subspace is a decomposition of the global image data, unlike localised features that have a reduced probability of disruption. One solution to overcome this is to introduce a robust hypothesise-and-test paradigm using only a subset of the image pixels [241, 242]. Another solution is to constrain the NMF decomposition to produce sparse basis vectors that represent more localised image information [243].

Overall, each of these approaches has advantages and disadvantages depending on the type of images used in each application, for example faces, objects, scenery, etc. Therefore, careful consideration is required when applying these techniques to the domain of spectrogram image processing. In the next section, an initial approach is introduced based on a content-based feature extraction from the spectrogram. This captures the visual signature of the sound event in a way that takes account of the stochastic nature of the sound, and can perform well in mismatched conditions.

3.3 Spectrogram Image Feature for Robust Sound Event Classification

Inspired by the idea of using visual techniques from image processing for SER, a novel feature extraction method for sound event classification is now presented. The motivation stems from the fact that spectrograms form recognisable images that can be identified by a human reader, forming the basis for the idea of spectrogram image processing. The proposed approach is called the spectrogram image feature (SIF) [5], which extracts a visual signature from the sound's time-frequency representation. The idea of the SIF is to capture the stochastic nature of the spectrogram image

content, while producing a feature that is robust to mismatched noise. To achieve this, the method first quantises the dynamic range of the spectrogram into regions, corresponding to the low, medium and high power spectral information in the sound event. This process is analogous to pseudo-colouration in image processing, which is commonly used to enhance the characteristic image information for human perception. However, for machine classification it is found to improve the discrimination of the extracted feature, by allowing higher weights to be assigned to the most reliable, high power spectral regions. The rest of this section first gives an overview of the approach, before describing in detail the image feature extraction and classification approach.

3.3.1 Overview

The idea behind the SIF is that the spectrogram naturally represents the joint spectro-temporal information contained in the sound event signal, unlike frame-based features that capture only a slice of frequency information at a given time. Therefore, it is possible to extract a global image feature from the spectrogram to characterise the sound event information for classification. This feature forms a distinct signature for each sound event, and the distance between image features provides a metric for classification.

Inspiration for the feature extraction can be found from previous work in the field of content-based image retrieval (CBIR), since this is the similar task of extracting a robust feature to characterise the low-level pixel information. A popular CBIR approach is to characterise the colour information in the image through the distribution of image pixel intensity [172, 194]. This is preferred over grey-level intensities as it provides increased illumination invariance and discriminative power for classification [199]. It is also beneficial to capture the distribution information in local regions of the image, since similar images will often have a consistent layout, such as the sky at the top with buildings below. One approach for this is the colour layout descriptor included in the MPEG-7 toolkit [198]. This partitions the image into local blocks, and captures the colour information through the mean of the colour in each block [198].

The spectrogram can easily be normalised into a grey-scale image, by scaling the dynamic range of the spectral information into the $[0, 1]$ range. Also, due to the stochastic nature of the sound events, characterising the spectrogram through the im-

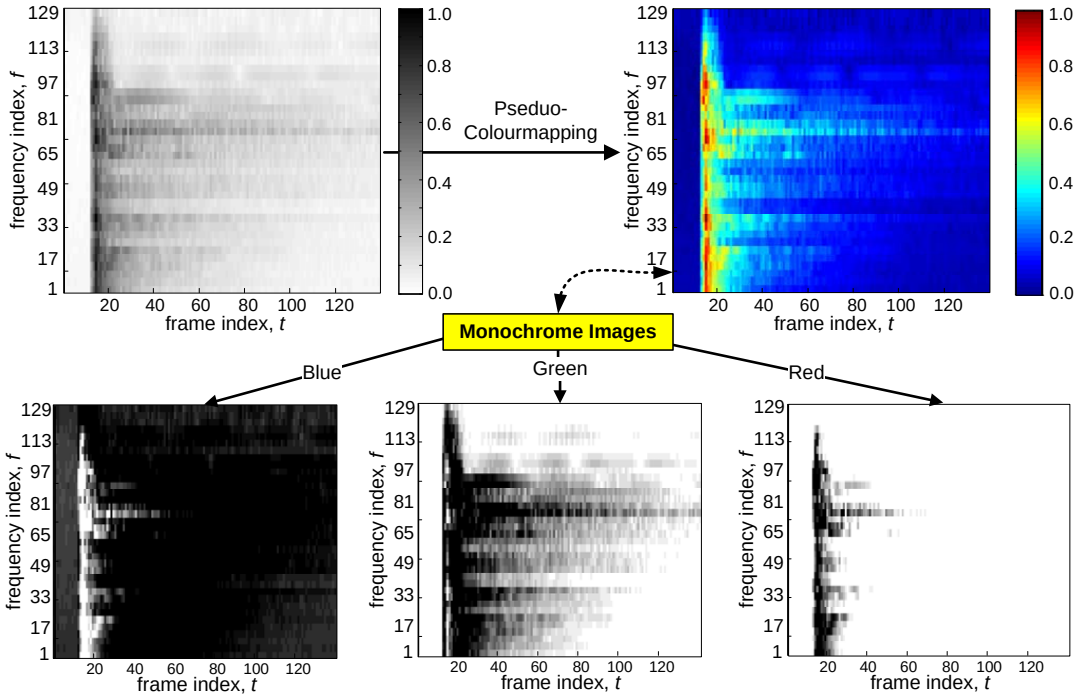


Figure 3.8: Overview of pseudo-colourmapping. The grey-scale spectrogram in the top left has been transformed to a colour representation using the "Jet" colourmap shown in Fig. 3.10a. This can then be broken down into the three RGB monochrome images shown, each of which represents the intensity of a single colour in the image.

age pixel distribution should provide a reliable way of capturing the sound information. It is also important to capture the sound event information in local time-frequency regions, rather than the global statistics over the whole spectrogram. Therefore, the colour layout approach could be applied to the grey-scale intensities to characterise the distribution of spectral information in the image pixels.

However, a further step is proposed here to enhance the discriminative ability of the SIF feature. The motivation comes from the pseudo-colourmapping procedure in image processing, which is used to enhance the visual perception of grey-scale images by increasing the contrast between regions of low and high intensity. A similar process can be applied to the spectrogram, which quantises the dynamic range into regions representing the low, medium and high power spectral information. The idea is to utilise the fact that the energy of many sound events is often concentrated in a limited number of time-frequency regions, due to the sparse nature of the spectral information.

Then, the robust high-energy peaks of the sound event will be quantised into a separate region of the dynamic range and mapped to different feature dimensions from the low-energy background noise. These can then be assigned a higher weighting by a discriminative classifier such as SVM.

An example of the mapping is shown in Fig. 3.8. Here, the grey-scale spectrogram is pseudo-coloured to enhance the most important information. The colour image is composed of the three RGB monochrome images as shown, where each can be seen to represent information extracted from a different region of the dynamic range. In particular, the "red" quantisation captures only the most important high-power spectral information, and should be the most robust in mismatched noise conditions. Together, the signal processing in the SIF can be summarised as follows:

1. The spectrogram is first normalised into a grey-scale image, with the dynamic range of the spectral information adjusted to a value in the range $[0, 1]$.

2. A process analogous to pseudo-colouration is then used to quantise the dynamic range into regions representing the low, medium and high power spectral information. Each region is then mapped to form a monochrome image representing each "colour" region of the dynamic range.

3. Finally, the SIF is formed by capturing the layout and distribution statistics of each monochrome image in a similar way to the colour layout feature. This requires partitioning each image into blocks, then extracting the pixel distribution statistics from each block to form the feature.

An overview of this process is shown graphically in Fig. 3.9, with a comparison made to the processing steps for conventional MFCC features. Classification is performed using SVM, which is preferred over other approaches such as kNN, since SVM assigns higher weights to the most discriminative components of the feature and therefore should lead to a more robust classification.

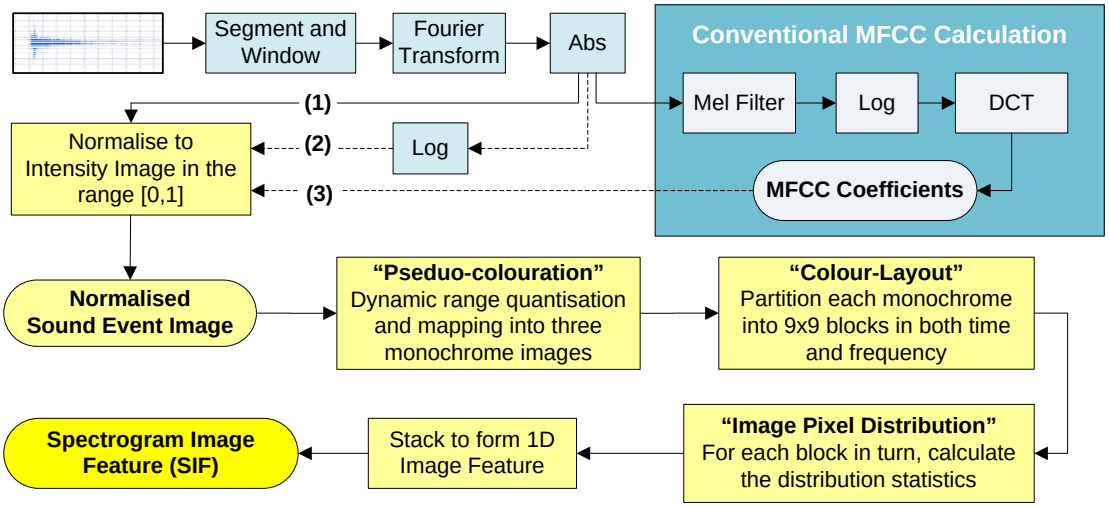


Figure 3.9: Overview of the SIF feature extraction algorithm, with conventional MFCC features as comparison. For the normalised spectrogram image, either (1) the linear spectrogram, (2) the log spectrogram, or (3) the MFCC cepstrogram is used as the input image.

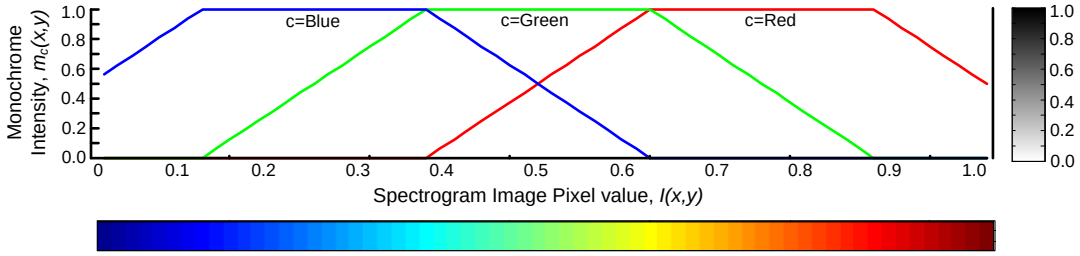
3.3.2 Image Feature Extraction

Starting from the sound event spectrogram, $S(f,t)$, the matrix is first normalised into a greyscale intensity image, with the range scaled between $[0,1]$ as follows:

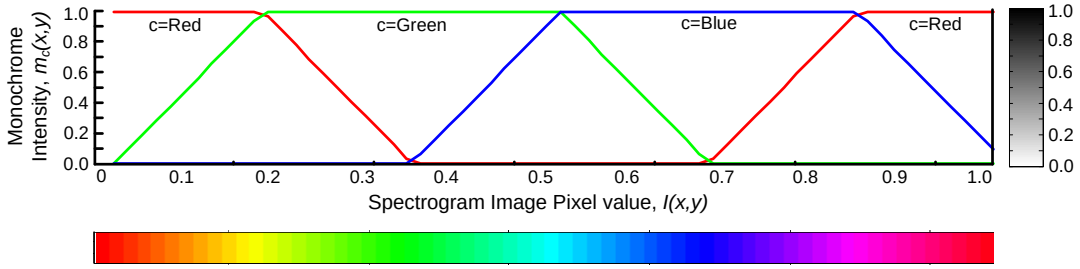
$$I(x=f,y=t)=\frac{S(f,t)-\min(S)}{\max(S)-\min(S)}. \tag{3.11}$$

where x,y are the vertical and horizontal indices of the image respectively, and $I(x,y)$ is a two-dimensional image representation of the sound event of size $X\times Y$. The spectrogram representation, $S(f,t)$, can be either the linear spectrogram, $S_{lin}(f,t)$ from (3.1), the log spectogram, $S_{log}(f,t)$ from (3.2), or the MFCC cepstrogram.

The dynamic range of the greyscale sound event image, $I(x,y)$, is now quantised into separate regions that are then mapped into a higher dimensional space. Examples of two common pseudo-colour mapping functions from image processing are shown in Fig. 3.10. These map each value in the input image to three (RGB) monochrome intensity values, representing information from a particular region of the dynamic range. The mapping function is referred to here as q_c, with the resulting monochromes



(a) The "Jet" mapping function, which quantises the image into three (RGB) monochrome values.



(b) The "HSV" mapping function. Note that the "red" mapping captures both the high and low energy values together.

Figure 3.10: Examples of common colourmaps from image processing. Here, each greyscale image input value, $I(x, y)$, is mapped to three (RGB) monochrome intensity values.

written as follows:

$$m_c(x, y) = q_c(I(x, y)) \quad \forall c \in (c_1, c_2, \ldots c_N) \tag{3.12}$$

where c represents the quantisation region of the dynamic range, and the output, m_c, are a set of monochrome images, each of which is the same size as the input image. Each quantised image is referred to here as a "monochrome" image, following the common definition in image processing meaning that it represents the intensity of single "colour" from a quantised region of the dynamic range.

The mapping operation can be seen as a generalisation of the pseudo-colourmapping procedure from image processing, since the quantisation is not limited to the three colours required in the colourmap for image processing. However, for the SIF, it was found from initial experiments that three was a good trade-off between the accuracy and computational cost, hence the three-level system is employed in the later experiments. Therefore, several standard pseudo-colour mapping functions are considered

here that are common in image processing. These are called "HSV" and "Jet", which can both be formally defined as follows:

$$q_c(I(x,y)) = \begin{cases} \frac{I(x,y)-l_1}{l_2-l_1}, & \text{for } l_1 < I(x,y) < l_2 \\ 1, & \text{for } l_2 \leq I(x,y) \leq u_1 \\ \frac{u_2-I(x,y)}{u_2-u_1}, & \text{for } u_1 < I(x,y) < u_2 \\ 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

where the parameter set $\{l_1, l_2, u_1, u_2\}$ defines the precise quantisations for each of the colours in the mapping. For the Jet mapping in Fig. 3.10a, the parameters are $c_{red} = \{\frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \frac{9}{8}\}$, $c_{green} = \{\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}\}$, and $c_{blue} = \{-\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}\}$. While for the HSV mapping in Fig. 3.10b, the parameters are $c_{red} = \{\frac{2}{3}, \frac{85}{100}, \frac{1}{5}, \frac{1}{3}\}$, $c_{green} = \{0, \frac{1}{5}, \frac{1}{2}, \frac{2}{3}\}$, and $c_{blue} = \{\frac{1}{3}, \frac{1}{2}, \frac{85}{100}, 1\}$.

An image feature can now be extracted to characterise the information in each of the monochrome images. The proposed feature is similar to the colour layout feature in image processing [198], expect that the central moments are used to better capture the local distribution of image pixel information. Each monochrome is first partitioned into two-dimensional local blocks, D_x, D_y, giving a total of $D_x \times D_y$ blocks, as shown in Fig. 3.11. Each block is therefore of size $(\frac{X}{D_x}, \frac{Y}{D_y})$. The pixel distribution information, $x_{i,j}$, is then extracted from each local block from each of the monochrome images. This is characterised here by the central moments of the distribution as follows, dropping the c notation for clarity:

$$x_{i,j}^{(k)} = \begin{cases} E[L_{i,j}], & \text{for } k = 1 \\ E\left[(L_{i,j} - E[L_{i,j}])^k\right], & \text{for } k = 2, 3, \ldots \end{cases} \tag{3.14}$$

where $L_{i,j}$ are the pixels in the local block $m_{i,j}(x,y)$ as shown in Fig. 3.11, E is the expectation operator, and for $k > 1$ the feature represents the k^{th} moment about the mean. Together, $x_{i,j}$, characterises the distribution of the monochrome image pixels in each block, such that both the time-frequency and dynamic range information is captured. This therefore is used as the final image feature, where the indices are concatenated to form a single high-dimension feature vector.

In preliminary experiments on the SIF, the second and third central moments, $k =$

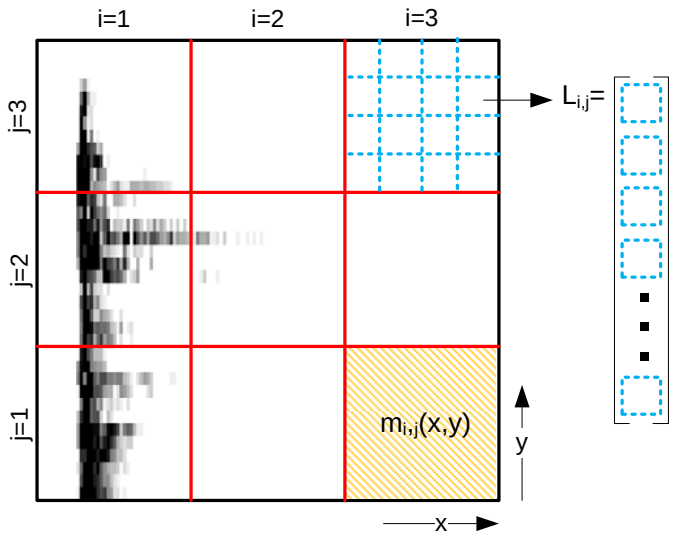


Figure 3.11: Schematic of the image feature partitioning used in the SIF with $D_x = D_y = 3$. The underlying monochrome is partitioned into $D_x \times D_y$ blocks with indices i, j, with each block referred to as $m_{i,j}(x, y)$. The pixels of each local block are also extracted into a vector, $L_{i,j}$.

$\{2, 3\}$, were found to produce a good overall performance, hence are used throughout. However, 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. This may be caused by the increasing noise energy causing a shift in the distribution statistics, while the shape of the distribution is less affected. In addition, it was found that partitioning each image dimension into $D_x = D_y = 9$ blocks gives a good tradeoff between performance and feature vector size. It should be noted that overlapping blocks could be used, however it was found that this did not significantly improve the results. Overall, the final SIF is a 486 (2x3x9x9) dimension vector, with three quantisation regions ($c \in \{c_{red}, c_{green}, c_{blue}\}$) and two central moments ($k \in \{2, 3\}$) to capture the distribution statistics.

3.3.3 Classification Approach

Several different classification methods were considered that are common in image processing, including kNN and SVM. In the following experiments, SVM is chosen, as it provides more discrimination through optimisation, and is an efficient encoding of

the class separation problem. It also provides a useful comparison to previous works which have also utilised SVM for sound event classification [38, 176, 185]. A linear SVM classifier is used [96], and the One-Against-One (OAO) binary configuration is employed to solve the multi-class problem, with the max-votes-wins voting strategy for classification. Note that the conventional non-linear Gaussian kernel was also tested and achieved a similar classification accuracy to linear SVM. However, since the Gaussian kernel is not well suited to high dimensional features, and has an increased computational cost, linear SVM is preferred and is used throughout.

Also, it should be noted that if the proposed feature extraction method is seen as a non-linear transform $\phi(x)$ from sample x, then the method can be considered as a novel SVM kernel, where $K(x_i, x_j) = \phi(x_j)^T \phi(x_i)$.

3.4 Experiments

In this section, experiments are conducted to compare the performance of the SIF method with a range of baseline methods for sound event recognition, including the baseline results established in Section 2.3. In addition to this, several recent SER methods that also draw inspiration from the image processing field are implemented, to provide a comparison between other methods more similar to the SIF. The methods are tested on the same standard database of environmental sounds in both clean and mismatched noise conditions, with training performed using only clean samples.

3.4.1 Experimental Setup

For comparison with the previous baseline methods, the same experimental setup is used as in Section 2.3. Hence, the same 50 sound event classes are selected from the RWCP database, and as before, 50 files are randomly selected for training and 30 for testing in each of the five runs of the experiment. The classification accuracy for the SIF is investigated in mismatched conditions, using only clean samples for training, with the average performance reported in clean and at 20, 10 and 0 dB SNR across four realistic noise conditions. The standard deviation is also reported ($\pm$) across the five runs of the experiment and the four different noise conditions.

SIF Evaluation Methods

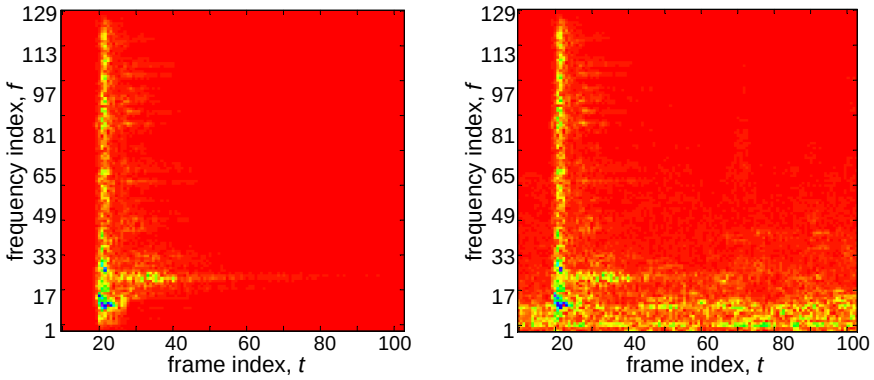
Since the SIF approach is not limited to any particular two-dimensional sound event image, three representations are explored here: the traditional log power spectrogram, the linear power spectrogram, and the cepstrogram, which is an image formed by stacking conventional frame-by-frame MFCC coefficients. This enables the experiments to investigate the following factors in generating the SIF:

1. Greyscale vs. Pseudo-Colour Quantised
2. Linear vs. Log Power Spectrogram
3. Spectral vs. Cepstral (MFCC image) representation

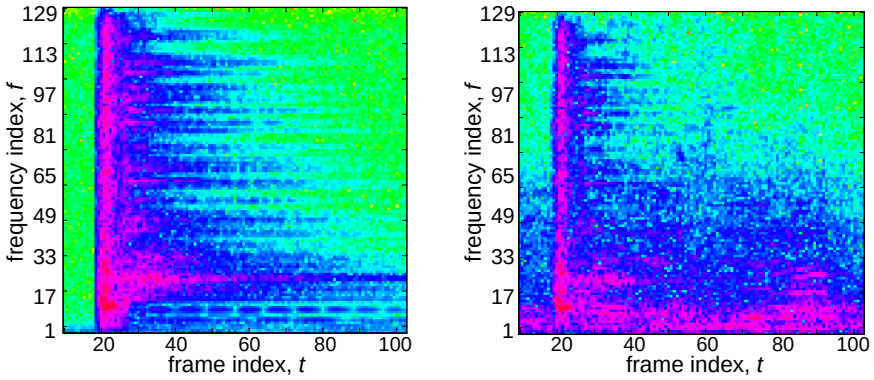
Examples of the three representations, in clean and noisy conditions, are shown in Fig. 3.12. In all cases, the same SIF extraction method is applied, as described in Section 3.3.2, using the "HSV" mapping to perform pseudo-colour quantisation. Preliminary experiments showed that the classification accuracy for the HSV mapping was marginally higher than for Jet, hence it is used throughout. This may be explained by the fact that HSV is a more discriminative mapping, with each input intensity value mapped to two colour channels as seen in Fig. 3.10b. On the other hand, the highest and lowest range of input values in Jet are mapped only to a single colour channel via a roll-off in output intensity, as seen in Fig. 3.10a, and this may reduce its discriminative ability.

Baseline Image Processing Inspired Methods

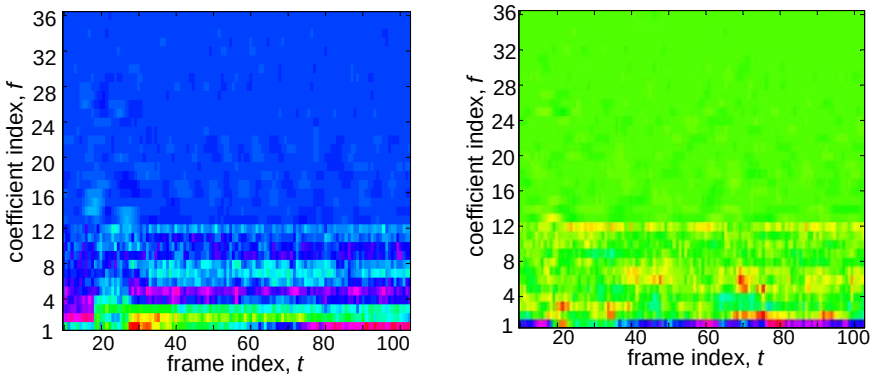
In addition to the baseline methods established in Section 2.3, a selection of previous image processing inspired methods are also implemented to provide a more complete comparison with the SIF. Since very few previous works have applied such methods to the task of SER, image processing approaches for speech and music tasks are included, as this enables a more complete comparison with a wider range of methods. Four methods are implemented to cover the categories of techniques as discussed in Section 3.1.4: global features, frame-based features, and local features, with and without temporal information. These are as follows:



(a) Linear spectrogram images



(b) Log spectrogram images



(c) Cepstrogram (MFCC) images

Figure 3.12: Examples of clean *(left)* and noisy *(right)* spectrograms from the "Cymbals" sound class, for each of the three different image representations considered for the SIF. The images have been pseudo-colourmapped using the "HSV" colourmap from Fig. 3.10b.

1. Spectrographic Seam Patterns [176]
   This is a global classification approach originally designed for word spotting. The method transforms the spectrogram into a binary image, by extracting high-energy vertical seams in the image, and then performs the straight line Hough transform to give a feature for SVM classification. The best performance was achieved using 25 seam patterns, as originally reported in [176].

2. Histogram of Oriented Gradients (HOG) Features [182]:
   This is a frame-by-frame approach originally proposed for ASR. Here, HOG features are extracted from each frame at fixed frequency locations, with the idea to capture more temporal information than delta-MFCCs. The parameters used are taken from [182], using 8 HOG descriptors per frame, each containing 32 features, and use PCA to reduce the dimension to 50 for classification using HMM.

3. SIFT Bag-of-Visual-Words (BOVW) descriptor [185]:
   This is a local feature approach designed for both music genre and sound event recognition. The method extracts 128 dimension SIFT features from each local 16×16 region on a regular 8×8 grid, and computes the codeword uncertainty for each feature to the codebook of 4096 visual words for classification using SVM.

4. Ordered Spectro-Temporal BOVW [186, 188]:
   This combines the techniques presented in [188] and [186]. Here, local time-frequency patches are extracted at random during training, and during testing the normalised minimum mean square error (MSE) is calculated, using a local pooling operator on the time-frequency location, to give a feature for classification using SVM. The implementation uses 1000 patches, extracted as in [186], but incorporates the idea in [188] of using ordered BOVW to capture temporal information, with a local pooling operator to find the minimum MSE in the local region to provide robustness during matching.

Each of these methods was implemented in Matlab, and found to give comparable performance to that reported by their authors on a similar sized dataset. For the HOG method, the same HMM configuration is used as for the other baseline experiments,

Image	Dynamic Range	Clean	20dB	10dB	0dB	Avg.
Linear	Greyscale	67.4 ± 0.4	67.4 ± 0.4	67.4 ± 0.6	61.7 ± 2.2	66.0
Log		74.9 ± 0.7	44.6 ± 7.5	21.3 ± 7.5	11.8 ± 4.2	38.2
Cepst		83.3 ± 1.1	33.0 ± 6.7	14.2 ± 3.3	6.1 ± 1.5	34.2
Linear	Colour Quantised	91.1 ± 1.0	$\mathbf{91.1 \pm 0.9}$	$\mathbf{90.7 \pm 1.0}$	$\mathbf{80.9 \pm 1.8}$	**88.5**
Log		$\mathbf{97.3 \pm 0.2}$	81.1 ± 5.5	53.5 ± 10.2	26.4 ± 8.8	64.6
Cepst		$\mathbf{97.3 \pm 0.5}$	45.7 ± 9.5	20.5 ± 4.4	6.1 ± 1.3	42.4

Table 3.1: Classification accuracy results for the spectrogram image feature (SIF) method, exploring the different sound event image representations that contribute to give the best performance.

with 5 states and 6 Gaussian mixtures, and both training and testing are carried out using HTK [161].

3.4.2 Results and Discussion

The results from the experiments on both the SIF and baseline methods are now presented. First, the important factors that contributed to the success of the SIF method are analysed: greyscale vs. colour quantisations of the dynamic range, linear vs. log power spectrogram images, and spectral vs. cepstral sound event image representations. Then, the performance of the best performing SIF method is compared to results achieved by the baseline methods.

Colour Quantised vs. Greyscale SIF

Here, the effect of the quantisation is analysed by comparing the results obtained for the greyscale and pseudo-colour quantised SIFs. It can be seen in Table 3.1 that the quantised SIF outperforms the equivalent greyscale SIF in both clean and mismatched conditions. This indicates that by mapping the dynamic range of the greyscale spectrogram into a higher dimensional space, in this case the three RGB quantisations, the separability between the 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

Sound Event	Green (Low)	Blue (Medium)	Red (High/Lowest)	Greyscale
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 3.2: Example distribution distances for greyscale and the three colour quantisations between clean and 0db noise conditions

the low-power region of the spectrogram's dynamic range. Therefore, the monochrome images mapped from the higher-power range should be largely unchanged, despite the presence of the noise. Since the discriminative components of the SIF should be assigned a higher weighting in the SVM classifier, the quantised SIF should be more robust than the greyscale SIF 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 can 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 [244] can be used as a measure:

$$H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\, e^{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}}. \tag{3.15}$$

Since the SIF feature uses the central moments to characterise the pixel distribution, this effectively mean-normalises the distribution. Hence, with $\mu_1 = \mu_2 = 0$, (3.15) simplifies to a ratio of the standard deviations. Example results are presented in Table 3.2, which show the mean distribution distances across the 9×9 SIF blocks, averaged over 50 samples using the linear power SIF. Although 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. It is suggested 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 Spectrogram Image

From the SIF results in Table 3.1, it can be seen that one important experimental outcome is the improved performance of the linear power SIF over the log power equivalent. For the colour-quantised SIF, the two representations achieved an average accuracy of 88.5% and 64.6% respectively. This is significant, since a linear representation is not commonly used in conventional audio processing systems, but performs significantly better here. It should however be noted that in clean conditions the performance of the linear power SIF is just 91.1%, compared to over 97.3% for the log power SIF. This can be explained by the differences in the linear and log power representations, as shown in Fig. 3.12. Here it can be seen that the linear representation is less affected by the noise, since the spectrogram is dominated by the sparse high-energy elements of the sound event that are an order of magnitude larger than the noise. This provides significant robustness to noise, but leads to confusion between the most similar sounds, which is reflected in the lower accuracy in clean conditions. On the other hand, the log power reduces the dynamic range of the spectrogram, revealing the detail from the low power frequencies. This provides better discrimination between sound events in clean conditions, but also causes the detail of the noise to be magnified, leading to changes in the feature that cannot be compensated for during classification. Hence, the linear power spectrogram produces a much more robust basis for the SIF extraction and classification in mismatched conditions, as the high power quantisation captures only the most characteristic sound event information.

Spectral vs. Cepstral Representations

An interesting comparison to examine is how the SIF concept applies to the "MFCC image". Table 3.1 shows that while the cepstrogram performs well in clean conditions, the accuracy falls rapidly for mismatched SNRs. This is explained by the way in which noise causes variations in the cepstrogram, especially in the lower order cepstral coefficients. This causes the cepstrogram image to appear significantly different, as can be seen in the clean and noisy examples shown in Fig. 3.12c. This is because the DCT transform mixes up the frequency components, breaking the assumption that particular frequency components will still be visible in the noisy spectrum. Therefore the performance is greatly degraded in mismatched conditions compared to that of the

Group	Method	Clean	20dB	10dB	0dB	Avg.
Base-line	ETSI-AFE	99.1 ± 0.2	89.4 ± 3.2	71.7 ± 6.1	35.4 ± 7.7	73.9
	Multi-Conditional	97.5 ± 0.1	$\mathbf{95.4 \pm 1.3}$	$\mathbf{91.9 \pm 2.7}$	67.2 ± 7.3	88.0
Image Proc.	Spec. Seams	87.0 ± 1.6	43.2 ± 7.8	27.7 ± 5.2	15.5 ± 4.0	43.4
	HOG frame-based	$\mathbf{99.2 \pm 0.1}$	68.9 ± 4.6	33.2 ± 7.2	9.3 ± 5.4	52.7
	SIFT BOVW	89.0 ± 0.5	45.4 ± 5.2	25.8 ± 4.5	14.6 ± 4.4	43.7
	Ordered BOVW	94.8 ± 0.6	63.3 ± 9.2	32.7 ± 7.7	12.8 ± 4.0	50.9
SIF (linear power, colour quantised)		91.1 ± 1.0	91.1 ± 0.9	90.7 ± 1.0	$\mathbf{80.9 \pm 1.8}$	**88.5**

Table 3.3: Experimental results comparing the classification accuracy of the best performing colour quantised SIF methods with both conventional and image processing baseline methods.

standard SIF.

SIF vs. Baseline

The results in Table 3.3 show that the system with the best overall performance is the linear power colour quantised SIF, with an average classification accuracy of 88.5%. This is a 15% improvement over the ETSI-AFE baseline method from Section 2.3, which uses noise reduction with only clean samples for training. It can also be seen that the linear power SIF even achieves a small improvement in performance over the multi-conditional MFCC-HMM baseline. The largest improvement is achieved in the most severe 0dB mismatched conditions, with an improvement of 13.7%. This result is significant, as the SIF only requires clean data for training, hence should achieve a similar performance across a wide range of noise conditions. On the other hand, the multi-conditional training method requires a large amount of data for training, and may not perform well when the noise conditions during testing are different to those observed during the training.

The performance of the SIF is also compared to the other image processing inspired approaches. From the results in Table 3.3, it can be seen the SIF compares well to these methods, as it appears that none of the image-based methods is particularly robust to mismatched noise. In addition, the spectrographic seams and SIFT

BOVW approaches appeared to be less discriminative than the conventional baseline methods, with a classification accuracy of only 87.0% and 89.0% respectively in clean conditions. While the ordered BOVW approach improves upon this by incorporating temporal information, this only gives a small improvement in performance. Among the four image-based baseline methods, the frame-based HOG method performed best, and achieves the highest overall performance in clean conditions. Similar to the conventional baseline methods, the HOG system uses a conventional HMM recogniser to capture the temporal information in the sound events. However, it is not as robust to noise compared to the SIF or conventional baseline methods. This should be expected, since the local spectral gradients that are extracted to form the feature will be severely corrupted by the noise.

A question that might be asked is why the SIF, a two-dimensional feature, can be compared with frame-by-frame approaches such as MFCCs. Simply comparing the feature dimensions, then at first this appears valid, as the SIF has 486 dimensions, while the frame-by-frame features have just 36. However, once combined with HMMs, the number of parameters increases dramatically. For the HMM system used in these experiments, with 5 states and 6 Gaussians, there are $36 \times 6 \times 2 \times 5 = 2160$ parameters for each HMM model, where the 2 represents the mean and variance of each Gaussian. In addition, such frame-by-frame methods are considered state-of-the-art for acoustic recognition tasks, particularly in speech. Therefore the experimental comparison carried out is considered to be sufficient.

3.5 Summary

This chapter introduced the idea of spectrogram image processing for SER. The idea is to overcome the drawbacks of the state-of-the-art techniques by naturally capturing the two-dimensional spectro-temporal information in the spectrogram image. To establish the background, both the state-of-the-art in image processing, and existing spectrogram image-based approaches for SER were reviewed. Motivated by these, the spectrogram image feature (SIF) was then proposed for sound event recognition, which quantises and maps the dynamic range of the spectrogram image to a higher-dimensional space to produce a robust feature for classification. This was demonstrated through a detailed set of experiments that compared the different aspects that con-

tributed to the success of the SIF, and showed a strong performance against a range of baselines from both conventional audio processing and image processing inspired approaches. However, it was found that while the linear power colour quantised SIF performed robustly in mismatched conditions, it could not match the performance of the baseline systems in clean conditions. In the next chapter, this aspect is further analysed, and a new sound event image is proposed to improve upon the existing SIF framework.

Chapter 4

Generating a Robust Sound Event Image Representation

In this chapter, a novel sound event image representation is proposed to improve upon the SIF that was introduced in the previous chapter. This is motivated by the drawbacks of the spectrogram image, where it is difficult to develop a reliable missing feature mask that can distinguish the sound event from noise. The proposed representation is called the subband power distribution (SPD) image [6, 7], which is a novel two-dimensional representation that characterises the spectral power distribution over time in each frequency subband. Here, the high-powered reliable elements of the spectrogram are transformed to a localised region of the SPD, such that they can be easily separated from the noise using a missing feature mask. The image feature framework is then applied to the SPD to generate a novel feature called the SPD-IF. Further to this, a non-stationary missing feature mask estimation process and a kNN missing feature classifier are also proposed to marginalise the noise-affected SPD-IF feature dimensions.

The chapter is organised as follows. Section 4.1 first describes the motivation for finding a robust alternative to the conventional spectrogram image representation. Section 4.2 then introduces the proposed SPD image, before detailing a novel noise estimation technique and the proposed missing feature classification system. Experiments are then carried out in Section 4.3 to compare the SPD-IF against the state-of-the-art baseline techniques.

4.1 Motivation

This section discusses the motivation to find an alternative sound event image representation that can improve upon the SIF approach from Section 3.3. The problem is that the SIF is based on the spectrogram image representation, which leads to two issues in mismatched noise conditions. Firstly, the noise directly changes the pixel values in the spectrogram, which will affect the extracted image feature. Secondly, the way noise affects the spectrogram makes it challenging to apply a robust classification system to the SIF, such as a missing feature approach. These two aspects are now discussed below, and lead to the development of the SPD sound event image representation in Section 4.2.

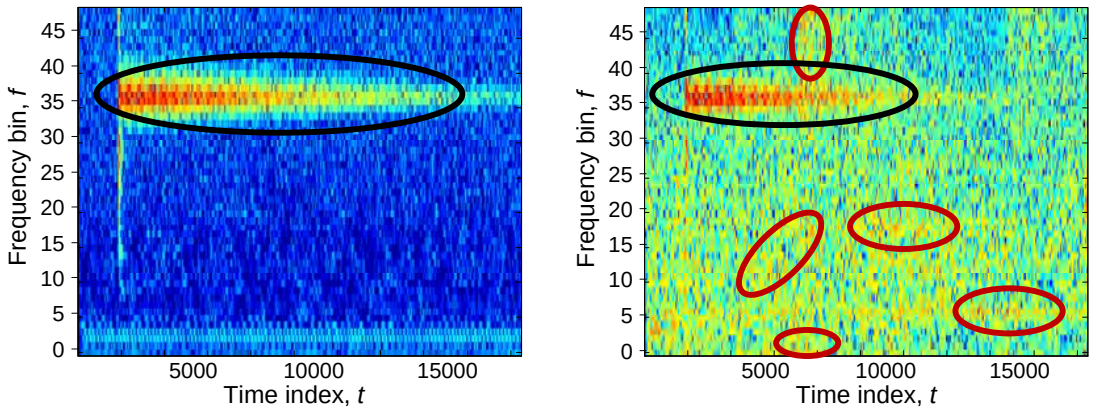
4.1.1 Effect of Noise on the Spectrogram

Due to the physical nature of many sounds, it is typical for the sound event spectrogram to be sparse. This means that the sound energy is concentrated in a small number of localised time-frequency regions that contain the most characteristic sound information. This is in contrast to diffuse noise, where the spectral information is typically spread more evenly across the frequency spectrum. When sound is captured under noisy environments, the interaction between the signals in the spectrogram is covered by the MixMax principle [98], as follows:

$$\log\left(|s_1| + |s_2|\right) \approx max\left(\log|s_1|, \log|s_2|\right) \tag{4.1}$$

where the stronger of the two signals, s_1, s_2, will dominate in the log-spectral domain. An example of this can be seen in Fig. 4.1. Here, the sparse harmonic of the bell sound can be clearly seen in both clean and noisy conditions, as highlighted by the black oval overlaid on the spectrograms. However, the diffuse noise masks some of the low power bell sound information, for example at the onset and towards the end of the harmonic.

In the previous work on the SIF, the procedure was to extract a feature by quantising and mapping the dynamic range of the spectrogram into separate regions, and characterising each region independently. The idea was that the high-power spectral information, such as that belonging to the bell sound in Fig. 4.1, would be mapped



(a) Clean, where it is simple to separate the bell sound from the background noise.

(b) Noisy conditions, where regions of noise may be mistaken for signal.

Figure 4.1: Example of the difficulty in generating a missing feature mask to separate the target signal from the background noise.

into a separate region that would remain robust in mismatched conditions. Then, while the low-power quantisations are likely to be affected by the mismatched noise, the SVM classification would still be robust as it assigns a higher weight to the more discriminative high-power components. However, in practice it was found that random fluctuations in the noise may also be mapped into the high-power region, such as those highlighted in red in Fig. 4.1b, which can affect the classification performance.

A further effect of the noise is that it makes the detection of the sound event much less reliable. Under real-world conditions, the performance of the detection algorithm may affect the onset and offset detection of the sound event from the continuous audio stream. This causes an issue with the practical implementation of the SIF method, as the length of the detected segment will affect the time-frequency partitioning of the spectrogram. This in turn will cause a shifting of the feature dimensions, which could lead to a mismatch in classification of the SIF against the clean training samples. One possible solution to this problem is to use a fixed-length sliding window detector as opposed to a feature-based detection algorithm. However, this may not be suitable in cases where the sound events have a wide variation in duration, and may lead to an increase in false detections. Hence an alternative solution is explored in this chapter.

4.1.2 Robust Classification

One solution to the problem of mismatch between clean and noisy conditions is to incorporate the SIF within a missing feature framework, referred to here as the MF-SIF. This approach aims to marginalise the feature dimensions that are affected by the noise, to leave only the reliable, high-power regions of the sound event as the basis for classification. Such a missing feature approach has been shown to improve performance in noisy conditions for tasks in both SER and ASR [103, 104], and was demonstrated by the missing feature experiments in Section 2.3.

However, the biggest challenge for missing feature approaches is the task of mask estimation, which determines the feature dimensions that have been corrupted by noise. Although this appears simple in theory, in practice missing feature mask estimation poses a significant problem, particularly when based on the time-frequency spectrogram. This is because the non-stationary nature of the noise across time, frequency and the dynamic range, makes developing a reliable mask challenging [104]. An example of this can be seen in Fig. 4.1, where the bell sound is shown in clean and noisy conditions. In clean conditions, it is simple to separate the regions of high power belonging to the bell sound from the background. However, when the noise power becomes comparable to the signal, as in Fig. 4.1b, the sound event is much more difficult to detect as the random noise fluctuations can easily be mistaken as reliable sound information. This severely reduces the classification performance under such conditions.

Another difficulty with conventional missing feature approaches, is that existing mask estimation methods may not work well with the wide variety of sound events. This is particularly true of classifier-based mask estimation [105], since such methods require prior knowledge about the characteristics of the target signal and noise. While this may work well for speech, it is difficult to design a feature set that can reliably discriminate the wide variety of sound event characteristics from noise. Hence, classifier-based methods will not work well for sound events in practice. The other common mask estimation approach uses local SNR estimates as a threshold for each element in the spectrogram [105]. However, in many cases it may be impossible to distinguish certain noise elements from the target sound, particularly in low SNR conditions when the noise can have comparable energy and sharp peaks to the signal. In

such cases, noise regions may be marked as reliable, and have a severe effect on the classification performance. Also, as local-SNR methods assume that the noise is stationary over time, these approaches will not work well for more realistic non-stationary noise environments.

Therefore, it is challenging to generate a missing feature mask which can be reliably combined with the SIF based on the spectrogram image representation. Hence, the approach taken in this chapter is to find an alternative sound event representation that can overcome some of these drawbacks. This is called the subband power distribution (SPD) image, and is introduced in the next section.

4.2 Subband Power Distribution Image Feature

In this section, the two-dimensional subband power distribution (SPD) image is introduced as an alternative sound event image representation. This is motivated by the challenges of extracting a robust image feature directly from the spectrogram, as discussed in the previous section. The SPD captures the stochastic distribution of spectral power over time in each frequency subband, such that the reliable, high-power spectral information is transformed to a localised region of the SPD image. This then allows an image feature to be extracted, called the SPD-IF, which can easily be combined with a missing feature framework for robust classification. This section first introduces the idea behind the SPD-IF framework, and then describes the algorithm for generating the SPD and the proposed missing feature classification system.

4.2.1 The SPD-IF Framework

The SPD-IF framework builds upon the success of the SIF for robust sound event classification using image feature extraction from the time-frequency spectrogram. However, instead of performing image feature extraction directly on the spectrogram, the sound event is first transformed into the SPD image representation before extracting the image feature. An outline of this framework is shown in Fig. 4.2, which compares the SPD-IF with the previous SIF method. The key aspect of this framework is the novel sound event image representation called the subband power distribution (SPD). This characterises the spectral power distribution over time, in each frequency

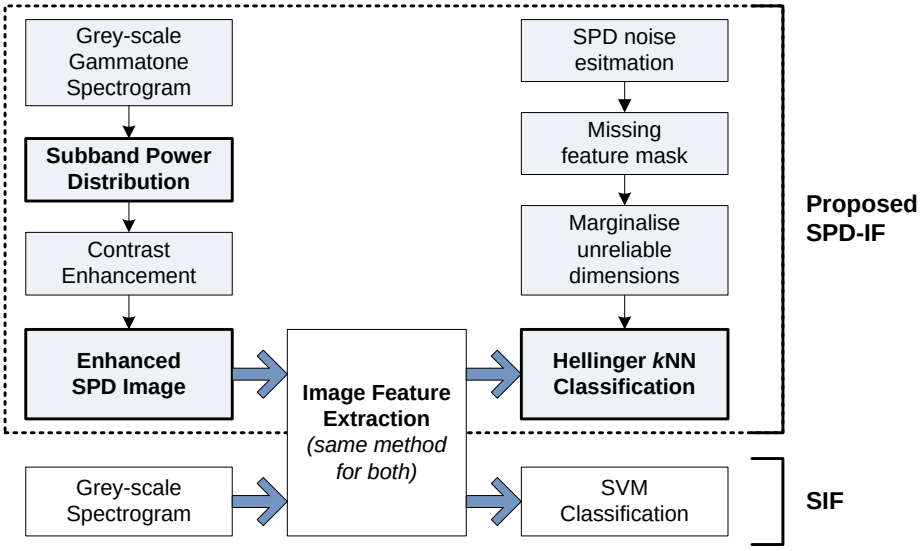
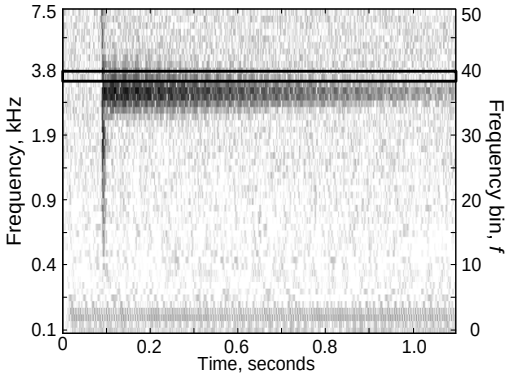


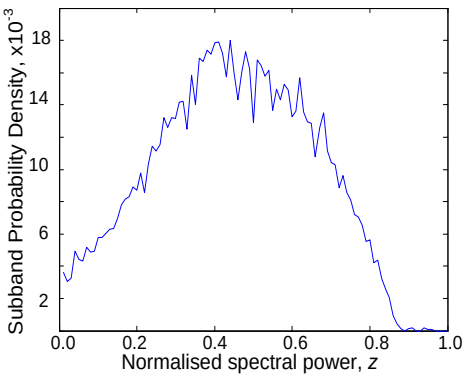
Figure 4.2: Overview of the proposed SPD-IF approach compared to the previous work on the SIF (the shaded boxes indicate the new contributions). First, a normalised SPD image of the sound is generated, before an image feature is extracted for classification. For the SPD, a missing feature kNN classification system is used, employing the Hellinger distance measure to compare the distribution distance between image feature dimensions.

subband, forming a two-dimensional representation of frequency against normalised spectral power. It can therefore be considered as a generalisation of the power spectral density (PSD), since the PSD is a one-dimensional representation of the average power present in the signal at each frequency bin [245].

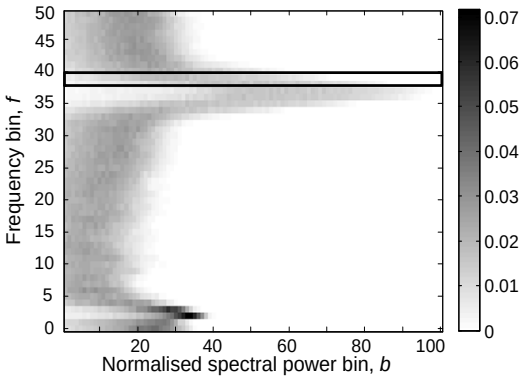
A graphical illustration of how the SPD is assembled is shown in Fig. 4.3. First, the distribution of normalised spectral power in each subband of the log-power spectrogram is calculated over time. In the example shown, the highlighted subband for the bell sound in Fig. 4.3a is characterised by the distribution shown in Fig. 4.3b. Then, each subband distribution is stacked together to form the two-dimensional SPD representation in Fig. 4.3c, which visualises the distribution of normalised spectral power against frequency. A final step is then taken is to enhance the contrast of the raw SPD representation, to give the final SPD image in Fig. 4.3d. This is performed to normalise the distribution information to a fixed range, and is important as it maximises the amount of information extracted from the SPD, as explained later in Section 4.2.2. The final SPD image forms the basis for image feature extraction, with the resulting



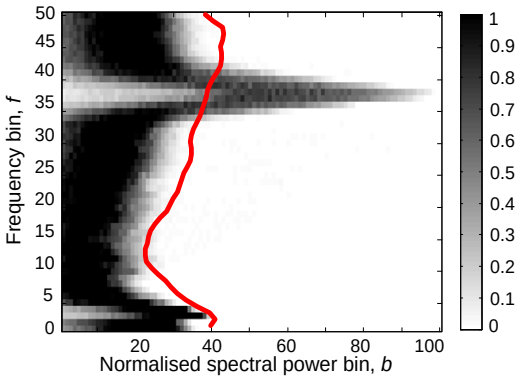
(a) Normalised gammatone spectrogram of a bell ringing sound event, $G(f, n)$.



(b) Raw subband probability distribution in subband, f=39.



(c) Raw SPD, $H(f, b)$, formed by stacking the subband distribution information across frequency.



(d) SPD Image, $I(f, b)$, after contrast enhancement has been performed.

Figure 4.3: Overview of generation of the SPD Image. The probability distribution is taken over each subband, as shown by the example in (b), and are stacked to form the raw SPD in (c). This undergoes contrast enhancement, to give the SPD in (d). The red line in (d) indicates an upper estimate of the noise over the clip. Areas to the right of this line are considered to contain only signal, while the rest is dominated by the noise distribution.

feature called the SPD-IF.

The advantage of the SPD representation over the spectrogram is twofold. Firstly, it is less affected by any time shifting of the sound event detection algorithm, as the temporal information is characterised by the distribution of the spectral power. This gives the SPD image a fixed size that is independent of the clip length, which is

unlike the spectrogram where one dimension explicitly represents the time information. Secondly, the signal and noise information are much more easily separated in the SPD representation compared to the spectrogram. This is because the high-powered reliable elements of the sound event are transformed to a localised region of the SPD. This is due to the physical characteristics of many sound events that produce a sparse spectrogram representation, meaning that a large proportion of energy is contained in only a few frequency bands. By comparison, many noise conditions are diffuse, with energy spread across the frequency spectrum. Hence, even for 0dB noise, the sparse signal components will still be much greater than the noise energy, thereby satisfying the MixMax principle and remaining separable from the noise.

From the SPD image, an image feature can be extracted using the same process as for the SIF. This proceeds by first quantising and mapping the dynamic range into separate regions, before partitioning the image into local blocks and extracting the pixel distribution statistics from each block. The next step is then to separate the reliable signal region from the noise in the SPD image, such that a missing feature mask can be estimated for robust classification. This process is significantly simplified in the SPD compared to the spectrogram, since the reliable high-power signal information is transformed to a localised region of the SPD and can be separated from the noise simply by a line. This is unlike the spectrogram where the boundary between noise and signal forms a complex two-dimensional time-frequency surface. An example of the separation between signal and noise in the SPD is demonstrated by the red line shown in Fig. 4.3d. The signal information is represented in the SPD region to the right of this line, while the noise dominates the region to the left of the boundary. This noise boundary can be estimated either using conventional stationary noise estimation methods, or using the SPD directly, as discussed later in Section 4.2.3. Any high-power noise peaks, which could be easily mistaken for the signal in the spectrogram, as in Fig. 4.1b, will appear much less significant after taking the distribution to form the SPD image. This is because such peaks occur randomly across time and frequency, and hence they will be assigned a very low value when the subband distribution is taken across the whole sound segment. Therefore, even if these noise peaks fall into the reliable region of the SPD image, they will not have such a significant effect on the image feature that is extracted or the classification performance of the SPD-IF.

The final step in the SPD-IF framework is to perform classification of the sound

event using a missing feature classification system. The idea is to marginalise the components of the SPD-IF belonging to the noise, such that only the robust signal information is used as a basis for classification. For classification, kNN is used with the Hellinger distance measure, which is chosen as it naturally measures the similarity between the pixel distribution information captured in the SPD-IF extraction process. Together, the SPD provides the basis for a classification system that will be significantly more robust than the equivalent system based on the spectrogram. The rest of this section now describes each of the steps in the SPD-IF framework in detail.

4.2.2 Subband Power Distribution Image

Starting from a time-frequency spectrogram representation of the sound, the SPD image is designed to represent the distribution of spectral power in each frequency subband over time. As this captures the long term temporal distribution statistics, it is desirable for the spectrogram to have a high time resolution, to better capture the distribution. Therefore, the gammatone filterbank decomposition is chosen as the model for time-frequency analysis, as previously introduced in Section 3.1.2. This has the advantage that there is no tradeoff between time and frequency resolution, which is a common drawback of the conventional short-time Fourier transform (STFT) representation. Note that the log-power spectrogram is used as it compresses the dynamic range of the sound event information in the SPD, which in turn increases the discriminative power of the SPD-IF.

The gammatone spectrogram, $S_g(f,n)$ from equation (3.8), is first normalised into a grey-scale image, $G(f,n)$, in the range $[0,1]$ as follows:

$$G(f,n) = \begin{cases} \frac{S_g(f,n)}{\max\limits_{f,n}(S_g(f,n))} & \forall\ S_g(f,n) \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where f represents the centre frequencies of the filters and n is the time index. A bank of $F = 50$ filters is used, with centre frequencies equally spaced between 100-8000 Hz on the equivalent rectangular bandwidth (ERB) scale [162].

The SPD is then formed by finding the distribution, $P_{G_f}(z)$, of the normalised spectral power, $z = [0,1]$, in each frequency subband over time. Here, P is the probability

density function, and G_f is a random variable representing the normalised spectrogram, $G(f,n)$ in the frequency subband, f. To estimate this distribution, the simplest solution is to use a non-parametric approach based on the histogram. This has the advantage of speed and simplicity, and benefits from the fact that the upper and lower bounds of the histogram bins are fixed by the normalisation of the spectral power to a grey-scale intensity. The subband distribution therefore becomes:

$$P_{G_f}(b) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}_b(G(f,n)), \quad \forall\ b = 1,\ldots,B \tag{4.3}$$

where N is the number of time samples in the segment, B is the number of histogram bins, and $\mathbf{1}_b$ is the indicator function, which equals one for the b^{th} bin if $G(f,n)$ lies within the range of the bin and is zero otherwise, as follows:

$$\mathbf{1}_b(G(f,n)) = \begin{cases} 1, & \text{if } y(b-1) < G(f,n) \leq y(b) \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

where $y(b) = \frac{b}{B}$ is used to calculate the edges of the histogram bins, which are spaced evenly across the $[0,1]$ range of the normalised spectral power, z.

The raw SPD, $H(f,b)$, is then the two-dimensional representation that is formed by stacking together each subband distribution vector over frequency to form a matrix, as follows:

$$H(f,b) = P_{G_f}(b) \qquad \forall f,b \tag{4.5}$$

where the result is an $F \times B$ matrix. This forms an image representation of the raw probability distribution information for each frequency subband over time, and is constrained to lie in the range $0 \leq H(f,b) \leq 1$. An example of this raw SPD representation is given in Fig. 4.3c, where a total of $B = 100$ bins are used.

However, although the raw SPD, $H(f,b)$, can already be considered as a grey-scale image for image feature extraction, it was found that most of the distribution information is contained within a small region of the dynamic range. This is due to the physical nature of many sound events, which have an attenuating or otherwise non-stationary spectrogram envelope. This means that, for a high enough number of histogram bins, it is unlikely for the signal to be stationary enough to give a high

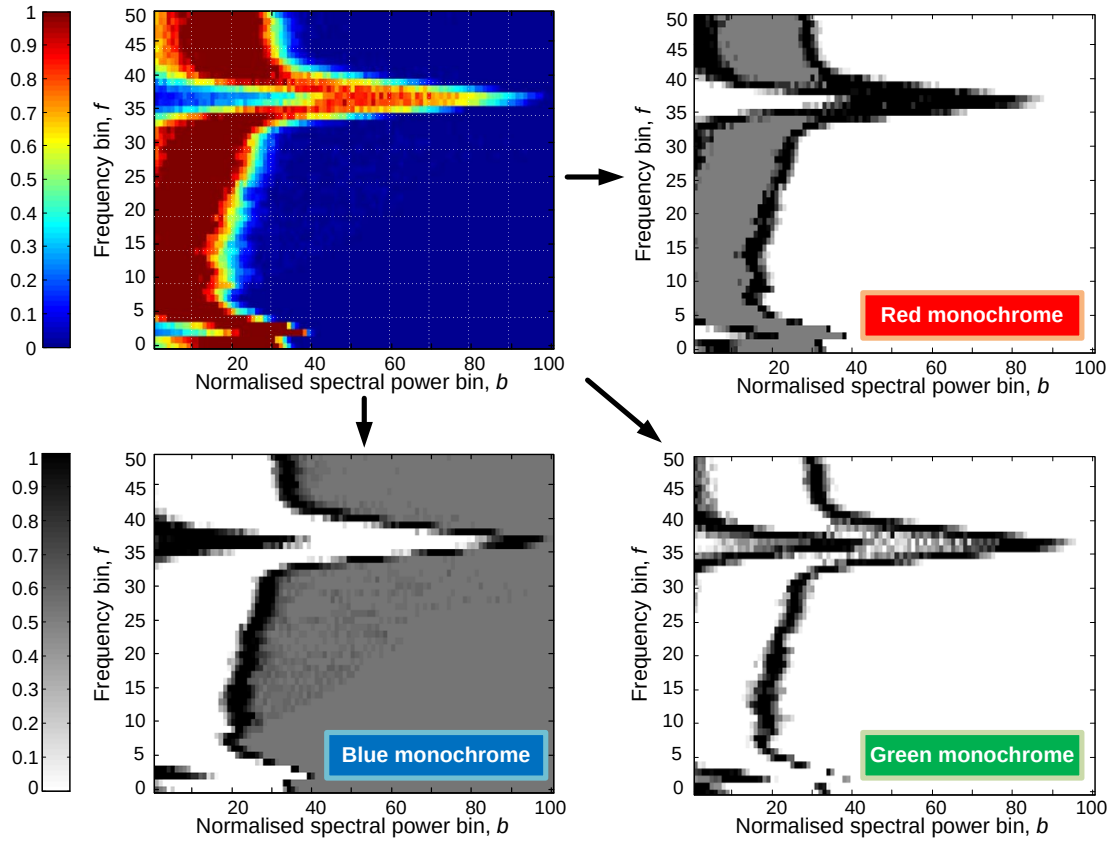


Figure 4.4: The three monochrome quantisations, $m_c(f, b)$, are shown for the SPD image, $I(f, b)$, of the bell sound above. These quantisations are labelled red, green and blue, to correspond with the RGB colours of the Jet mapping function from Fig. 3.10a. The dark areas of the monochromes indicate higher values of the quantisation, while the white area contains zeros.

density value in any given bin. Therefore, it is desirable to enhance the contrast of the raw SPD, to produce an enhanced SPD image, $I(f, b)$, that better represents the important signal information for classification. In image processing, this is referred to as "contrast stretching" [164], and is performed here as follows:

$$I(x = f, y = b) = \begin{cases} H(f, b) \times h, & \text{if } H(f, b) < \frac{1}{h} \\ 1, & \text{otherwise} \end{cases} \tag{4.6}$$

where h is an appropriate constant, and x, y represent the image indices for image fea-

ture extraction, as in (3.12). This operation does not affect fully stationary subbands, as these are still assigned a high value in the SPD image, hence this step was found to significantly improve the classification performance over a broad range of sound classes. Empirically, it was found that using $h = 50$ provides a sufficient enhancement in contrast. Comparing Figs. 4.3c and 4.3d demonstrates the improvement provided by the contrast stretching process. Here it can be seen that the bell sound information is much more clearly represented in the SPD image in Fig. 4.3d, as the distribution information has been stretched to cover the full dynamic range of the image.

The two-dimensional SPD image, $I(f, b)$, now forms the basis for image feature extraction in the same way as for the SIF, as previously introduced in Section 3.3.2. For the dynamic range quantisation in (3.12), the "Jet" mapping function is used, as this was found to give better results that the HSV mapping in preliminary experiments. This may be explained by the confusion in the "red" HSV mapping, which represents both high and low pixel values. An example of the RGB monochromes produced by the "Jet" quantisation for a bell sound are shown in Fig. 4.4. It can be seen that the "blue" quantisation captures the more non-stationary information from the sound event, as shown by the black areas around the edge of the signal region. On the other hand, the more stationary signal information is captured in the "red" quantisation, as shown by the region of grey covering the region of the SPD corresponding to the stationary background noise. The "red" quantisation should also be less susceptible to random non-stationary noise fluctuations, which should also make it more robust for classification.

The final step in the image feature extraction is to characterise the pixel information in the monochromes using their distribution statistics. For the SPD-IF, the same extraction process is used as for the SIF, which was detailed previously in Fig. 3.11 and equation (3.14). It was found in preliminary experiments that partitioning each SPD dimension into $D = 10$ blocks, and using the mean and variance to represent the distribution ($k = \{1, 2\}$ from (3.14)) gave the best trade-off between performance and feature vector size. Therefore, the total feature vector length is $10 \times 10 \times 3 \times 2 = 600$, since there are 100 local blocks, three monochrome mappings ($c = 3$), and two k parameters representing the distribution of image pixels in each block.

4.2.3 SPD Noise Estimation

One advantage of the SPD representation is that the noise and signal are transformed to localised regions of the SPD image. They can then be separated simply by a one-dimensional line, as discussed previously in Section 4.2.1. To find this separating line, an estimate of the upper bound of the noise in the clip is required, as shown in the previous example in Fig. 4.3d. Although it is possible to use conventional approaches to find this estimate, in this section a novel method to estimate the noise using the SPD representation is introduced. This has the advantage that it can adapt to non-stationary noise conditions, and works by finding the cross-correlation between a noise-only SPD and the SPD of the sound event in noise. This can then be used to generate a missing feature mask for the SPD, which forms part of the missing feature classification system described in the next section.

The most common approach for noise estimation in conventional audio processing is to generate a noise model using an initial audio segment without any sound events present [158]. From this, a missing feature mask is commonly estimated in the spectral domain based on the local SNR, in a similar way to spectral subtraction [99]. This is typically done by setting a threshold on the local SNR, δ, and denoting spectral information that falls below this value as noise [103]. This then forms the boundary between signal and noise, and can be written as follows:

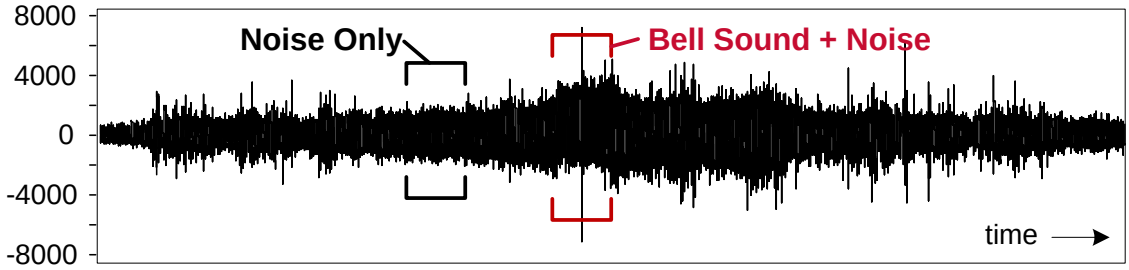
$$n(f) = \mu\left(G(f, n_N)\right) + \delta(f) \tag{4.7}$$

where $\mu\left(G(f, n_N)\right)$ is the noise estimate obtained from the spectrogram, $G(f, n_N)$, over the noise-only segment with time samples n_N. The threshold $\delta(f)$ can simply by a constant, however a better estimate can be obtained by setting it to two times the standard-deviation of the noise in each frequency subband:

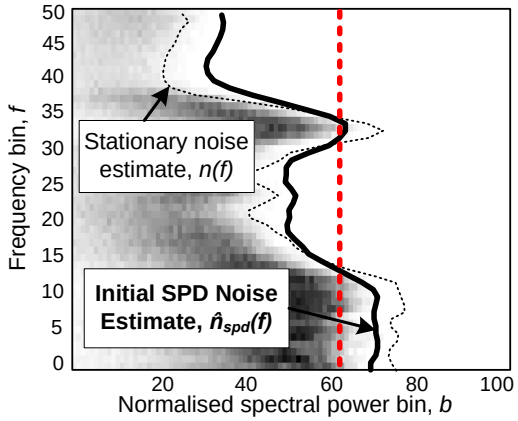
$$\delta(f) = 2\sigma\left(G(f, n_N)\right) + \Delta \tag{4.8}$$

where $\sigma\left(G(f, n_N)\right)$ is the standard deviation of the spectral information over the noise-only time samples, and Δ is a small constant added to account for unseen noise fluctuations.

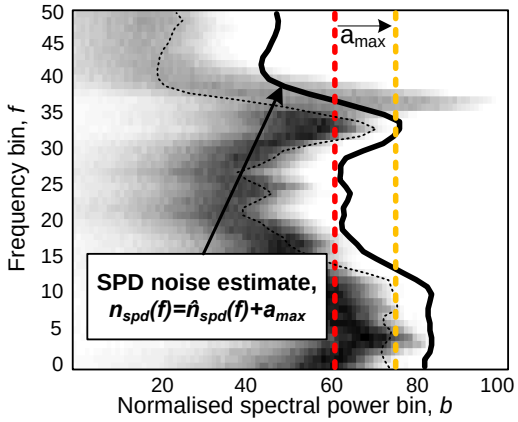
The problem with this type of noise estimation is that it is "stationary", since it



(a) An example of non-stationary wind noise, where the noise level increases and then decreases over time.



(b) The noise-only SPD, $H_N(f,b)$. An initial noise estimate, $\hat{n}_{spd}(f)$, is found directly by finding the maximum bin in each frequency subband.



(c) Noisy bell sound SPD, $H(f,b)$. The SPD noise estimate, $n_{spd}(f)$, is estimated based on the maximum subband cross-correlation, a_{max}.

Figure 4.5: Overview of the SPD noise estimate approach. The SPD noise estimate, $n_{spd}(f)$, is shown by the solid black line in (c), and has been shifted by a_{max} to reflect the increase in noise intensity. This can be compared to the stationary noise estimate, $n(f)$ from (4.7), shown by the dotted lines, which cannot adapt to the change in non-stationary noise intensity over time.

is estimated based on an initial noise segment, and hence cannot track the noise well in non-stationary noise environments. This is demonstrated in Fig. 4.5, which shows a clip containing non-stationary wind noise and a bell sound. Here, the stationary noise estimate is obtained from the noise-only period in Fig. 4.5a, and can be seen initially to fit the noise profile well, as shown by the dotted line in the noise-only SPD in Fig. 4.5b. However, as the noise is not re-estimated when the signal is present, the stationary noise estimate provides a poor fit when the bell sound occurs, as shown in

Fig. 4.5c. This may result in misclassification of the bell sound, since regions of the SPD containing noise will be marked as reliable when it comes to the missing feature classification.

To improve upon this, a noise estimation approach is proposed here that utilises the information in the SPD representation, and can adapt to changes in non-stationary noise conditions. This idea is based on the assumption that the characteristics of the noise distribution remain the same over time, despite changes in the non-stationary noise intensity. In the SPD representation, this change in intensity can be approximated simply as a shift in normalised spectral magnitude of the noise distribution. Therefore, the same initial noise segment provides an estimate of the shape of the noise distribution, and then the correlation between the noise-only SPD, $H_N(f,b)$ and the noisy-signal SPD, $H(f,b)$ can be used to estimate the change in non-stationary noise intensity. This is illustrated in Fig. 4.5b and 4.5c, where the increase in non-stationary wind noise is labelled a_{max}, and the SPD noise estimate can be shifted to better track the noise boundary in the SPD. Note that the raw SPD, $H(f,b)$, is used for the noise estimate as opposed to the SPD image, $I(f,b)$. The reason is that in this case the noise is more important than the signal, and hence there is no need to apply the contrast enhancement step, which may also distort the noise information in certain circumstances and result in a lower overall performance.

To generate the proposed SPD noise estimate, the first step is to obtain a noise-only SPD, $H_N(f,b)$, from a segment containing only noisy time samples, n_N, calculated using (4.3). An upper bound of the noise in the noise-only SPD is then estimated to provide an initial noise estimate. This is calculated as the maximum occupied bin for each frequency subband, as follows:

$$n_{max}(f) = \underset{b}{\operatorname{argmax}}(H_N(f,b) > 0) \tag{4.9}$$

where b represents the bins across the normalised spectral power dimension of the SPD, and $n_{max}(f)$ is the bin that represents the upper bound of the noise region. This is further smoothed to avoid sharp discontinuities across frequency using a moving

average filter, as follows:

$$\hat{n}_{spd}(f) = \frac{1}{M} \sum_{i=f-M/2}^{i=f+M/2} n_{max}(i) \tag{4.10}$$

where M is the order of the filter, and $\hat{n}_{spd}(f)$ is the initial SPD noise estimate representing the upper bound noise estimate for the noise-only segment, as shown in Fig. 4.5b.

Then, given an noisy-signal SPD containing a sound event clip, $H(f, b)$, the noise intensity change, a_{max}, is found by calculating the subband cross-correlation between $H_N(f, b)$ and $H(f, b)$, as follows:

$$[a_{max}, f_{max}] = \underset{a,f}{\operatorname{argmax}} \Big(H(f, b) \star H_N(f, b + a) \Big), \quad \forall \ -B < a < B \tag{4.11}$$

where $\star$ represents the cross-correlation across the subband bins, and the range of a is determined by the size of the raw SPD image. Note that the cross-correlation is performed separately on each SPD subband, f, as $H(f, b)$ is a mixture of the noise and signal distributions and hence the highest correlation should occur between two noise-dominated subbands. The final SPD noise estimate, $n_{spd}(f)$ is then simply:

$$n_{spd}(f) = \hat{n}_{spd}(f) + a_{max}. \tag{4.12}$$

where $\hat{n}_{spd}(f)$ is the initial noise estimate calculated from the noise only segment in (4.10). Note that a_{max} is not limited to be positive, as the noise intensity can both increase or decrease, as shown by the shape of the wind noise example in Fig. 4.5a.

4.2.4 Missing Feature Classification

A missing feature classification system is now proposed for the SPD-IF. The idea is to utilise the SPD noise estimate found in the previous section to generate a missing feature mask for the SPD image. A missing feature classifier is then used to marginalise the unreliable feature dimensions, such that classification is based only on the reliable signal information. These two processes are now described below.

Mask Estimation

When generating the SPD image, $I(f, b)$, the sparse, high-power signal components are transformed to a continuous region of the image. Therefore, there exists a boundary, ∂I, between clean and noisy regions, where the image region above this boundary is derived only from the signal. To generate the missing feature mask, an approximation of this noise-signal boundary is required. However, it is clear that the reliable SPD boundary, ∂I, can be simply approximated by the noise estimate in the clip, $n_{spd}(f)$, since this is also an upper bound on the noise distribution in the SPD image. Therefore, the reliable region of the SPD, $I_r(f, b)$, can be assigned as follows:

$$I(f,b) \rightarrow \begin{cases} I_r(f,b), & \text{if } b > n_{spd}(f). \\ I_u(f,b), & \text{otherwise.} \end{cases} \tag{4.13}$$

where the subscripts r, u denote reliable and unreliable image regions respectively.

This mask can now be applied to the SPD-IF feature. If a sub-block of the SPD image, denoted L_{ij}, is intersected by the noise estimate, $n_{spd}(f)$, the whole block must be assumed to be unreliable. This is because the feature, x_{ij}, is based on the distribution statistics of the image pixels within the block L_{ij}, which will be affected by any noise pixels contained within the block. Hence, the reliable feature vector, x_r, can be assigned as follows:

$$x_{ij} \rightarrow \begin{cases} x_{r,ij}, & \forall I(p,q) \subset L_{ij} \rightarrow I(p,q) \in \{I_r\}. \\ x_{u,ij}, & \text{otherwise.} \end{cases} \tag{4.14}$$

where the sub-block with indices ij can only be considered reliable if all the pixels in the block, $I(p, q) \subset L_{ij}$, belong to the set of reliable pixels $\{I_r\}$. The reliable feature vector, x_r, is subsequently concatenated to form a single vector. Meanwhile, the unreliable feature dimensions, x_u, can now be marginalised as they do not contain useful signal information.

Classification

Here, kNN is used for classification, which, although uncommon in the acoustic field, is relatively common in image processing and can achieve comparable performance with SVM [228]. However, an important advantage of kNN is that it can be easily combined with a missing feature framework, as this is not straightforward for the SVM classifier. It also offers flexibility in the choice of distance measure, and here the Hellinger distance is chosen as it measures the similarity between two distributions derived from the data. This fits naturally with the SPD-IF, which models the distribution of pixels in each monochrome image subblock. This is preferred over the conventional Euclidean distance, which measures the distance between mean and variance parameters independently as elements of the feature vector, rather than using them to calculate the distribution distance. Among the probabilistic distances, the Hellinger distance has superior properties, as it is bounded, symmetric non-negative and skew insensitive [246], hence is preferred for this task.

As the feature information in the SPD-IF is characterised by a normal distribution, using the mean and variance of the image pixels, the Hellinger distance between two SPD-IF vectors, x_1, x_2, can be written as follows:

$$\mathrm{d_H}(x_1, x_2) = \sum_{k=1}^{N_r} \left(1 - \sqrt{\frac{2\sigma_{1,k}\sigma_{2,k}}{\sigma_{1,k}^2 + \sigma_{2,k}^2}} e^{-\frac{1}{4}\frac{(\mu_{1,k}-\mu_{2,k})^2}{\sigma_{1,k}^2+\sigma_{2,k}^2}} \right)^{\frac{1}{2}} \tag{4.15}$$

where N_r is the number of reliable dimensions, and μ, σ are the mean and variance parameters respectively that are extracted as part of the SPD-IF. Note that for certain feature dimensions, the variance may be small or even zero. Therefore, a floor is applied to the variance such that a measure of the similarity between these dimensions can still be obtained. This is set at $1e^{-3}$, which is equivalent to having only a single non-zero pixel.

The conventional Euclidean distance between the SPD-IF vectors can also be calculated as follows:

$$\mathrm{d_E}(x_1, x_2) = \frac{1}{N_r} \left[\left(\sum_{k=1}^{N_r} (\mu_{1,k} - \mu_{2,k})^2 \right)^{\frac{1}{2}} + \left(\sum_{k=1}^{N_r} (\sigma_{1,k} - \sigma_{2,k})^2 \right)^{\frac{1}{2}} \right] \tag{4.16}$$

This measures the distance between mean and variance parameters independently as elements of the feature vector, rather than using them to calculate the distribution distance. The Euclidean distance is used for experimental comparison with the proposed Hellinger kNN in the next section.

4.3 Experiments

In this section, experiments are conducted to compare the performance of the SPD-IF method with both the SIF and baseline methods for SER. The same database of environmental sounds is used as in Section 2.3, with testing carried out in both clean and mismatched noise conditions to simulate a more realistic testing environment. Training is performed using only clean samples, while testing is carried out across a range of noise conditions.

4.3.1 Experimental Setup

For comparison with the previous baseline methods, the same experimental setup is used as in the previous experiments in Section 2.3. Hence, the same 50 sound event classes are selected from the RWCP database, and as before, 50 files are randomly selected for training and 30 for testing in each of the five runs of the experiment. The classification accuracy for the SPD-IF is investigated in mismatched conditions, using only clean samples for training, with the average performance reported in clean and at 20, 10 and 0 dB SNR noise conditions. The standard deviation is also reported ($\pm$) across the five runs of the experiment and the four different noise conditions.

SPD-IF Evaluation Methods

This set of experiments is designed to analyse the SPD-IF method in detail, and to compare each step of the process separately. Therefore, the following experiments are carried out:

1. SPD image representation vs. Spectrogram (SIF)
2. Hellinger vs. Euclidean distance measure

3. SPD vs. Stationary noise estimate

In all cases, the same image feature extraction and kNN classifier is used, with the parameter $k = 5$ set for the class decision, which is used throughout. For the first case of SPD vs. Spectrogram, it is important to ensure a fair comparison by generating a missing feature mask for the spectrogram, with the approach called MF-SIF. The noise mask for the MF-SIF is derived in an analogous way to that of the SPD-IF, albeit across the time, frequency and dynamic range dimensions of the log-power STFT spectrogram. For testing, the "Factory Floor" noise condition is used, as this was found to be the most challenging noise condition for each of the above methods.

Baseline Methods

The performance of the SPD-IF is also compared against both the SIF and the best performing baseline methods from the evaluation in Section 2.3, as follows:

1. Spectrogram Image Feature (SIF), using the best performing raw-power STFT spectrogram and SVM classifier.

2. Baseline MFCC-HMM with Advanced Front End (AFE) noise reduction [100].

3. Baseline multi-conditional MFCC-HMM.

Both of the MFCC-HMM methods above use 36-dimension frame-by-frame MFCCs, with 12 cepstral coefficients, without the zeroth component, plus their deltas and accelerations. The HMM uses 5 states and 6 Gaussian mixtures, with both training and testing carried out using HTK [161].

The two baseline methods are chosen to provide a meaningful comparison with the previously evaluated techniques. In particular, the AFE method is chosen to provide a fair comparison with the SIF and SPD-IF, as it uses the same clean data for training and gave the best performance in the earlier baseline experiments. In addition, the multi-conditional MFCC-HMM method was found to give the best baseline performance overall, although it requires noisy data for training, as opposed to only clean data for the proposed SPD-IF.

4.3.2 Results

The results from the experiments on the SPD-IF are now presented. The important factors contributing to the success of the SPD-IF method are first analysed, before comparing the best performing SPD-IF method against the baseline techniques.

SPD vs. Spectrogram

The results comparing the SPD-IF and MF-SIF, using the same kNN missing feature classification approach with Euclidean distance, are shown in Fig. 4.6. It can be seen that the average improvement for the SPD-IF over the MF-SIF is around 6%, and this is found to be consistent across all noise conditions, as shown in the detailed results from Table 4.1.

The reason for this result is that the SPD representation transforms the reliable, high-power sound event information to a continuous region of the SPD image, unlike the spectrogram where the same information is spread across time and frequency. The signal region of the SPD can then be easily separated from the noise region, as the boundary is simply the estimated upper bound of the noise in the clip. However for the MF-SIF, the mask must be applied across both time, frequency, and the dynamic range mappings of the image feature, making it less reliable in non-stationary noise. In addition, the MF-SIF has only a coarse dynamic range quantisation, with just three mapping dimensions as in pseudo-colouring in image processing. Hence, it was found that the noise estimate could label some signal regions as unreliable, even when only a small amount of noise corrupted that region of the spectrogram. On the other hand, the SPD-IF has a finer partitioning of the dynamic range, hence the noise mask can better separate the noise and signal regions of the SPD.

Hellinger vs. Euclidean Distance Measure

From the results in Fig. 4.6 and Table 4.1, it is also possible to compare the effect of utilising the Hellinger distance measure for the SPD-IF. It can be seen the Hellinger distance measure gives an average improvement of around 3% over the Euclidean distance, but in particular gives a larger increase in severe noisy conditions. For example, for the SPD noise estimate, an improvement in classification accuracy of 8% is observed for the 0dB noise condition when using the Hellinger distance measure.

Method	kNN	Noise Mask	Clean	20dB	10dB	0dB	Avg.
SPD-IF	Hellinger	SPD	98.8 ± 0.3	98.0 ± 0.3	$\mathbf{96.5 \pm 0.5}$	$\mathbf{88.4 \pm 0.7}$	**95.4**
		Stationary	$\mathbf{99.0 \pm 0.3}$	$\mathbf{98.4 \pm 0.2}$	$\mathbf{96.5 \pm 0.4}$	82.6 ± 0.8	94.1
	Euclidean	SPD	98.9 ± 0.2	97.8 ± 0.1	94.5 ± 0.2	80.4 ± 0.9	92.9
		Stationary	97.0 ± 0.4	95.9 ± 0.2	91.5 ± 0.7	79.9 ± 0.9	91.1
MF-SIF	Euclidean	SPD	92.8 ± 0.3	90.2 ± 0.5	89.8 ± 0.4	74.3 ± 0.7	86.8
		Stationary	95.6 ± 0.6	90.4 ± 0.8	77.3 ± 0.8	69.1 ± 0.8	83.6

Table 4.1: Results comparing the performance of the image feature methods in mismatched noise conditions for the "Factory Floor" noise. The experiments also explore the proposed SPD vs. Stationary noise mask and Euclidean vs. Hellinger kNN classification.

These improvements should be expected, since the image features are capturing the pixel distribution information. It is therefore more natural to compare the distribution distance between each block of the image feature, rather than measuring the Euclidean distance between the parameters of the distribution. In the noisy conditions, it is suggested that while the noise may shift the pixel distributions, the change in distribution distance is less than the Euclidean distance since the reliable parts of the distribution are unchanged.

SPD vs. Stationary Noise Estimates

Comparing the results in Fig. 4.6, it can be seen that the proposed SPD noise estimate consistently outperforms the conventional stationary noise estimate by around $1-2\%$. In addition, from Table 4.1, it can be seen that the most significant improvement is found at 0dB, where the improvement for the kNN Hellinger distance SPD is almost 6% using the SPD noise estimate. This highlights the ability of the SPD noise method to adapt to non-stationary noise conditions, since these experiments are carried out in the most challenging Factory Floor noise condition.

SPD-IF vs. Baseline

The results in Table 4.2 compare the performance of the proposed SPD-IF against both the SIF and the best performing ETSI-AFE and multi-conditional MFCC-HMM

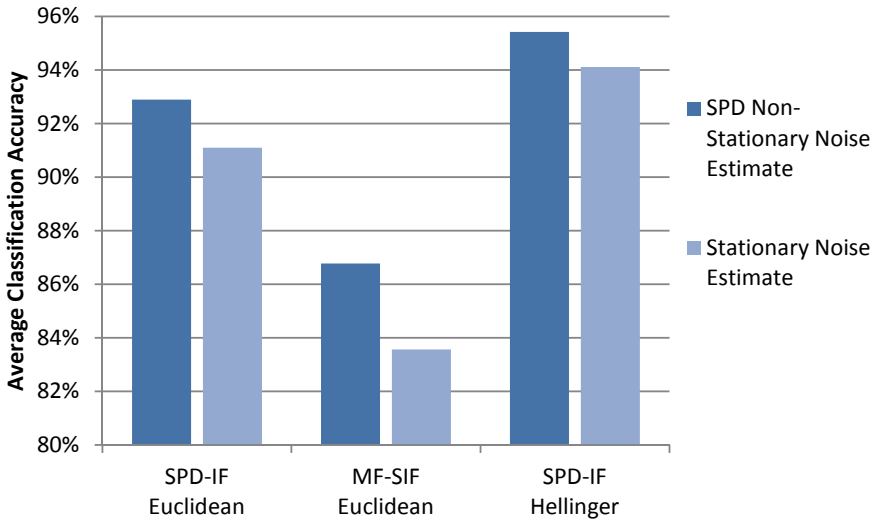


Figure 4.6: Classification accuracy results analysing the components of the SPD-IF method. Results comparing the SPD vs. spectrogram are represented in the SPD-IF and MF-SIF results. The SPD-IF is also implemented using both the Euclidean and Hellinger distance measures for kNN classification, while the stationary and SPD noise estimates are also compared from equations (4.7) and (4.12) respectively.

baselines. These results demonstrate that the SPD-IF performs significantly better than all three baselines, with an average classification accuracy of 95.9%. This is an improvement of over 6% compared with the best performing SIF and multi-conditional MFCC-HMM baselines. For the individual noise conditions, it can be seen that the SPD-IF outperforms the baselines in almost every case.

Importantly, the SPD-IF overcomes the drawback of the lower performance of the SIF in clean conditions, which was a result of the linear-power spectrogram used in the SIF, as opposed to the log-power representation used in the other methods. The SPD-IF also comes close to the performance of the AFE MFCC-HMM in clean conditions, with a difference of around 0.3%. This is significant given the strong performance of this baseline method in matched conditions. This reflects the discriminative nature of the SPD representation for a wide variety of sound event classes, since a large database of 50 sound event classes was used in these experiments.

The performance of the SPD-IF in noisy conditions also demonstrates the robustness of the representation, combined with the ability to easily separate the noise and

Group	Method	Clean	20dB	10dB	0dB	Avg.
Baseline	SIF	91.1 ± 1.0	91.1 ± 0.9	90.7 ± 1.0	80.9 ± 1.8	88.5
	ETSI-AFE	$\mathbf{99.1 \pm 0.2}$	89.4 ± 3.2	71.7 ± 6.1	35.4 ± 7.7	73.9
	Multi-Conditional	97.5 ± 0.1	95.4 ± 1.3	91.9 ± 2.7	67.2 ± 7.3	88.0
SPD-IF (Hellinger-kNN and SPD noise estimate)		98.8 ± 0.3	$\mathbf{98.0 \pm 0.3}$	$\mathbf{96.6 \pm 0.4}$	$\mathbf{90.3 \pm 2.0}$	**95.9**

Table 4.2: Classification accuracy results comparing the SPD-IF with the SIF and best performing baseline methods.

signal regions of the SPD. For example, in 0dB noise, the SPD-IF achieves an average accuracy of over 90%, which is around 10% higher than than the next closest result for the SIF, and is a 23% improvement over the multi-conditional MFCC-HMM baseline technique. Overall, the SPD-IF equals or outperforms the best performing baseline at every noise condition, with the proposed SPD noise estimate and Hellinger kNN classification system contributing significantly to the demonstrated performance.

4.3.3 Discussion

Several interesting aspects of the SPD-IF system are now discussed, including the effect of the spectral content of sound events, the missing feature classification system, and how the SPD-IF can be applied in an online recognition system.

Spectral Content of Sound Events

While the SPD captures the long-term temporal statistics of the sounds through the subband distribution, it does not explicitly model the temporal structure of the sound. Therefore, it may be possible to generate artificial sounds, such as simple upward and downward transients, that have similar SPD representations but could be distinguished easily in the spectrogram. However, in practise, natural sound phenomena rarely conform to such simple examples, as they typically have distinct increase/release energy cycles, which may be different across frequency subbands. Therefore, classification using the SPD-IF is able to distinguish between a wide variety of natural sounds, which is demonstrated by the experiments on a large database containing 50 classes

	SPD-IF	**Conventional Techniques**	
		Marginalisation	Imputation
Classifier	kNN	HMM-GMM	
Distance Measure	Hellinger	Mahalanobis	
Mask	Two-dimensional	Frame-by-frame	
Missing Features	Remove unreliable dimensions		Replace
Bounded	No	Yes (optional)	
Computation	Low	High	

Table 4.3: Comparison between missing feature approaches

of sound events. The SPD-IF also performs well in noisy conditions, provided that the sound spectrogram contains a few characteristic, high-power components that can be mapped to give a reliable region of the SPD for classification. Due to the physical nature of sounds and noise, this should be the case down to very low SNRs, as the noise energy is diffuse across the spectrum. This is demonstrated by the performance of the SPD-IF in 0dB non-stationary noise conditions, where it achieves a classification accuracy of over 90%.

Missing Feature Classification Systems

Table 4.3 shows a comparison between the proposed SPD-IF missing feature classification approach and the two conventional techniques that can be applied for HMM-GMM. It can be seen that the SPD-IF approach is fundamentally different, as a single feature is extracted from the two-dimensional SPD image to represent the whole sound clip. This is opposed to the conventional HMM-GMM methods that operate frame-by-frame in the time-frequency spectrogram domain. One advantage of the SPD-IF system is that the mask estimation is much simpler in the SPD domain compared to the spectrogram, and hence the approach can overcome the key drawback of conventional missing feature classification. Another advantage is the low computational complexity, since the HMM-GMM approach must evaluate the whole model at each time frame, and also incurs a considerable cost in integrating out the missing feature values. Together, the SPD-IF can achieve an improved classification performance with a low computational complexity, as shown in the previous experiments.

Online Recognition System

Many applications of SER require real-time performance. For investigation of the performance of the SPD-IF in real-time systems, it has been implemented in C#. Here, it is found that the feature extraction and classification of the SPD-IF, with 50 sound classes, runs around 10 times faster than real-time on a 2.0GHz Intel Core 2 Duo processor. This is comparable in speed to the MFCC-HMM baseline using HTK. For comparison with conventional missing feature approaches, the implementation of missing feature marginalisation in the CASA-Toolkit (CTK) [247] was found to run at around one times real-time on the same experiment. This means that a one second sound clip requires a further one second before the classification result becomes available. This may not be acceptable depending on the application.

An additional advantage of the SPD-IF for online systems is that retraining is exceptionally fast, since the kNN classification algorithm only requires feature extraction and storage. It also does not require a large number of samples for training, which makes the SPD-IF suitable for applications requiring a quick initial training setup, followed by additional training to be added later. This is not possible with HMM methods, which require the whole model to be recalculated, and in general require more time and data for training. In addition, there are algorithms which can speed up high-dimensional kNN searches, and while these are not employed here, they can solve the potential problem of having a very large training database.

4.4 Summary

This chapter proposed the subband power distribution (SPD) image representation, to improve upon the previous work on the SIF to represent sounds through a two-dimensional image feature. The SPD is a two-dimensional representation of the distribution of normalised spectral power over time against frequency, where the temporal information is captured implicitly through the distribution information. The advantage of the SPD over the spectrogram is that the sparse, high-power elements of the sound event are transformed to a localised region of the SPD, unlike in the spectrogram where they may be scattered over time and frequency. This enables a missing feature mask to be easily applied to the SPD-IF, and the missing elements of the image fea-

ture can be marginalised in a kNN missing feature classification system. Experiments were carried out to validate the proposed approach, and compare the results to those previously achieved by both the SIF and the best performing baseline techniques. This demonstrated that the SPD-IF is both discriminative in clean conditions, and robust to noise, achieving an average classification accuracy of almost 96%, which is a significant improvement over the baseline techniques. The biggest improvement in performance was in severe noise, where the SPD-IF achieved an accuracy of over 90% in 0dB noise, which is a 23% improvement over the best-performing multi-conditional MFCC-HMM baseline.

Chapter 5

Simultaneous Recognition of Overlapping Sounds

In this chapter, we turn our attention to the challenging task of simultaneous recognition of overlapping sound events from single channel audio. This problem naturally occurs in the unstructured environments found in SER tasks, in addition to the problem of noise robustness that has been studied in the previous chapters.

The approach taken in this chapter is motivated partially by the limitations of the current state-of-the-art techniques, in particular those of frame-based approaches where each time frame contains a mixture of information from multiple sources that is difficult to separate. Additional inspiration comes from the field of image processing, where the problem of object detection in cluttered environments can be seen to have many similarities with detecting overlapping sounds embedded in background noise. Together, this leads to the development of a solution based on local spectrogram features (LSFs) in the spectrogram, which capture the joint spectro-temporal information surrounding keypoints detected in the spectrogram [8–10].

The chapter is organised as follows. Section 5.1 first provides the motivation for using spectrogram image processing to address the challenging task of overlapping SER. Section 5.2 then introduces the proposed LSF recognition system based on local features and the generalised Hough transform (GHT) detection mechanism. Experiments are then carried out in Section 5.3 to evaluate the performance on an overlapping SER task with noise and channel mismatch.

5.1 Motivation

Acoustic information occurring in real-life unstructured environments is unlikely to be captured as a stream of isolated sound events. Hence, it becomes a challenging task to detect and segment overlapping sounds from a single continuous audio stream. Many state-of-the-art SER systems are not designed for this purpose, since they generate a model of the sound that assumes each event will occur in isolation. Those that do address the problem, such as Factorial HMMs (FHMMs) [248, 249], often have limitations such as computational complexity or a fixed number of overlapping sources.

It should be noted that an alternative approach to the problem of overlapping signals is to make multiple simultaneous recordings of the auditory scene using a microphone array. This enables multi-microphone techniques, such as beam-forming [76] and statistical independence [250], to be used. However, such techniques often require assumptions about the nature of the environment, and the performance may also degrade rapidly in noise [251]. In addition, the methods are often designed to recover the sources for human listening, and therefore may contain considerable distortions that make them unsuitable for the task of SER. Therefore, these multi-microphone techniques are not considered any further, since the focus in this chapter is the arguably more challenging task of simultaneous recognition of overlapping sound events from a single microphone.

The rest of this section now discusses the motivation for addressing the problem of overlapping SER, including the limitations of the state-of-the-art and inspirations that can be found from both human hearing and image processing.

5.1.1 Problem Description

Two sounds that occur simultaneously will be received as a mixture of the two by a listening device such as a microphone. This happens frequently in practically every real-life environment, including applications occurring in indoor spaces such as offices or meeting rooms, and outdoor places such as restaurants or train stations. For a human listener, the task of separating and recognising these overlapping sounds is intuitive and simple. For speech, it is commonly referred to as the "cocktail party effect", where a person is able to follow a particular conversation from a room filled

with competing speech and sounds [252].

For sound events, the problem remains essentially the same. For example, an acoustic surveillance system may have the task of detecting gunshot sounds from a crowded public place with a range of background noise. However, the problem of overlapping sounds is often overlooked in most SER systems, where the more general problem of noise robustness is more thoroughly researched. An example of this trend can be seen in results for CLEAR evaluation of sound events occurring in the meeting room environment [13, 27]. Here, it was found that overlapping segments were responsible for 70% of the errors produced by the majority of the competing systems [38]. This result should have been expected, since the majority of the systems were not designed for anything other than isolated sound events. This is despite the significant implications of developing a system that can achieve simultaneous recognition of overlapping sound events in practical applications.

The difficulty faced by conventional systems for SER is that the training often occurs in a controlled environment, without any examples of overlapping sound events. This means that the sound event model captures only information about the sound in isolation, and hence overlapping sound events will produce a low score against all of the trained models. In addition, most conventional systems use frame-based features such as MFCCs, which represent a complete slice of the frequency spectrum at each increment in time. This leads to a problem, since the combined spectral information is made up of contributions from each source according to the MixMax interaction between the two signals [98], as previously detailed in equation (4.1). The resulting feature therefore represents a mixture of the different sources, which can be extremely difficult to separate using conventional mask estimation methods without prior knowledge of one of the signals.

An example of the overlapping problem is given in Fig. 5.1. This shows the spectrogram of three sound events – bell, horn and whistle – which are strongly overlapped in time, and additionally are captured in the presence of background noise. The spectrogram shows that the harmonic of the bell becomes completely overlapped by the stronger whistle sound at around frame time $t = 40$, such that the bell sound is completely masked and can no longer be heard. In addition, this mixture is overlapped with a horn sound between frames $t = 30 - 100$, although most of the horn's energy is occurring at a different frequency. This type of challenging scenario is the focus of

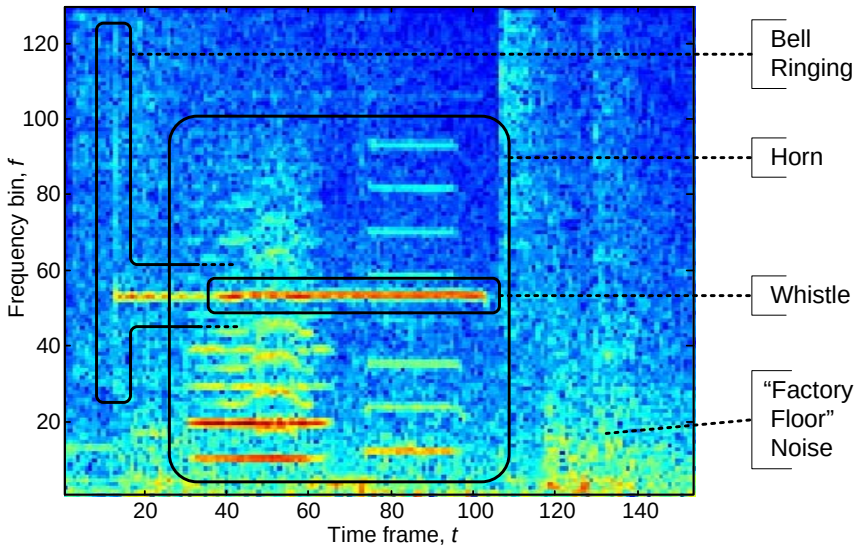
this chapter.

5.1.2 Limitations of the State-of-the-Art

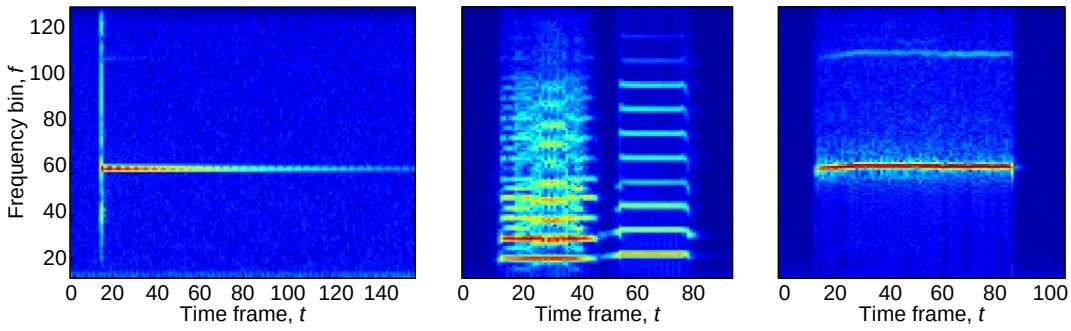
Although the topic of simultaneous recognition of overlapping sound events from a single audio stream is important, it has not been the subject of extensive research [38]. The small number of previous works that do exist can be separated into three different methodologies: direct classification, spectrogram decomposition and computational auditory scene analysis (CASA). These approaches are introduced in this section, along with a discussion of their limitations that may reduce their effectiveness in practical applications.

Direct Classification This is where a conventional acoustic modelling technique is used, but either the training or testing approach has been modified to compensate for the presence of overlapping sound events. An early example of this, originally developed for overlapping speech in ASR, is Factorial HMMs (FHMMs) [248, 249]. A conventional GMM-HMM model is used, but the Viterbi decoding process is modified to find the best combination of hidden mixture states to explain the observed feature. It can also be simplified as a MixMax-GMM, which is equivalent to FHMM but using only one hidden state [253]. The disadvantage is that the combinatorial nature of the problem results in extremely high computational complexity, which limits the number of simultaneous sources that can be recognised in practice.

A simpler approach is therefore to modify the training to include a category containing different combinations of overlapping sound events, then perform conventional classification [254]. This is the approach taken in [38], where the first SVM classification stage assigns the input as either isolated events or a combined "overlapped" class. This is then expanded in a second hierarchical SVM to identify the overlapped combination. The disadvantage of this approach is that it requires sufficient training samples of the overlapped sounds in advance. These should cover each possible degree of overlap, hence may not always be available. It also requires a different class for all the possible overlapping sound event combinations, which in practice limits the number of simultaneous sounds



(a) Example of bell, horn and whistle sound events overlapping with non-stationary background noise. Here, the harmonic of the bell sound is covered by the whistle around frame 40, which in turn is completely overlapped with the horn which is present between frames 40-100.



(b) The three isolated sounds from the mixture above. Left to right: bell ringing, horn and whistle.

Figure 5.1: Example of three overlapping sounds in the presence of non-stationary background noise. This demonstrates the challenging problem of simultaneous recognition of overlapping sound events.

to just one or two due to the increasingly large number of combinations. Other approaches include detecting ridges in the spectrogram and extracting features by combining different segments to form hypotheses [255, 256]. An alternative idea is to fuse audio and video features to provide complimentary information during overlapping regions [39]. More recently, an approach has been developed to transform the probabilistic distribution of the subband information to a new domain that is sparse and additive [257]. The idea is that overlapping sound events form separate peaks in the new representation, such that SVM can be used to detect these peaks corresponding to the overlapping sound events, within a confidence interval. However, it is noted that this approach may not be suitable for all combinations of sound event classes, especially those which have similar subband spectral distributions.

Spectrogram Decomposition This is a form of blind source separation that uses factorisation to decompose the input signal into its constituent sources. Since the log-spectral values in the spectrogram can easily be made to be fully positive, the most common approach is to use non-negative matrix factorisation (NMF). This is an unsupervised decomposition similar to PCA, but with different constraints. The approach in [258] is to use NMF to decompose a spectrogram containing overlapping sound events into four components, where different sound events may be separated into different components for recognition. This is shown to improve upon a similar system that performs simple recognition without factorising the input audio stream [43]. A more recent NMF approach decomposes the frames in the spectrogram into a set of templates, such that the activation of these templates during testing can be used as a measure to detect the overlapping sound event classes [259]. The approach also applies additional constraints, such as sparsity, to the NMF to improve the decomposition. This can help to improve the factorisation during testing, particularly in cases where the system is presented with mixtures containing unknown sound events or noise. It is noted however in [258] that the problem of controlling the outcome of the factorisation is one of the major difficulties with the NMF approach, and is an ongoing topic of research.

Computational Auditory Scene Analysis This was previously introduced in Sec-

tion 2.2.2 as a state-of-the-art method for robust SER. However, it can also be applied to the problem of overlapping sound events. The idea is to generate a set of masks that can segment the spectrogram into regions corresponding to the different overlapping sources [20]. The segmentation is typically achieved by grouping the spectrogram elements based on their observed properties and cues, for example regions that share a common onset and offset time [21]. The masks can then be used in a missing feature recognition system [103, 104], with one pass required for each sound source to be recognised. As discussed in Section 2.2.2, the problem with such systems is that it is difficult to reliably generate the mask, and errors in this stage significantly affect the subsequent recognition. The problem of mask estimation is also magnified for simultaneous sound event sources, as the problem cannot be simplified by assuming a single source in the presence of background noise.

Together, each of the techniques discussed above has their advantages and disadvantages. Therefore the challenging task of simultaneous recognition of overlapping sound events remains an open and interesting research topic.

5.1.3 Inspiration from Object Detection

Although the spectrogram has important characteristics that make it different from conventional images, some of the fundamental problems faced in SER are similar to those in image processing. In particular, the problem of recognising sound events that are overlapped with other sound events and noise can be seen as being similar to the problem of detecting overlapping objects in a cluttered scene. This is the task of object detection, which has been extensively studied in the field of image processing.

An example of a typical object detection problem is given in Fig. 5.2. This image shows a scene containing several objects placed on a chair in the corner of a room. When compared to the spectrogram in Fig. 5.1, which contains three sound events overlapping in the presence of background noise, the following similarities can be observed:

location – in both cases the location of the object or the sound event is not known in advance, hence must be detected. It should be noted however that while

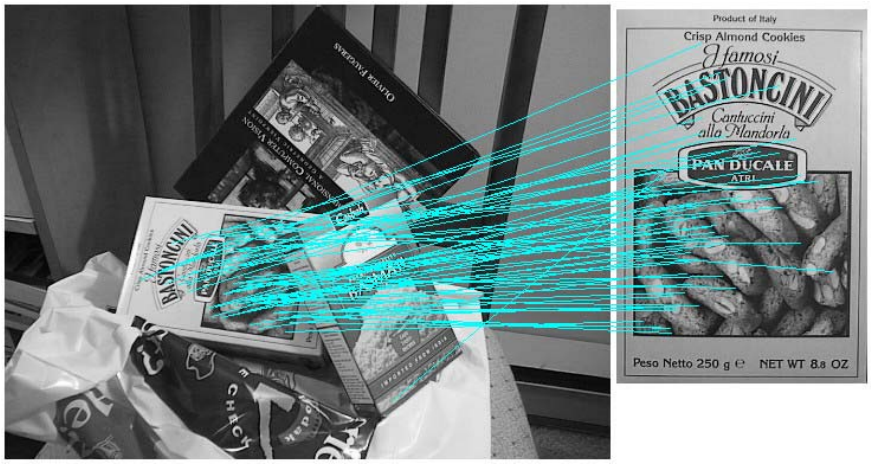


Figure 5.2: Example of the problem of object detection. Here, the box can be detected from the cluttered scene using the SIFT method from [207], where the blue lines indicate the matches with the training image.

the object location is typically two dimensional, the location of the sound in the spectrogram is largely constrained by the frequency content, hence only the onset of the sound in time needs to be detected.

occlusion – for conventional images there is occlusion when objects overlap with each other, such that the closer object to the viewpoint will physically obscure any that are further away. In the spectrogram, sound events merge with each other as opposed to occlude each other. However, due to the MixMax principle [98], the sound event with the highest energy in a given time-frequency region can be assumed to mask any lower energy sounds in a way that is equivalent to occlusion.

background – while the stochastic noise forms the variable background of the spectrogram, the background of a conventional image may also vary depending on the location of the scene. In both cases, the background is often less important than the sound event or object to be detected.

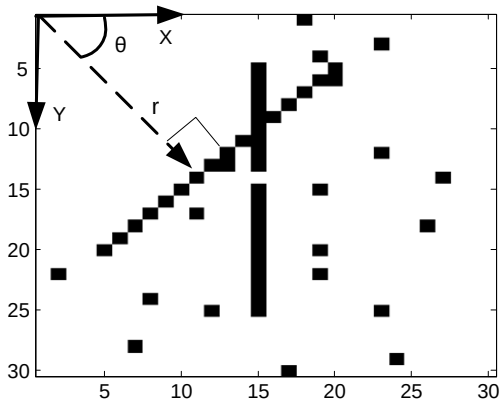
Each of these factors introduce challenges in designing a suitable recognition system. However, many of these have already been addressed by research in the field of object detection. An example of one such approach is shown in Fig. 5.2, where the SIFT

system from [207] can be used to detect the box in the scene. The method finds correspondences between the training image and the new scene, as demonstrated by the blue lines overlaid on the figure, and can allow for changes in size or rotation of the target object. The key to SIFT, and several other state-of-the-art approaches, is to use local features that are extracted from the area surrounding keypoints detected in the image [206, 217–219]. This is important, because an individual local feature is less likely to be affected by either occlusion or changes in the background, compared to one that characterises the pixel information across the whole image. The problem of recognition is then to find the geometrical correspondences between the sets of local features extracted in training and testing. One solution is called the generalised Hough transform (GHT), which can find the geometrical correspondences using a voting procedure based on the independent local features extracted from the image [177, 215]. The idea is that a distribution function can be associated with each local feature, and then used as a voting function that is summed in the Hough accumulator. A local maxima in the accumulator will then be generated if a number of local features belonging to the same object all vote for the same pose. As the GHT is fundamental in both SIFT, and several other state-of-the-art object detection systems [206, 218, 260], a brief review of the technique is now given below.

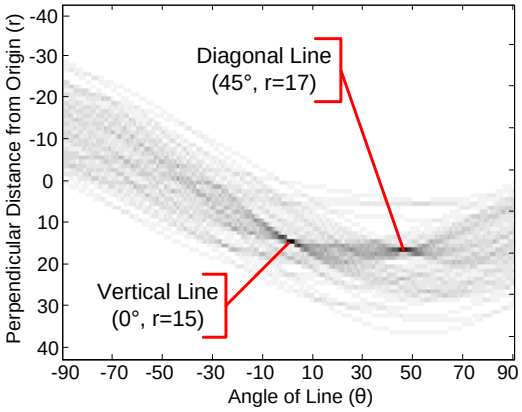
Hough Transform

The Hough transform was originally designed to detect parametrised lines and curves [177], and was only expanded later to cover arbitrary shapes through the GHT [215]. To understand the detection mechanism, consider the simple example shown in Fig. 5.3a, which contains two lines against a noisy background. Here, each point $P_i = [x_i, y_i]$ is considered as a local feature, and therefore casts votes into the Hough accumulator in Fig. 5.3b. The voting function is simply the distribution of all possible straight lines that the point could belong to, covering all possible rotations: $-90 < \theta < 90$. Using the polar equation of a straight line, the Hough accumulator, $H(r, \theta)$, is therefore as follows:

$$H(r,\theta) = \sum_{P_i} \begin{cases} 1, & \forall r = x_i \cos\theta + y_i \sin\theta \\ 0, & \text{otherwise,} \end{cases} \tag{5.1}$$



(a) Image containing two lines and noise.



(b) Hough accumulator after the straight line Hough transform.

Figure 5.3: Simple example of the Hough transform for overlapping straight lines in noise. The result is two strong local maxima in the Hough accumulator indicating the hypotheses for the two lines.

where r is the perpendicular distance from the origin and θ the angle from the horizontal axis, as shown in Fig. 5.3a. Local maxima in the Hough accumulator correspond to the combined evidence, from the individual points lying on the line, for a line hypothesis with a given (r, θ). Importantly, the Hough accumulator space is sparse and separable, such that two distinct hypotheses will not overlap, even if they are overlapping in the original image space. This is desirable for both object detection and overlapping sound event recognition.

The extension of the Hough transform to the GHT allows for the detection of arbitrary shapes that cannot be represented as an analytical equation. The GHT requires a codebook of local feature information to be learnt during training, which stores both the local feature template and a geometrical voting function. This voting function models the geometrical distribution of the codebook entry in the training images, relative to an anchor point, for each class of object to be detected. Then, during testing, the matched codebook entry for each local feature casts votes for possible locations of the anchor point into the Hough accumulator for each object class. Then, as before, the local maxima in the accumulator correspond to hypotheses for a particular object. This therefore maintains the key principles of the Hough transform, such as independent voting of local features and a sparse and separable accumulator space.

Extension to Overlapping Sound Event Recognition

A similar process can be used to develop a sound event recognition system that is robust to overlapping sounds, noise and distortion. This could use local features, similar to those used in object detection systems [206, 260], and then use their geometrical information to connect independent glimpses of sound events occurring across disconnected regions of the spectrogram [261]. However, it may not be appropriate to directly apply such object detection techniques to overlapping SER, due to the differences between the spectrogram and conventional images. Therefore, the following extensions can be used to take advantage of the information available in the spectrogram:

reference point – assuming that the sound event cannot shift in frequency, or rotate like an conventional object, the reference point can simply be the onset of the sound, since this often carries important information about the sound.

keypoint detection – this can be simplified, since the most important and reliable keypoints in the spectrogram are simply the sparse peaks of the sound event that carry the most energy.

three-dimensional modelling – the geometrical distribution of the local features (LSFs) can be modelled over time, frequency and spectral power. This captures more information than simply the two image dimensions, as it models the full trajectory of the sound including increasing and decaying spectral profiles.

local missing feature mask – as background noise may affect the spectral information in the LSF extracted from the keypoint, a local missing feature mask can be estimated to allow reliable matching with the codebook. Unlike traditional mask estimation, this can be estimated independently for each keypoint, which means that any errors in the noise mask will only affect a single keypoint and have a negligible effect on the overall result.

mixmax occlusion – regions of the sound event that are missing can be evaluated according to the MixMax relation between overlapping signals. This is different from object detection, where opaque occlusion between objects does not allow for such additional reasoning.

Combining these aspects can produce a more robust system for overlapping sound event recognition, as introduced in the next section. It also provides a novel and significant departure from conventional frame-based approaches that are common in the domain of audio processing. It also has an advantage over conventional HMM recognition systems, in that a sound can still be recognised even when a proportion of features is missing or corrupted due to noise or overlapping sounds. This is because the GHT is a summation of independent evidence, unlike HMM where the likelihoods are multiplicative, such that noise or overlapping sounds affecting one part of the feature has an adverse affect on the whole recognition. It also has further advantages compared to state-of-the-art techniques for overlapping SER, in that it does not require any assumptions about the possible overlapping combinations or require training on samples with different degrees of overlap between different sound events [38, 253, 257]. The idea of using local features from the spectrogram also has parallels with research into the human understanding of speech [262]. Here it is suggested that the human auditory system may be based on the partial recognition of features that are local and uncoupled across frequency. Together, this inspiration can therefore be used to develop a recognition system that is robust to noise and distortion, by connecting glimpses of sound events occurring across disconnected regions of the spectrogram [261]. This is the approach taken in this chapter, which is further developed in the next section.

5.2 Local Spectrogram Feature Approach

In this section, an approach for simultaneous recognition of overlapping sound events is introduced, based on the idea of using Local Spectrogram Features (LSFs). The aim is to overcome the limitations of the current state-of-the-art methods, which commonly have to make limiting assumptions about either the overlap between the sound events, the decomposition of the spectrogram, or the missing feature mask. The LSF approach is developed based on the concept of spectrogram image processing, and inspired by previous works on object detection. This section first gives an overview of the LSF approach, followed by a detailed description of each step in the algorithm, including keypoint detection, local feature extraction and the generalised Hough transform (GHT) recogniser that is key to the success of the system.

5.2.1 Overview

The LSF approach takes inspiration from several state-of-the-art image processing techniques that use local features combined with the generalised Hough transform (GHT) to detect objects in cluttered real-world scenes [171,206,260]. The advantage of such approaches is that local features are less likely to be occluded or distorted by other objects in the image, compared to taking a global or segment-based view of the image. When applied to sound event recognition, the idea is that each local spectrogram feature (LSF) should contain a glimpse of the spectral information coming from a single sound source [261]. This provides a key advantage over state-of-the-art audio processing techniques, where global or frame-based features can contain information from multiple sources that may be difficult to separate without prior knowledge of the overlapping sources. It is also a natural progression of the work in this thesis, by moving from the global view of the sound event image in the SIF and SPD methods, to a local view of the information in the spectrogram in the LSF approach.

The challenge then is to recognise the sound event based on a set of LSFs extracted from the spectrogram. This is solved through the use of the GHT, where each sound event is modelled through the geometrical distribution of the LSFs in the spectrogram, relative to the sound onset. During recognition, the GHT performs a summation of the independent evidence from each LSF into the Hough accumulator space. When a set of LSFs are similar enough to the geometrical distribution observed during training, a sharp peak will be observed in the Hough accumulator corresponding to the onset of the sound event. As there is a separate Hough accumulator for each sound event model, an arbitrary combination of overlapping sound events can be detected simply by finding local maxima in each accumulator space.

The LSF approach can now be summarised by breaking the approach down into separate training and testing processes. Fig. 5.4 gives an overview of the training process, which consists of the following steps:

1. *Local feature extraction*: in this step, "keypoints" are first detected in the spectrogram to locate characteristic spectral peaks and ridges. For each keypoint, an LSF and local missing feature mask are extracted to represent the local spectral region.

2. *LSF clustering*: this is performed to generate a codebook of local feature infor-

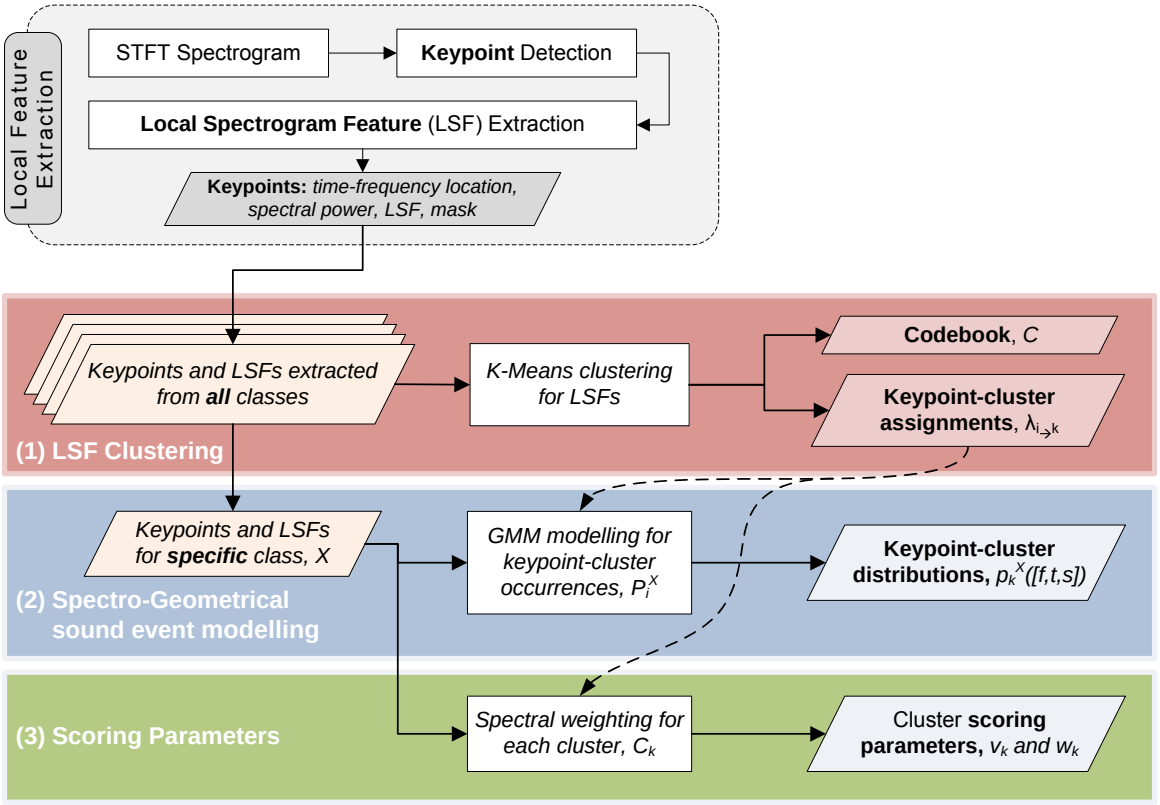


Figure 5.4: Overview of the geometrical sound event modelling used in the LSF approach. Here, the extracted LSFs are first clustered to form a codebook, then the geometrical distribution of each codebook is modelled over time, frequency and spectral power in a GMM.

mation.

3. *Geometrical sound event modelling*: each sound event is now modelled through the geometrical distribution of the codebook clusters in the training spectrograms over time, frequency and spectral power, relative to the sound event onset.

4. *Scoring parameters*: these are also extracted for each cluster to provide a threshold for verification during testing.

For recognition, it is assumed that the combination of sound events in each clip is not known in advance, and that the onset and relative magnitude of the sound events may have changed between training and testing. Hence, it is not possible to directly fit the geometrical distribution models learnt during training to the observed

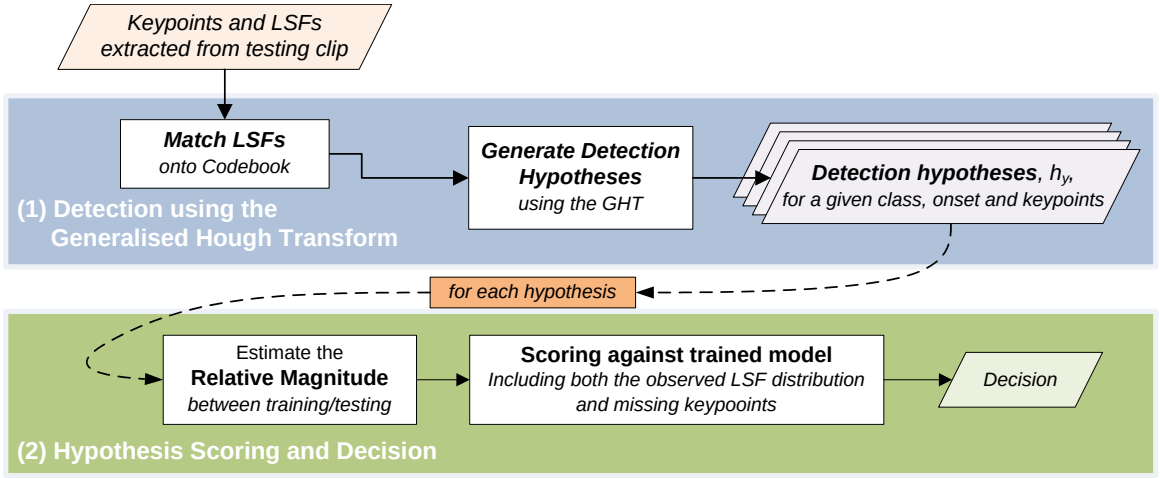


Figure 5.5: Overview of the proposed LSF recognition system.

keypoints, without first detecting the onset of the sound. However, the clustering in step (2) of the training enables each LSF to be represented by the information contained in the closest matching codebook entry. In particular, the geometrical distribution information associated with the codebook entry can be used as a voting function to provide a detection mechanism using the GHT. This must be followed by further scoring and verification that can take account of changes in magnitude between training and testing, and any keypoints that may be missing due to overlapping sound events. The testing process therefore consists of the following steps, as outlined in Fig. 5.5:

1. *Detection using the GHT*: the LSFs are first matched onto the codebook to provide a mapping to the geometrical distribution information learnt during training. This is then used as a voting function for the GHT, which generates sound event hypotheses by finding local maxima in the Hough accumulator space. Each hypothesis consists of a class label, the estimated onset time of the sound event, and the set of keypoints that contributed to the hypothesis.

2. *Hypothesis scoring and decision*: each hypothesis is examined by first estimating the relative magnitude of the sound event between training and testing. This is performed using a second GHT based on the conditional distribution of the observed log-spectral power difference given the hypothesised sound event onset. Finally, the keypoints that contributed to the hypothesis can then be scored

against the full sound event model over time, frequency and spectral power. Missing keypoints are also allowed to contribute to the hypothesis score by using prior information about each cluster that is learnt in training. Together, the output score is compared to the thresholds obtained during training to produce a decision.

The idea is that a sound event can only be recognised if both the LSFs match to the correct codebook clusters, and the keypoints have the same geometrical distribution as found during the training. Then, even if random fluctuations in the stochastic noise incorrectly match a particular codebook cluster on the background, this will not generate a strong hypothesis as they will not combine with other keypoints to match the localised geometrical distribution of the clusters found during training. The rest of this section now describes each of the training and testing steps in detail.

5.2.2 Local Spectrogram Feature Extraction

Extraction of Local Spectrogram Features (LSFs) is based on the log-power Short-Time Fourier Transform (STFT) representation of the sound, $S(f,t)$, where f represents the frequency bin and t is the time frame. Each spectrogram, $S(f,t)$, is of size $F \times T$, where the frequency dimension, $F = 129$, is determined by the sampling frequency of 16kHz and a 16ms time window with a 50% overlap. The time dimension of the spectrogram, T, varies according to the length of the clip, with approximately 125 frames per second. The LSF extraction then consists of two steps: (1) detecting keypoints in the spectrogram and then (2) characterising the local region surrounding each keypoint with an LSF.

In image processing, many previous methods have been developed for both keypoint detection [165] and local feature extraction [213]. However, these approaches may not be the best for the spectrogram, as there are important differences between spectrograms and conventional images, as discussed previously in Section 3.1.2. In particular, the intensity of each cell in the spectrogram is directly related to the energy and information contained in the signal. This is unlike in image processing where local gradients are typically more important than intensity or colour values [165]. Therefore, keypoints can be detected simply by finding the local maxima in the spectrogram that

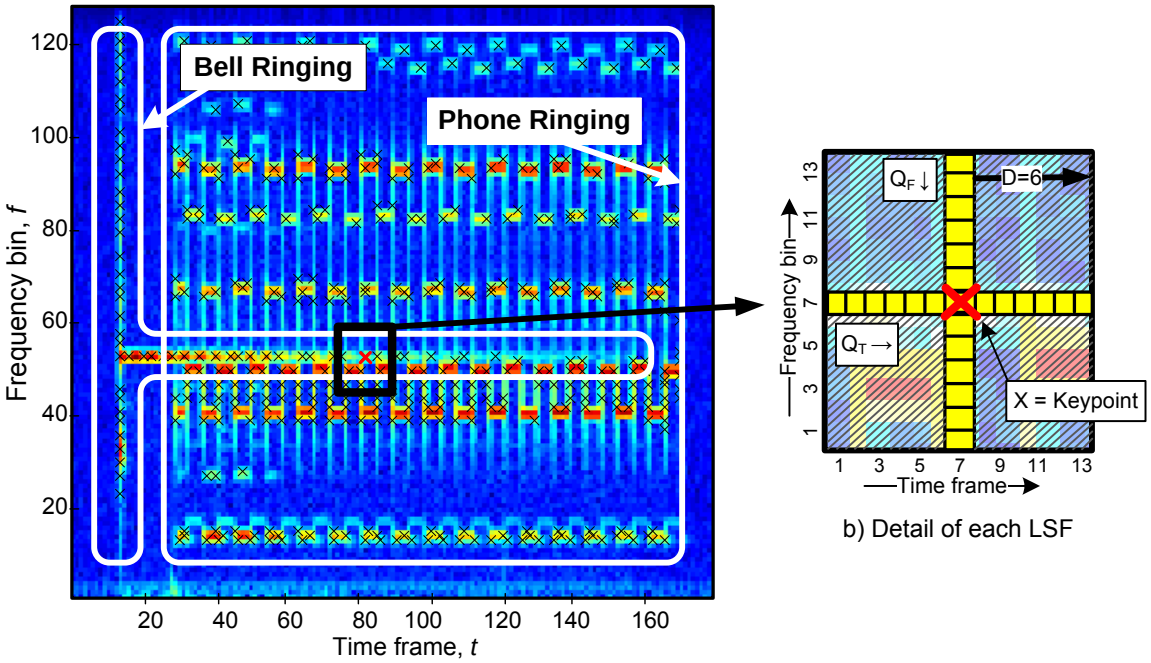
correspond to the characteristic sparse peaks in the sound event. A local SNR can also be used as a measure for filtering out the less important keypoints.

The sound information can also be characterised directly in a feature based on the local spectral region surrounding each keypoint. Therefore, both keypoint detection and LSF extraction can be combined into a single extraction step that is based on the same local spectral information. The shape of the local region is then important for defining the local information that is extracted from the spectrogram. In particular, it is important to ensure that the LSF contains information from only a single sound source, even in the case of overlapping sounds, as otherwise the LSF may not match against the correct codebook entry. Therefore, it is not suitable to capture the surrounding 2D region using a square or circle shape as this will contain information from overlapping sounds. An example of this can be seen clearly in Fig. 5.6a, where the bell and phone ringing sounds are overlapped. Therefore, here it is proposed to use a "plus-shaped" local region to form the basis for both the keypoint detection and LSF extraction, as shown schematically in Fig. 5.6b. The idea is that the plus shape is composed of the local spectral and temporal shape separately, such that it gives a "glimpse" of the local spectrogram information in two dimensions [261]. This ensures that keypoints can be detected on both short impulsive sounds, which appear as vertical lines in the spectrogram, as well as on harmonic sounds that appear as horizontal lines. It is also more likely to capture information from a single sound from an overlapping mixture due to the spars nature of the sound events. An example of this can be seen for the bell sound in Fig. 5.6a. Here, although the local region is dominated by the phone sound, the keypoint highlighted in red can still be detected on the harmonic of the bell, and the extracted LSF provides a glimpse of the bell sound from the overlapping mixture.

The plus-shaped local region for the LSF is composed of the local horizontal and vertical spectral shapes within a radius D of the central point, as follows:

$$\begin{aligned} Q_T(f,t,d) &= S(f,t+d) \\ Q_F(f,t,d) &= S(f+d,t) \end{aligned} \qquad \text{for } -D \leq d \leq D \tag{5.2}$$

where Q_T and Q_F separately capture the local time and frequency dimensions respectively. From preliminary experiments it was found that $D = 6$ was small enough to



a) Spectrogram showing a bell and phone sound event overlapping.

Figure 5.6: Example of bell and phone ringing sounds overlapped, where × represents the detected keypoints. The highlighted region on the right gives an example of the proposed plus-shaped LSF, where the yellow boxes indicate the local horizontal and vertical spectral information that is used to form the feature. In the example shown, the LSF is able to provide a glimpse of the bell from amongst the mixture with the phone sound.

localise the spectral peaks, but large enough to provide a feature for clustering, hence is used throughout. With a radius of $D = 6$, each LSF represents a local region with a frequency range of approximately 800Hz and a time window of 100ms.

The task of keypoint detection is then defined as selecting a set of points in the spectrogram as follows:

$$P_i = \{f_i, t_i, s_i\}\,, \tag{5.3}$$

where i is the index of the keypoint and $s_i = S(f_i, t_i)$ is the log-spectral power. These are detected at locations that are local maxima across either frequency or time, subject to a local signal-to-noise ratio (SNR) criterion, as follows:

$$S(f_i, t_i) \geq \begin{cases} \max\limits_{d} \Big(Q_T(f_i, t_i, d) \Big), & or \\ \max\limits_{d} \Big(Q_F(f_i, t_i, d) \Big), & and \\ \eta(f_i, t_i) + \delta_{SNR} \end{cases} \quad \forall\ 1 \leq |d| \leq D \tag{5.4}$$

where $\eta(f, t)$ is the local noise estimate, and $\delta_{SNR} = 5dB$ is the local SNR threshold [103] that must be exceeded for the keypoint to be detected. The local noise estimate is generated as follows:

$$\eta(f, t) = \frac{1}{2D} \min \Big(\sum_{1 \leq |d| \leq D} Q_T(f, t, d), \sum_{1 \leq |d| \leq D} Q_F(f, t, d) \Big). \tag{5.5}$$

This represents the minimum of the two means over the horizontal and vertical local spectral dimensions, such that an approximate local estimate of the noise can be obtained.

For each detected keypoint, P_i, an LSF, L_i, is extracted from the local spectrogram region surrounding the detected keypoint. Here, the normalised plus-shaped local region is used, such that the LSF characterises only the local spectral shape and not the magnitude of the sound. This is important, as the magnitude of the sound may vary between training and testing. Instead, the magnitude information is captured in the geometrical distribution model of the sound, which is described later. The proposed LSF can therefore be written as follows:

$$L_i(d) = \left[\frac{Q_T(f_i, t_i, d)}{s_i}, \frac{Q_F(f_i, t_i, d)}{s_i} \right], \quad \forall\ 1 \leq |d| \leq D \tag{5.6}$$

where $s_i = S(f_i, t_i)$ is the spectral power at the keypoint. This is concatenated into a single vector, $L_i(z)$, of length $4D$, where $z = 1, \ldots, 4D$ is a new variable introduced to denote the dimensions of the LSF vector.

As noise may distort the local spectral information between training and testing, a missing feature mask, M_i, is also extracted for each LSF as follows:

$$M_i(z) = \text{sign} \left(L_i(z) - \frac{\eta(f_i, t_i)}{s_i} \right) \tag{5.7}$$

where $z = 1, \ldots, 4D$ is the variable representing the LSF dimensions and $\eta(f, t)$ is the

local noise estimate from (5.5). Note that $M_i(z) = -1$ denotes the unreliable LSF dimensions.

5.2.3 Geometrical Sound Event Model

The next step is to train a geometrical model of each sound class, based on the keypoints and LSFs extracted from the spectrogram. This process consists of three steps, as shown previously in Fig. 5.4: (1) all extracted LSFs are clustered to form a codebook of local spectral information that is independent of the sound class, and then (2) the geometrical distribution of keypoints assigned to each cluster are modelled for each sound class to provide a voting function for the GHT during recognition, and finally (3) several scoring parameters are extracted to provide a threshold to control the verification of matches during testing. While this algorithm is inspired by the object detection approach of [206], unlike in image processing it is possible here to model the sound geometry over three dimensions: frequency, time and spectral power. This captures the full trajectory of the sound in the spectrogram, and enables the approach to distinguish between rising and falling tones occurring with the same LSF patterns.

Codebook Clustering

For this first step, K-means clustering is used, where the output is a set of K codebook entries, C_k, where $k = 1, \ldots, K$, such that $\lambda_{i \to k}$ denotes the assignment of LSF L_i to cluster C_k. Each dimension of the codebook entries, $C_k(z)$, is modelled as a Gaussian distribution, with the mean, μ_k, and variance, σ_k^2. The idea is to capture information about the distribution of spectral values associated with each dimension of the LSF codebook entries. This enables the use of missing feature marginalisation to perform robust matching when noise corrupts the local signal information. The codebook entries are calculated as follows:

$$\begin{aligned} \mu_k(z) =& \frac{1}{n_k} \sum_{\lambda_{i \to k}} L_i(z) \\ \sigma_k^2(z) =& \frac{1}{n_k} \sum_{\lambda_{i \to k}} \Big[L_i(z) - \mu_k(z) \Big]^2 \end{aligned} \tag{5.8}$$

where $z = 1, \ldots, 4D$ are the LSF dimensions, and n_k represents the number of LSFs assigned to cluster C_k. The number of clusters, K, must be chosen such that the codebook is able to model the LSF patterns sufficiently well. In preliminary experiments, it was found that as long as K is large enough, for example $K = 200$, the performance did not vary significantly and the clustering produced compact clusters with a small variance. Hence this number is used throughout.

Geometrical Keypoint Distribution Model

The next step is to model each sound class, X, through the geometrical distribution of the observed keypoints, P_i^X, assigned to each cluster in the training samples. For this, a reference point for the temporal distribution of keypoints is required, as this normalises the distribution information from different samples of the same sound class. In object detection, typically a central point on the object is used as a reference point. However, for sound events, the onset time is a more suitable reference point, since many sounds are impulsive and carry important information at the onset. Therefore, the keypoints in each training sample are first normalised to have the same onset, such that:

$$t'_i = t_i - t_{ON}, \tag{5.9}$$

where t_{ON} is the onset time. As clean isolated samples are used for training, the first keypoint detected in each sample is used as the sound onset.

The geometrical cluster occurrence distribution of sound class X is then modelled over frequency, f, time, t, and spectral power, s. This is achieved using a three-dimensional GMM probability density function (PDF), which can be written as follows:

$$p_k^X\left([f,t,s]\right) = \sum_{m=1}^{m_k} c_{km}\mathcal{N}\left([f_i,t'_i,s_i];\nu_{km}^X,\Sigma_{km}^X\right) \tag{5.10}$$

where m_k is the number of mixture components in cluster k, c_{km} is the weight of the m^{th} component and $\mathcal{N}([f,t,s];\nu_{km}^X,\Sigma_{km}^X)$ is a multivariate Gaussian model, with mean vector ν and covariance matrix Σ. The PDF is estimated using the algorithm from [263], which uses a Kurtosis-based mode splitting method that does not require any prior knowledge about the number of mixtures required to model each cluster.

A graphical illustration of the geometrical modelling is given in Fig. 5.7 for the bell

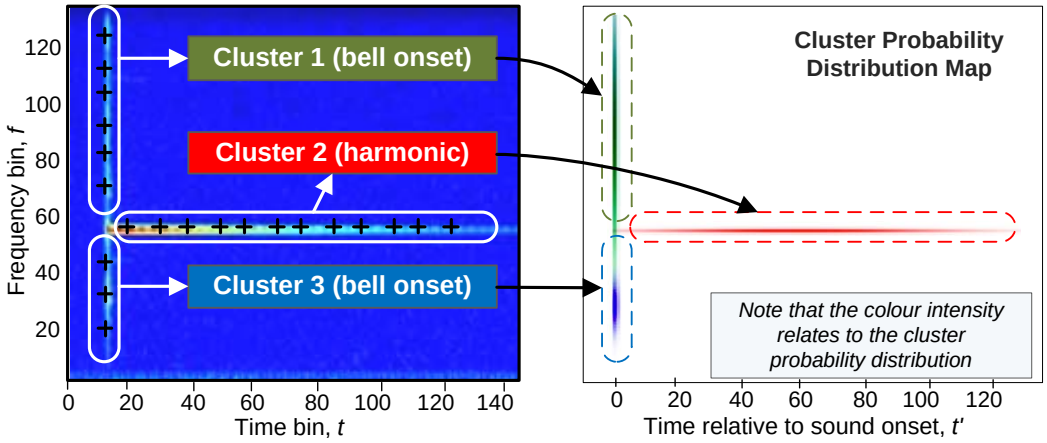


Figure 5.7: Example of the cluster geometrical occurrence distributions (marginal over time and frequency) for the top three clusters from the model of a bell sound.

ringing sound event that formed part of the previous overlapping mixture in Fig. 5.6. The figure shows both the matching of local features onto the codebook, and a model of their geometric distribution, for the top three clusters from the sound event model. It can be seen that both the onset and the harmonic of the bell can be neatly modelled by a small number of localised clusters, as it is found that the neighbouring LSFs have a similar spectral shape. This shows that it is possible to characterise sound events firstly through the matching of the LSFs onto a codebook, and secondly the geometrical distribution of the corresponding keypoints in the spectrogram relative to the sound onset.

Due to the fact that the codebook clustering takes place over LSFs extracted from all sound event classes, not every cluster in the codebook will appear for every class. Therefore, it is only necessary to model a subset of clusters that contributed the highest numbers of keypoints for the given class, n_k^X. This ensures that only the most consistently occurring clusters are used to model the sound. Here the cut-off for the number of keypoints is chosen such that the trained model explains 95% of the observed keypoints in the training samples.

Cluster Scoring Parameters

The final step in the modelling process is to extract scoring parameters, which are used as a threshold for hypothesis verification during testing. The first parameter is

the voting count of the cluster, v_k^X, which represents the average log-spectral power assigned to the cluster:

$$v_k^X = \frac{1}{N} \sum_{\lambda_{i \to k}^X} s_i \tag{5.11}$$

where N is the number of training samples provided for the class, X, and $\lambda_{i \to k}^X$ represents LSFs from class X assigned to codebook cluster, C_k.

The second is the cluster score, w_k^X, which represents the relative weight that the cluster contributes to the sound class:

$$w_k^X = \frac{v_k^X}{\sum_{k=1}^K v_k^X} \tag{5.12}$$

where K is the number of clusters in the codebook. Note that the spectral power is used, as opposed to simply the number of keypoints, as this gives more weight to the peaks in spectrogram, which carry more sound information than those with a lower spectral power.

During recognition, v_k^X is used as a cluster decision threshold, which must be obtained before the cluster k can be determined to have existed in the sound clip. Then, w_k^X is used to score the hypothesis by summing together the scores of the detected clusters. Since $\sum_{k=1}^K w_k^X = 1$ for each sound class, a threshold can be set for accepting a hypothesis to provide the desired tradeoff between false rejection and acceptance.

In addition to this, there may be keypoints missing during testing due to overlapping regions with other sound events or noise. Therefore it is desirable to enable information from these missing keypoint locations to contribute to the cluster score. The solution employed here is to calculate the expected spectral magnitude of each cluster at time-frequency keypoint locations that have the highest likelihood in the geometrical sound event model. This expected magnitude value is then used as a threshold during testing to determine if a keypoint may be missing. In particular, if the observed spectral magnitude at these locations is greater than the threshold, without a keypoint being detected, then the expected keypoint magnitude is allowed to contribute to the cluster score. This scoring process is detailed further in Section 5.2.5, while the expected magnitude of each cluster, $S_k^X(f,t)$, is calculated during training over a set of

time-frequency locations as follows:

$$S_k^X(f,t) = \begin{cases} \underset{s}{\text{argmax}}\ p_k^X\Big([f,t,s]\Big), & \text{if } p_k^X([f,t]) > \beta^k \\ 0, & \text{otherwise.} \end{cases} \tag{5.13}$$

where β^k is a likelihood threshold that is set such that only the most likely 75% of cluster locations are searched for possible missing keypoints during testing. This was to found to produce good results in preliminary experiments.

5.2.4 Detection using the Generalised Hough Transform

Given the sound event model from the previous section, the GHT is now employed to perform sound event detection. It does this by performing a summation of the distribution functions from each LSF-cluster match into the Hough accumulator. The idea is that all keypoints belonging to the same sound event in the spectrogram will share a common onset reference point. Hence, it is possible to accumulate evidence for the sound event based on the geometrical distribution models of each sound class. This is because the distribution will only have a maximum value at the correct onset time when the keypoint geometry matches the distribution found in the training.

A graphical illustration of this idea is shown in Fig. 5.8 for the example of a bell sound that is heavily overlapped with the sound of a phone ringing. The process consists of two steps: (1) match each LSF onto the codebook, and then (2) sum the cluster distribution functions in the Hough accumulator. The figure shows that the LSFs can be matched to the correct clusters from the geometrical model in Fig. 5.7, despite the overlap with the phone sound event. It can also be seen that strong evidence for the bell sound onset has been accumulated, as indicated by a sharp peak in the Hough accumulator corresponding to the onset time of the sound event. In this way, it allows the system to recognise an arbitrary combination of sound events in the spectrogram, including cases where two different sound events occur at the same time, or two instances of the same sound class overlap in each other with a small time offset. The two steps are now described in detail below.

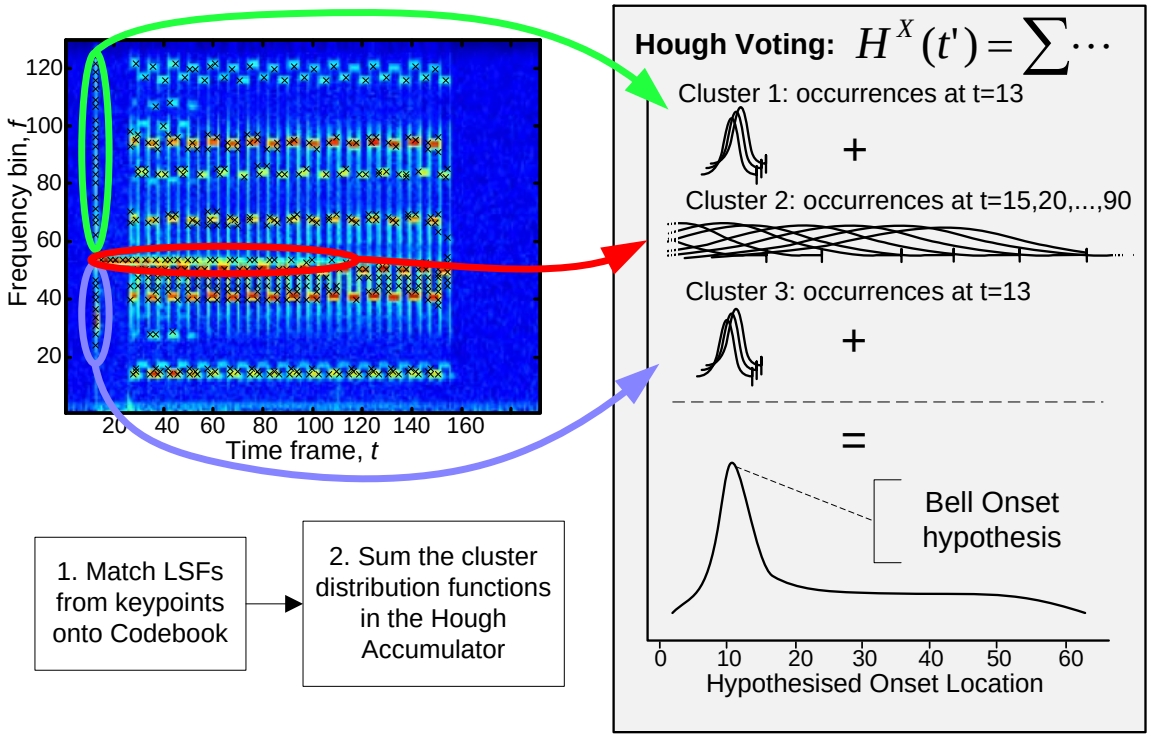


Figure 5.8: Schematic of the GHT voting process for the top three clusters of the bell sound shown in Fig. 5.7. The method proceeds by first matching the LSFs onto the codebook, then performing a GHT by summing the geometrical cluster distribution in the Hough accumulator.

Codebook Matching

The first step is to match the LSFs against the codebook generated in equation (5.8). The idea is to assign each LSF to the closest cluster, such that each keypoint can now be represented by the information contained in the codebook entry. In particular, the geometrical keypoint distribution model of the cluster, for each sound class, will later be used as a voting function for the GHT.

To ensure a consistent performance in mismatched conditions, each LSF, L_i, has an associated missing feature mask, M_i from (5.7), where $M_i(z) = -1$ represents the unreliable dimensions that may have been corrupted due to noise or overlap. To utilise this information, bounded marginalisation is performed against the codebook. This uses the observed spectral information as an upper bound for the probability distribution, and has been shown to perform better than completely marginalising

the missing dimensions [103]. Together, the reliable and unreliable dimensions sum together to give an overall log-likelihood score, $l_{i,k}$, as follows:

$$l_{i,k} = \sum_{z \in \{M_i(z)=1\}} \log \mathcal{N}\left(L_i(z); \mu_k(z), \sigma_k^2(z)\right) + \sum_{z \in \{M_i(z)=-1\}} \log \int_{-\infty}^{L_i(z)} \mathcal{N}\left(\Lambda; \mu_k(z), \sigma_k^2(z)\right) \, \mathrm{d}\Lambda \tag{5.14}$$

where Λ represents the normalised spectral power in the LSF, as calculated in (5.6). Each LSF, and its associated keypoint, P_i, is then assigned to the winning codebook cluster, as follows:

$$\lambda_{i \to k} = \underset{k}{\operatorname{argmax}} \ (l_{i,k}) \tag{5.15}$$

where $\lambda_{i \to k}$ denotes keypoint P_i assigned to cluster C_k.

GHT Detection Mechanism

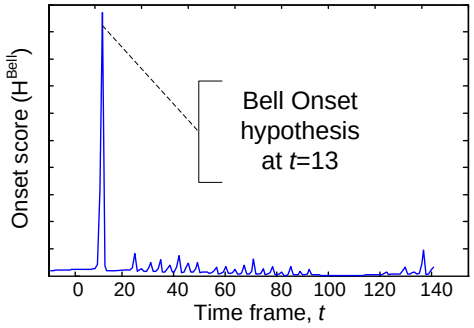
Given the assignments between each keypoint and a codebook cluster, $\lambda_{i \to k}$, the class-specific geometrical distribution model associated with the cluster, p_k^X, can now be used as a voting function for the GHT. For each observed keypoint occurring at time t_i, the distribution of the reference onset time, relative to the keypoint, can be written as $p_k^X([f, t_i - t, s])$. Since the relative magnitude between training and testing is still unknown, the marginal distribution over frequency and time is used as the voting function for the GHT. This voting function can be written as follows:

$$g_i^X(t') = p_k^X\Big([f, t_i - t] \mid f = f_i\Big) \tag{5.16}$$

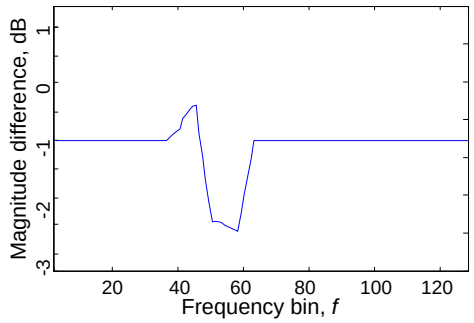
where p_k^X is the geometrical model from (5.10), f_i, t_i, are the frequency and time coordinates of observed keypoint, and $t' = t_i - t$ is the variable representing the hypothesised onset time relative to the keypoint.

The Hough accumulator, $H^X(t')$, for class X at time t', is then the summation of the voting functions as follows:

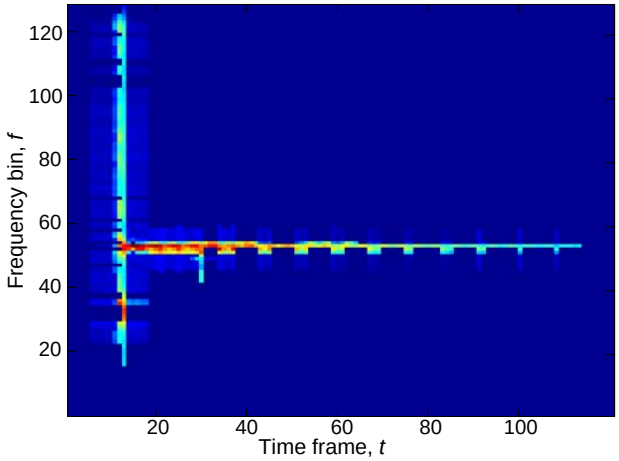
$$H^X(t') = \sum_{\lambda_{i \to k}} g_i^X(t') \tag{5.17}$$



(a) Example of the Hough accumulator for the bell sound from the mixture in Fig. 5.8.



(b) The relative magnitude estimated for the bell onset in (a) using a second GHT.



(c) The matching clusters for the recognised bell sound event. This demonstrates how the bell sound can be recognised from isolated glimpses between the overlapping regions of the sound.

Figure 5.9: Example of the output from the LSF recognition system for the bell sound extracted from the mixture of two sounds in Fig. 5.8. The detection hypothesis is indicated by the sparse peak in the Hough accumulator in (a) at $t = 13$, which is then scored against the trained model given the estimated relative magnitude from (b).

where $\lambda_{i \to k}$ are the keypoint-cluster assignments. Local maxima in the Hough accumulator correspond to the combined evidence over a set of independent keypoints for a sound onset. An example is shown in Fig. 5.9a, where the Hough accumulator for the bell sound clearly shows a sparse peak indicating the onset from the mixture. These local maxima can be found using the gradient-based approach of [264], since the Hough accumulator is a sum of weighted GMM distributions. Each local maxima generates a hypothesis, h_y, as follows:

$$h_y = \{X_y, t_{ON,y}, P_i^y\} \tag{5.18}$$

where y is the index for the hypothesis, and each hypothesis is specified by its class, X_y, onset time, $t_{ON,y}$, and set of contributing keypoints, P_i^y. A minimum threshold

is also set at 20% of the mean peak values obtained during training. This defines a minimum amount of observed evidence for a hypothesis to be generated. In addition, the contributing keypoints allow for a segmentation of the detected sound event to be generated. An example of this is shown in Fig. 5.9c for the bell sound from the mixture, and is generated by reconstructing the spectrogram using the codebook centres from the keypoint-cluster assignments.

5.2.5 Hypothesis Scoring and Decision

The output of the GHT detection is a set of sound event hypotheses, h_y, including their onset times and contributing keypoints. However, the relative magnitude transfer function of the sound between training and testing is still unknown, and must be estimated before the hypotheses can be scored against the full distribution model from training. The advantage of this step is that it enables the LSF system to detect sound events even when the mixing proportions are unknown, or in the presence of an unknown channel distortion. The scoring process is therefore as follows: (1) the relative magnitude transfer function between training and testing is estimated for each hypothesis, then (2) the hypotheses are scored against the sound event model, and finally (3) a decision is made by comparing the score against the thresholds from training. These steps are now described in detail.

Relative Magnitude Estimation

Assuming the sound is subjected to an unknown convolutive channel distortion in the time domain, this becomes additive in the log-power STFT domain, as follows:

$$S(f,t) + R(f,t) \approx S(f,t) + R(f) \tag{5.19}$$

where $S(f,t)$ represents the log-power STFT of a clean training sample, and R is the transfer function between the observed spectrograms in training and testing. Here, it is assumed that the response time of the channel is short, hence $R(f,t)$ can be approximated as $R(f)$, such that the channel distortion does not vary with time. Note however that the transfer function is permitted to vary across frequency, since this is common in many real-life room or microphone impulse responses.

The relative magnitude, $R(f)$, can be estimated using the trained sound model, p_k^X from (5.10). Since the onset for the hypothesis is already known, the conditional distribution of the difference in log-spectral power is used as a voting function for a second GHT. For each keypoint with index i, this is written as:

$$g_i^y(s') = p_k^{X_y}\Big([f, t, s_i - s] \mid f = f_i, t = t_i'\Big) \tag{5.20}$$

where $s' = s_i - s$ is the variable representing the spectral power difference and f_i, t_i', s_i is the three-dimensional location of the keypoint i in frequency, time and spectral power, where $t_i' = t_i - t_{ON,y}$ is the time relative to the hypothesised onset.

As in (5.17), the Hough accumulator, $H_f^y(s')$, is the summation of the voting functions. As $R(f)$ may vary over frequency, the summation of the voting function is performed separately for each subband, as follows:

$$H_f^y(s') = \sum_{\lambda_{i\to k}^y} g_i^y(s'),\ \forall f = f_i \tag{5.21}$$

where $\lambda_{i\to k}^y$ are keypoint-cluster assignments for hypothesis h_y.

The maximum value of the accumulator in each frequency subband can now be found, which corresponds to combined evidence for a given relative magnitude transfer function, $R(f)$, in that subband. This is written as:

$$R(f) = \underset{s'}{\operatorname{argmax}}\ H_f^y(s') \tag{5.22}$$

Note that $R(f) = 0$ represents an observed sound event with the same spectral power as the training, and that values are permitted to be positive or negative since this indicates a sound with a high or lower intensity relative to the training.

As some frequency subbands may contain very few keypoints, they do not represent reliable evidence to estimate $R(f)$. Therefore, subbands with less than 10% of the maximum are replaced with the mean of the transfer function across the remaining reliable values. Finally, the transfer function is smoothed using a moving average filter of radius $D = 6$, since it is expected that there will be a smooth variation in $R(f)$ over frequency. An example is shown in Fig. 5.9b, where the bell sound is estimated to be 1dB less than the training samples on average, with the bell's harmonic up to 2.5dB

quieter than in training.

Hypothesis Scoring

Starting with the hypothesis, h_y, that explains the largest number of keypoints, the next step is to evaluate the observed keypoints and spectrogram for the hypothesised class, X_y, and onset time $t_{ON,y}$ against the trained model from (5.10). However, due to the masking effect that occurs between overlapping sounds or noise, some keypoints and LSFs may be missing. Missing keypoints are defined here as regions of the cluster distribution where the expected magnitude, $S_k^X(f,t)$ as calculated in (5.13), is less than the observed magnitude, $S(f,t)$. To account for this, the scoring process allows missing keypoint locations to contribute to the cluster score. Therefore, the cluster scores are calculated separately for the observed keypoints, $v_{k,O}^y$, and the missing keypoint locations, $v_{k,M}^y$, where the indices O, M refer to the observed and missing keypoints respectively. The weighted sum of the scores is then used to find the final hypothesis score, which can be compared to the score obtained during training to make a decision, as described below.

First, the cluster voting score for the observed keypoints, $v_{k,O}^y$, is calculated by summing together the spectral power of keypoints that contributed to the hypothesis. This is analogous to the score obtained during training in (5.11), except keypoints are not considered if they have a likelihood less than a threshold. This ensures that low quality matches do not bias the output score, as these are more likely to have been matched on the background. The cluster score is therefore calculated as:

$$v_{k,O}^y = \sum_{\lambda_{i\to k}^y} \begin{cases} s_i', & \text{if } p_k^{X_y}\left([f_i, t_i', s_i']\right) > \gamma_k^{X_y} \\ 0, & \text{otherwise,} \end{cases} \tag{5.23}$$

where $\lambda_{i\to k}^y$ are keypoint-cluster assignments for hypothesis y, $\gamma_k^{X_y}$ is a likelihood threshold, and both $t_i' = t_i - t_{ON,y}$ and $s_i' = s_i + R(f_i)$ are set to align the keypoints with the trained model using the hypothesised relative magnitude and onset time. The threshold, $\gamma_k^{X_y}$, is set for each cluster based on the likelihood distribution of keypoints found during the training, such that 95% of keypoints are matched.

Next, missing keypoint locations that had a high likelihood in the training are

allowed to contribute to the cluster score. As the keypoint is missing, the vote is based on the expected keypoint magnitude for the cluster at that time-frequency location, as follows:

$$v_{k,M}^{y} = \sum_{S_k^{X_y}(f,t)>0} \begin{cases} S_k^{X_y}(f,t), & \text{if } S(f,t') > S_k^{X_y}(f,t) \\ 0, & \text{otherwise.} \end{cases} \tag{5.24}$$

where $t' = t - t_{ON,y}$ is the time relative to the hypothesised sound onset, and $S_k^X(f,t)$ is the expected cluster magnitude, as calculated in (5.13), that must be exceeded for the keypoint to be determined to be missing.

The final hypothesis score, $score(h_y)$, is then calculated as the sum of the cluster weights, w_k^X, but only for clusters that exceed a threshold in the voting score compared to training. This is calculated as follows:

$$score(h_y) = \sum_{k=1}^{K} \begin{cases} w_k^X, & \text{if } \left(v_{k,O}^{y} + \alpha_1 v_{k,M}^{y}\right) > \alpha_2 v_k^X \\ 0, & \text{otherwise.} \end{cases} \tag{5.25}$$

where v_k^X, w_k^X were found during training using (5.11) and (5.12), $\alpha_1 = 0.8$ is a weighting factor to balance the observed, $v_{k,O}^{y}$, and missing keypoints, $v_{k,M}^{y}$, and $\alpha_2 = 0.5$ is a threshold that defines the minimum cluster score value required for the cluster to be considered matched in the spectrogram.

Decision

If the hypothesis score exceeds a threshold:

$$score(h_y) > \Omega, \tag{5.26}$$

then the hypothesis is accepted. This threshold can be varied to control the tradeoff between false rejection and acceptance of hypotheses. In the following experiments, $\Omega = 0.5$ is used, such that at least half of the clusters must be matched for the hypothesis to be accepted. This was found to provide a good trade-off in preliminary experiments.

If the hypothesis is accepted, the given sound class X_y is considered to have been detected at the onset time, $t_{ON,y}$. The keypoints that contributed to the hypothesis

are then removed from further matches, and the next best hypothesis is evaluated until all valid hypotheses in the clip have been tested.

5.3 Experiments

In this section, experiments are conducted to evaluate the performance of the LSF system on a database of overlapping sound events in mismatched conditions. Several baseline methods are also implemented and evaluated to provide a comparison for the experimental results. The database is generated using a random overlap between different classes of sound events, and both the noise and volume of the sound events are adjusted to simulate real-world experimental conditions.

5.3.1 Experimental Setup

Database

As there is no standardised database of overlapping sound events, it was necessary to simulate the experimental database using sound samples from the same RWCP Sound Scene Database used in the previous chapters [159]. The following five classes are selected: horn, bells5, bottle1, phone4 and whistle1. The isolated sound event samples have a high signal-to-noise ratio (SNR), and are balanced to give some silence either side of the sound. The selected categories are chosen to provide a significant amount of overlap during testing. For example, the harmonics of the bell and whistle sounds occur at the same frequency, the impulsive onset of the bell and bottle sounds are similar, while both the horn and phone sounds span a wide time-frequency region of the spectrogram. Amongst the sounds, the bottle1 class contains the most variation, with five different bottles being struck by two different objects, although there is some variation across all classes.

For each event, 20 files are randomly selected for training and another 50 for testing. Given the $5 + \binom{5}{2} = 15$ overlapping combinations for testing containing either one or two sound events, this gives a total of 100 and 750 samples for training and testing respectively, with each experiment repeated in 5 runs. For overlapping samples, the onset times of the two sound events are randomly chosen for each clip to ensure that

the temporal overlap is between $50 - 100\%$ and that the order of the sound events is randomised amongst the testing set.

Evaluation Conditions

For each experiment, training is carried out only in clean conditions using the 20 isolated samples from the database. The performance of each method is then evaluated under the following conditions, which are chosen to better simulate real-world experimental conditions:

1. **Clean**: this is evaluated separately for both the isolated and overlapping sound event samples.

2. **Mismatched noise**: "Factory Floor 1" [160] noise is added to the testing samples at 20, 10 and 0 dB SNR. This noise is chosen for its challenging, non-stationary nature.

3. **Change of Volume**: the waveform of both sound events is pre-multiplied by one of the factors $\{0.5, 0.75, 1, 1.5, 2\}$ prior to combining them to form an overlapping sound event signal. This simulates a channel transfer function that is closer to the conditions observed in real-life applications.

As evaluation measure, the recognition accuracy (TP) and false alarm (FA) are calculated over each of the sound classes, over 5 runs of the experiment. TP is calculated as the ratio of correct detections to the number of clips containing occurrences of that class. Analogously, FA is the ratio of incorrect detections to the number of clips not containing that class.

Baseline Methods

For comparison with the proposed LSF approach, two state-of-the-art frame-based baseline classification approaches are implemented. The first method is called MixMax-GMM [265], which requires only isolated samples for training. The second is based on the approach of [38], and is referred to here as Overlap-SVM. This method requires both isolated and overlapping samples for training, hence provides an interesting comparison between the above methods. It is also notable that the LSF system performs

recognition as opposed to the simpler task of classification. However, the baseline classification methods are chosen to provide a well-performing benchmark and their results should represent an upper-bound of their equivalent recognition systems.

The MixMax-GMM is based on a mathematical combination of two GMM models, taking into account the MixMax approximation of two overlapping sound events [98]. The approach can be seen as a simplification of the full Factorial HMM approach using a single HMM state [265]. Therefore, for overlapped class $Z = X + Y$, where X and Y are two clean classes, the PDF, p_Z, of the overlapped class can be decomposed as follows [257]:

$$p_Z(\alpha) = p_X(\alpha)c_Y(\alpha) + p_Y(\alpha)c_X(\alpha) \tag{5.27}$$

where c_X is the cumulative density function (CDF) of class X and α represents the 36 dimension log-power Mel-frequency spectral coefficient (MFSC) features. Here, the PDF is modelled using a 6-component GMM, and the maximum log-likelihood, summed across all frames in the clip, is taken as the classification result.

The second baseline is referred to as Overlap-SVM, and is based on the approach proposed by [38]. The same frame-based features are used as in [38], consisting of a 16 dimension MFSC, plus deltas and accelerations, the zero-crossing rate, short time energy, spectral centroid, spectral bandwidth and 4 sub-band energies and spectral flux. The mean and variance of the 60-dimension frame-based features is taken over the clip, giving a final feature with 120 dimensions. This gives a total of 60 features for each frame, and the final feature is the mean and variance of these features across all frames in the clip, giving 120 feature dimensions in total. The method requires samples for each of the 10 overlapping combinations for training, which are generated from the isolated samples selected in the same way as described previously. The evaluation is then carried out using a two-stage SVM. The first stage classifies the clip into either the 5 isolated classes or an amalgamated "overlapping" class, containing all of the remaining 10 overlapping combinations. If the output of the first stage classifies the clip as overlapping, a second stage then determines the specific overlapping combination. Here, a conventional one-against-one SVM classification is used for the second stage as opposed to the tree-SVM structure used in the original method. This was done as it was found that there was insufficient training data to benefit from the full tree-structure.

Experiment Setup		Proposed LSF		Overlap-SVM		MixMax-GMM	
		TP	FA	TP	FA	TP	FA
Isolated	Clean	99.3 ± 2.7	0.4 ± 2.4	$\mathbf{100} \pm 0.0$	1.5 ± 3.4	99.6 ± 1.4	1.3 ± 5.8
Overlap		$\mathbf{98.0} \pm 3.4$	0.8 ± 3.6	96.5 ± 7.3	1.3 ± 2.8	84.0 ± 29.3	5.2 ± 17.0
Overlap: + Noise	20dB	$\mathbf{97.2} \pm 5.0$	0.7 ± 3.2	76.9 ± 39.0	18.6 ± 35.1	52.8 ± 44.9	27.8 ± 42.6
	10dB	$\mathbf{95.5} \pm 9.1$	0.9 ± 3.5	74.7 ± 40.9	20.9 ± 36.8	37.8 ± 42.9	25.1 ± 41.2
	0dB	$\mathbf{90.2} \pm 17.6$	2.5 ± 8.2	65.7 ± 41.9	25.8 ± 36.1	22.9 ± 38.8	20.9 ± 35.7
Overlap: + Vol. Change	$\times 0.5$	$\mathbf{98.1} \pm 3.0$	0.7 ± 3.3	84.0 ± 24.8	1.5 ± 5.0	56.0 ± 43.4	12.4 ± 27.8
	$\times 0.75$	$\mathbf{98.4} \pm 2.9$	0.5 ± 1.8	92.8 ± 13.1	1.1 ± 2.9	80.6 ± 30.0	4.4 ± 13.7
	$\times 1.5$	$\mathbf{98.4} \pm 2.7$	0.6 ± 2.1	95.9 ± 9.9	4.0 ± 11.4	82.0 ± 29.8	8.3 ± 21.6
	$\times 2$	$\mathbf{98.0} \pm 3.3$	0.7 ± 2.0	94.1 ± 14.7	7.0 ± 18.3	68.7 ± 40.7	23.7 ± 39.3
Average		**97.0%**	**0.9%**	86.7%	9.1%	64.9%	14.3%

Table 5.1: Experimental results across the various testing conditions. The values for TP/FA (%) are averaged over 5 runs of the experiments, with the standard deviation also reported ($\pm$). For the isolated experiment, the results are averaged over the 5 sound classes, while for the overlapping experiments, the results are averaged over the 15 overlap combinations. The entries in bold indicate the best result amongst the three methods in each row of the table.

5.3.2 Results and Discussion

The results from the experiments on the LSF recognition system are now presented and compared against those achieved by the baseline methods. First, the performance of the methods in clean conditions is analysed, with both isolated and overlapping results reported separately for comparison. Then, the performance is analysed in both mismatched noise and volume change. The results demonstrate the superiority of the LSF system across a range of experimental conditions, as described below.

Clean Conditions

The performance for each of the three methods in clean conditions is reported in the first section of Table 5.1. For isolated sounds, it can be seen that the proposed LSF approach performs well, achieving a TP of 99.3% for an FA of only 0.4%. Although the TP is marginally lower than the two baselines, it is still within 0.7% of the best Overlap-SVM baseline, which is a good result. In addition, the average FA for the

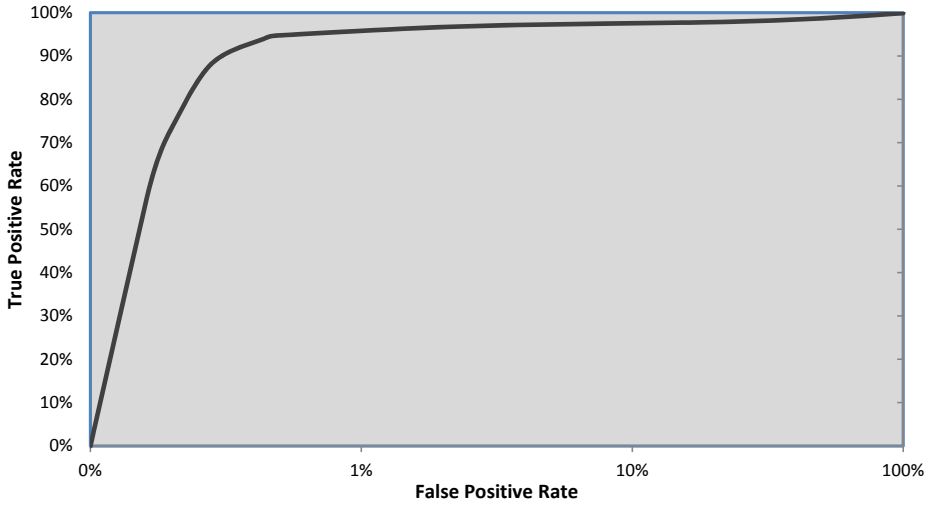


Figure 5.10: ROC curve showing the TP/FA experimental results in 10dB noise when varying the detection threshold Ω from (5.26).

LSF approach is an improvement of around 1% over the baselines, which indicates the other methods are less good at rejecting false matches.

For overlapping sounds, the LSF approach outperforms the two baseline methods, achieving a TP of 98.0% and an FA of only 0.8%. This is an improvement of 1.5% over the Overlap-SVM baseline, and significantly better than the MixMax-GMM approach, which achieves a TP of just 84.0% with an FA of 5.2%. The performance of the LSF system is also significant, considering that the Overlap-SVM baseline requires overlapping sounds samples for training. Therefore, while Overlap-SVM is performing classification in matched training and testing conditions, the LSF approach is performing recognition of the sound events and only requires isolated samples for training.

Mismatched Noise

The performance in mismatched conditions can be found in the second section of Table 5.1. The results show that the performance of both baseline methods declines rapidly with increasing noise. In particular, the MixMax-GMM method approaches a TP of 20% at 0dB, which is close to a random guess. However, the proposed LSF approach performs consistently well across all four conditions, and can still achieve a TP of 90.2% in 0dB conditions, for an FA of just 2.5%. This is an absolute improvement in

Sound Event	Horn	Horn				Bells	Bells			Bottle	Bottle		Phone	Phone	Whistle
		Bells	Bottle	Phone	Whistle		Bottle	Phone	Whistle		Phone	Whistle		Whistle	
Horn	**100**	**100**	**100**	**95.6**	**99.6**	0	0	10.8	0	0	8.4	0	10.8	6.8	0
Bells	0	**97.2**	0	0	0.4	**100**	**100**	**63.2**	**80.8**	0	0	3.2	0	0	0
Bottle	0	0	**98.4**	1.2	0	0	**96.0**	0.8	0	**100**	**82.2**	**96.8**	1.2	0.8	0
Phone	0	0	0	**100**	0	0	0	**100**	0	0	**100**	0	**100**	**100**	0
Whistle	0	0	0	0	**95.6**	0	0	0.4	**94.4**	0	0	**95.6**	0	**93.6**	**96.8**

Table 5.2: Detailed experimental results for the LSF method in 10dB Factory Floor noise, showing the results for each of the 15 overlapping combinations. The values (%) represent the percentage of clips with the detected sound event. Correct TP detections are highlighted in bold.

both TP and FA of over 20% compared to the best performing baseline method, where the Overlap-SVM method achieved a TP of 65.7%. Significantly, the Overlap-SVM also has an FA rate of 25.8% in the 0dB conditions, which is much higher than the proposed method, and would be intolerable in any practical system.

One reason for the poor baseline performance is that their frame-based features cannot separate the overlapping signals and noise that occur at the same time instance. To improve the results, it may be possible to use multi-conditional training with similar noise conditions, but given the wide variety of overlapping combinations this may not work in practise. Another way may be to use frame-based missing feature masks, however reliable estimation of the mask is challenging, as discussed previously in Section 5.1.2. The LSF system overcomes these problems by using features that are local across frequency, and by utilising a local missing feature mask, hence recognition can still be performed in the presence of competing signals.

Further analysis of the results for the LSF system is shown in the ROC curve in Fig. 5.10, which shows the performance of the system in 10dB noise when varying the threshold parameter Ω from (5.26). It can be seen that the system is able to achieve a high accuracy for a low false alarm, although some sounds are difficult to detect even with a low threshold value, as seen by the flattening off at the top of the ROC curve. This may be caused by the physical overlap between certain sounds that are difficult to separate and hence do not trigger a detection in the LSF system. This effect can be seen more clearly in Table 5.2, which shows the average performance for each of the

overlapping combinations at the default threshold of $\Omega = 0.5$. It can be seen that the LSF approach has the most difficulty identifying other sounds in mixtures containing the phone sound. In particular, the bell and bottle sound events are identified correctly 63.2% and 82.2% respectively when the phone sound was present. This result is because the spectral information in the phone sound is spread over time and frequency, as can be seen in the previous example in Fig. 5.8. This means that fewer keypoints for the overlapped class will be detected and assigned to the correct clusters, hence recognition is more difficult. In addition, the phone sound consistently caused false alarms for the horn sound in around 10% of the clips, as there was a sufficient number of keypoints incorrectly matched to produce the horn hypothesis. In future, this could be improved by enhancing the scoring criteria to reject detections that completely overlap other detections when there is insufficient evidence present.

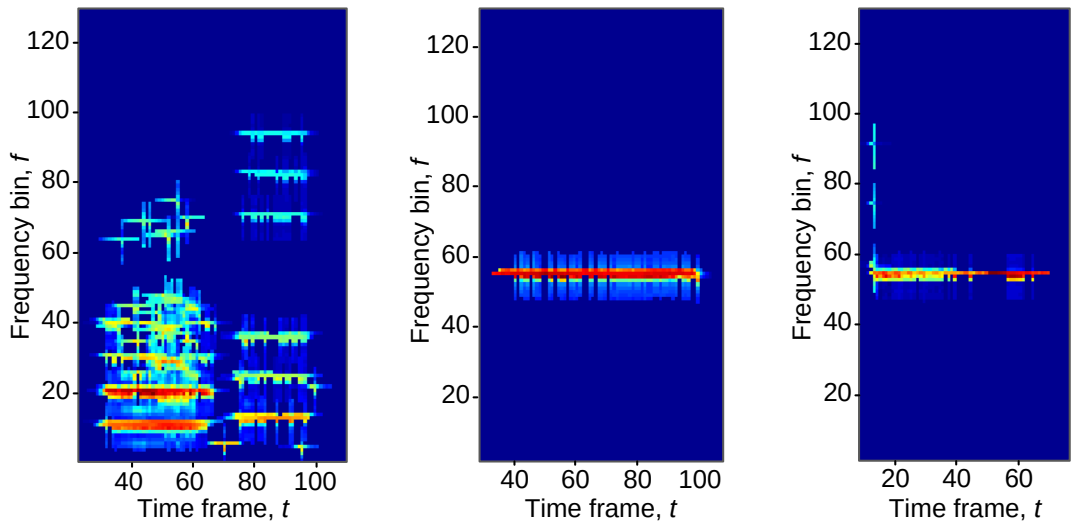
Volume Change

The final experiment examined the performance under changing volume, and the results can be found in the last section of Table 5.1. The results show that the LSF system performs consistently well, maintaining a TP above 98% for an FA below 0.7%. This compares well to the baseline methods, where the TP drops and the FA increases as the difference between the training and testing volumes increases.

Comparing the baselines, Overlap-SVM again performs significantly better, achieving a TP of 84% at 0.5× volume change compared to just 56% for MixMax-GMM. This may be due to the fact that the feature set for the Overlap-SVM includes perceptual features, such as zero-crossing rate and spectral centroid, which are less affected by the simple change in volume. However, both method perform significantly worse than the LSF method, which is able to estimate the unknown transfer function using the GHT voting mechanism detailed in Section 5.2.5. In addition, the LSF transfer function is able to vary across frequency, meaning it should be able to adapt to a wider range of real-world conditions.

Sound Event Reconstruction

It should be noted that the LSF approach can reconstruct the observed sound event based on the keypoints and clusters that contributed to the hypothesis score. This



(a) Horn. (b) Whistle. (c) Bell Ringing.

Figure 5.11: Example LSF reconstructions of the three overlapping sounds from Fig. 5.1, using the assigned codebook clusters to reconstruct the spectrograms of the sound events. This also demonstrates that the LSF approach is not limited to a just two sounds, since all three overlapping sounds can be recognised without modification to the algorithm.

is possible because each of the detected keypoints is matched against the codebook, and the codebook centre can be used to reconstruct the local spectral information. An example of this was shown previously in Fig. 5.9c for a bell and horn overlapping, and another example is shown below in Fig. 5.11 for each of the three overlapping sounds from Fig. 5.1. These figures demonstrate that the observed regions of the sound events can be well reconstructed, such that the output represents the spectral information from the sound event in clean conditions. Additionally, it may be possible to utilise information regarding the missing keypoints to reconstruct the entire overlapped sound, however the focus in this chapter is on recognition, hence reconstruction is left for a future work.

Three Overlapping Sounds

A final example is to demonstrate the ability of the proposed LSF system to perform recognition of more than two overlapping sounds, without any modifications to the

algorithm. This is possible because no assumptions are required about the number of sound events that may occur simultaneously, since the GHT generates a set of detection hypotheses based on the observed evidence from the extracted keypoints and LSFs.

An example is shown in Fig. 5.11, where the bell, horn and whistle are recognised and reconstructed from the noisy spectrogram example shown previously in Fig. 5.1. Here it can be seen that even though the harmonic of the bell is similar in spectral shape to the whistle, the segmentation is largely correct since the temporal distribution of the whistle sound is different from the bell.

For the baseline methods, classification of this sample would not be possible, as the Overlap-SVM requires training on all expected combinations in advance, and the MixMax-GMM is currently only derived for two overlapping combinations. While it may be possible to modify these methods to allow for the additional overlaps, this involves increased training for Overlap-SVM, and increased computational effort for MixMax-GMM. In addition, the additional overlapped classes would inevitably reduce the performance of both methods compared to the results presented in this section. However, without modification the LSF system can still recognise each of the three sounds, with very few LSFs incorrectly attributed to the wrong sound. This highlights a significantly benefit of the LSF approach over many of the state-of-the-art baseline methods.

5.4 Summary

This chapter addressed the challenging task of simultaneous recognition of sound events in overlapping and noisy conditions. This is motivated by both the challenging nature of the problem, and the limitations of the state-of-the-art, which typically use conventional frame-based systems or rely on the automatic decomposition of the spectrogram. However, the problem of recognising overlapping sound events in noise has many parallels with object detection in image processing, where overlapping objects may obscure each other and be set against an unknown background. Combined with inspiration from human perception of sound, based on partial recognition of spectral information across frequency, this provides the basis for the LSF approach introduced in this chapter. The idea is to detect keypoints in the spectrogram, and then characterise the sound jointly through both the LSF and the geometrical distribution of

the matching codebook cluster in the spectrogram. For recognition, this distribution model can be used as a GHT voting function, where the independent information, accumulated across the set of keypoints in the spectrogram, provides evidence for sound events in a space that is sparse and separable in challenging overlapping conditions. Experiments were carried out on a simulated database of overlapping sounds, with two baseline frame-based techniques implemented for comparison. The results demonstrated that the LSF system performed significantly better than the baseline methods in overlapping and mismatched conditions. In addition, the approach is able to detect an arbitrary combination of overlapping sounds, including two or more different sounds or the same sound overlapping itself, which is an important improvement over the baseline techniques.

Chapter 6

Conclusions and Future Work

This thesis has focussed on the topic of sound event recognition (SER), where the aim is to detect and classify the rich array of acoustic information that is present in many environments. This is a challenging task in unstructured environments where there are many unknowns, e.g. the number and location of the sources, and any noise or channel effects that may be present. However, many state-of-the-art SER systems perform poorly in such situations, particularly those that rely on frame-based features, where each feature can contain a mixture of multiple sources or noise. This thesis has developed novel approaches to address some of the challenges faced, in particular by using inspiration from image processing as a foundation. The motivation stems from the fact that spectrograms form recognisable images, with a characteristic spectral geometry, which can be identified by a human reader. Hence, the idea is to base the approaches on spectrogram image processing, with the time-frequency spectrogram of the sound interpreted as a special case of a conventional image. To conclude the work, Section 6.1 first summarises the contributions proposed in this thesis. Finally, Section 6.2 discusses some of the future directions that can be explored, and the challenges that are still to be faced.

6.1 Contributions

The approach taken in this thesis has been to interpret the sound event spectrogram as an image, where each pixel in the image represents the spectral power at a particular

time and frequency. By taking this two-dimensional approach, the extracted features naturally capture both spectral and temporal information together. This is different from conventional frame-based audio features, which typically extract a feature that represents only the spectral information contained within each short-time frame. By combining image processing-inspired feature extraction from the spectrogram, with techniques for noise robust recognition, the resulting methods can significantly improve upon the state-of-the-art methods across a range of challenging experimental conditions. This idea of using spectrogram image processing has formed the basis for the contributions presented in this thesis, which are summarised below.

6.1.1 Spectrogram Image Feature

The idea of the SIF is to extract a visual signature from the sound's time-frequency spectrogram for classification, with the work published in IEEE Signal Processing Letters [5]. The method first maps the grey-scale spectral values into a higher dimensional space, by quantising the dynamic range into separate regions in a process analogous to pseudo-colourmapping in image processing. This is inspired by the visual perception of the spectrogram, where pseudo-colourmapping enhances the discrimination between the high-power spectral peaks and the low-power stochastic noise in the background. An image feature is then extracted from the quantised spectrogram that characterises the local pixel distribution in the image, based on the colour layout feature from image processing. This naturally captures the joint spectro-temporal information contained in the sound event signal, hence can improve upon the frame-based spectral features found in conventional state-of-the-art audio processing systems.

An evaluation was carried out using 50 sound event classes from the RWCP sound scene database, with the SIF compared to the best-performing baseline methods drawn from a wide range of different techniques. The results showed that the SIF achieved an average classification accuracy of 88.5% across the four noise conditions, which is an average improvement of 14.6% over the comparable ETSI-AFE baseline. The ETSI method uses the same clean samples for training, and relies on noise reduction in mismatched conditions to enhance the features prior to classification. Another baseline method, based on MFCC-HMM with the multi-conditional training approach, achieved a comparable performance to the SIF with an average accuracy of 88%. However, while

multi-conditional training is a popular approach for mismatched conditions, it requires a large amount of data for training, and may only achieve a good performance under certain noise conditions.

Detailed experiments were also carried out to compare the different aspects that contributed to the success of the SIF. These included varying the type of spectrogram image used, and performing classification with and without the proposed dynamic range quantisation. The results demonstrated that the quantisation step was key to achieving a good result, with image feature extraction performed directly on the grey-scale spectral values achieving an average classification accuracy of only 66.0%. Additionally, it was found that using the linear power spectrogram image was significantly better than both the log power or cepstral representations. This is due to the sparsity of the linear power representation, where the only the most reliable and characteristic elements fall into the highest dynamic range quantisation, therefore producing a more robust feature for classification.

6.1.2 Subband Power Distribution Image

Extending the previous work on the SIF, a novel sound event image representation called the SPD was proposed, with the work published in the IEEE Transactions on Audio, Speech and Language Processing journal [6], and in the Interspeech 2011 conference [7]. The SPD is a two-dimensional representation of the distribution of normalised spectral power over time against frequency, which gives the SPD a fixed dimension that is independent of the length of the clip. The temporal information is captured implicitly through the distribution information, such that a characteristic pattern will be visible in the spectrogram for different sound events. The advantage of the SPD over the spectrogram is that the sparse, high-power elements of the sound event are transformed to a localised region of the SPD, unlike in the spectrogram where they may be scattered over time and frequency. This enables a missing feature mask to be easily applied to the SPD, and the missing elements of the extracted image feature can simply be marginalised in a missing feature classification system. Here, it was proposed to generate the mask directly from the SPD representation, and then to use the kNN method for classification, utilising the Hellinger distance measure to naturally compare the distribution distance between image features.

Experimental validation was carried out to analyse the performance of the proposed SPD approach against both the SIF and the best performing baseline techniques, using the same experimental database for comparison. The results showed that the SPD-IF, requiring only clean samples for training, was both highly discriminative in clean conditions and robust to noise in mismatched conditions. This was demonstrated through an average classification accuracy of almost 96% over the four noise conditions, and achieving over 90% in the challenging 0dB noise condition. This is a significant average improvement of 7.4% over even the best-performing multi-conditional MFCC-HMM benchmark, with the improvement increasing to over 23% in the 0dB noise condition. The result reinforces the idea of using two-dimensional feature extraction techniques for sound event recognition, since the performance exceeds even the state-of-the-art multi-conditional training method in both clean and challenging mismatched noise conditions.

6.1.3 Local Spectrogram Features

The final work on the LSF focussed on the challenging task of simultaneous recognition of overlapping sounds, with the work published in Pattern Recognition Letters [8], and in the Interspeech 2012 and ICASSP 2013 conferences [9, 10]. In the unstructured environments that are commonly found in SER applications, it is much more likely for multiple signals to be received simultaneously at a distant microphone, unlike with the close-talking microphones commonly used for ASR. For a human listener, the task of separating and recognising these overlapping sounds is intuitive and simple, and is commonly referred to as the "cocktail party effect" for speech mixed with other competing speech and sounds. However, conventional frame-based methods aren't well suited to the problem, since each time frame contains a mixture of information from multiple sources. The difficulty faced by these methods is that the training only captures information about the sound in isolation, hence when two sound events overlap, the spectral mixture in each frame will produce low scores against both of the trained models.

Looking at the image processing domain, the problem of object detection in cluttered environments can be seen to have many similarities with detecting overlapping sounds embedded in background noise. This is because overlapping objects in an image

may obscure one other, and be set against an unknown background. This is comparable to a spectrogram containing overlapping sounds, which will mask each other through the MixMax criteria, with unknown stochastic noise in the background. Therefore, the LSF approach is developed, which is based on a distribution model of the local spectro-temporal information extracted from the spectrogram, with detection based on the generalised Hough transform (GHT). This uses the distribution model as a voting function, which sums together the information over a set of independent keypoints to produce sparse and separable onset hypotheses. The result is that the approach can detect any arbitrary combination of sound events in the spectrogram, including two or more different sounds or the same sound overlapping itself. This is an important improvement over the baseline techniques, which typically make assumptions about the number of sounds present or the amount of overlap. The final step of the algorithm is to score the detection hypotheses against the trained model, which includes the ability to estimate the relative transfer function between training and testing, such that the method can take account of unknown channel distortion.

Experiments were carried out on a simulated database of overlapping sounds, with two different baseline techniques implemented for comparison. The results showed that the LSF system could achieve state-of-the-art results across a range of challenging conditions. In clean conditions, the method achieved an average accuracy of 99.3% and 98.0% in isolated and overlapping conditions respectively. This compared well with the best-performing Overlap-SVM baseline, which achieved 100% and 96.5% respectively in the matched conditions. However, the most significant result was demonstrated in mismatched conditions, where the LSF system achieved an accuracy of 90.2%, for a false alarm of only 2.5%, in the overlapping experiment with 0dB noise. This is an improvement in accuracy of almost 25% compared to the best baseline method, and also a reduction in false alarm of over 20%. Finally, a simple experiment was carried out to simulate a varying channel distortion by changing the volume of the received sound signal. Again, the LSF system demonstrated the best performance, with almost no change in accuracy observed for the simultaneous recognition across the different volume conditions. This is compared to the Overlap-SVM baseline, where the accuracy under 0.5× volume was over 10% less than the results achieved in matched conditions.

6.2 Future Directions

The goal of this thesis has been to develop novel algorithms for sound event recognition across a range of challenging experimental conditions. This has resulted in the development of several techniques that address aspects of this problem, such as noise robustness and simultaneous recognition of overlapping sounds. Particularly, the performance of the SPD-IF method is shown to significantly exceed the state-of-the-art for classification, hence lends itself to real-time implementation for practical application. This itself introduces challenges, such as detection of sound events and rejection of false alarms, and in addition there are possible improvements in the modelling of the subband distribution and in estimating the noise mask in the SPD. However, the progression in this thesis has been to shift from extracting global features from the spectrogram, such as the SIF and SPD-IF, to extracting local features that can be combined to perform recognition. Therefore, the following areas for future work focus on areas of enhancement for the LSF approach, which will enable this promising method to applied in applications with a larger range of sound classes and real-world conditions. These are now discussed below.

6.2.1 Modelling

The distribution of the LSF codebook clusters is modelled over time, frequency and spectral power to characterise the sound event spectrogram. In essence, this forms a template for recognition of the cluster information during testing. However, certain sounds have more variation in spectral content, or consist of several spectral patterns separated by a short silence. An example of this would be a stapler, where sound is generated by both pressing and releasing the mechanism. The problem is that this leads to a less sharp sound event model, as different timings causing a blurring of the geometrical information. To overcome this, a method could be found for grouping regions of similar or repeating spectral content. Each region can then be modelled separately and linked together to create a sharper overall sound event model. This requires an initial grouping mechanism that could be developed based on similar ideas from CASA.

Another factor that could improve the modelling is to relax the assumption that the

sound event has a fixed frequency content. Currently, the sound event model is only allowed to shift along the time and spectral power axes during recognition. However, sounds from the same class may have a similar spectral content, but can be shifted or scaled in frequency. An example of this would be a bell sound, where a similar bell should have a similar geometry, but with the harmonic at a different frequency. This would not be recognised under the current scheme, unless an example of the sound event had been present in training. To improve the generalisation, clusters from the geometrical model could therefore be also allowed to shift along the frequency axis. Constraints would have to be set to prevent clusters of the same sound event from overlapping each other, and to penalise too greater shift away from the trained model. However, improving the generalisation ability of the sound event model should give an overall benefit to the approach.

6.2.2 Scoring

At present, the scoring mechanism for the LSF system treats each hypothesis in the spectrogram separately. Starting from the strongest hypothesis, it scores each in turn against the trained model until all hypotheses have been tested. When the score exceeds a threshold, the hypothesis is accepted and the keypoints that contributed to the recognised sound event are removed from the remaining hypotheses. The advantage of this is that the trade-off between false acceptance and rejection is controlled with a simple threshold, with the stronger hypotheses more likely to exceed this threshold.

An alternative approach is to consider the interaction between all of the hypotheses simultaneously. In this way, the cost of assigning a region of the spectrogram to one sound event or another can be better controlled. This may reduce the recognition errors observed when two sound event hypotheses, of which only one is correct, completely overlap each other. This is because assigning the same region to two different sound events can be penalised in the revised scoring mechanism. An approach for this has been previously developed for object detection, where the hypotheses are evaluated against a minimum description length (MDL) criteria in the spectrogram [206]. To bring this idea to sound event recognition, the method will need to be adapted to allow for regions of overlap that are caused by masking. However, the advantage is that it should lead to fewer false alarms, which is an important factor for real-life

implementation of the recognition system.

6.2.3 Segmentation and Reconstruction

It was discussed in Section 5.3.2 that the LSF approach may allow for the possibility of reconstructing spectral information from the recognised sound event. This is enabled through the matching of the extracted LSFs with the codebook clusters, allowing each LSF to be replaced by the corresponding cluster centre. This represents information learned during training on the clean sound event samples. Therefore, it should be possible to reconstruct a clean sample of a sound event, even if it is recognised in noisy conditions. However, further work is required to study the best method for reconstructing the spectral information from the LSFs, which may overlap each another in the spectrogram. In addition, since the reliable regions of the sound event should remain intact, a method needs to be found to ensure that the observed and reconstructed sound events are similar. One possible advantage of the LSF approach is that the model contains information about missing regions that may have been masked by other sounds. In this way, reconstruction of missing areas of the spectrogram is made possible by finding the most likely cluster that could have occurred in a particular region.

The problem of reconstructing is closely linked to the issue of segmentation of the recognised sound event from the spectrogram. Using the LSF approach, a local missing feature mask is extracted for each local region. This can be combined with the distance between the observed LSF and the cluster centre, to generate a segmentation based on the model of the sound event in clean conditions. The output would be similar to the process of mask estimation for missing feature recognition. However, in this case the segmentation is being extracted in a top-down manner, since the sound event is first recognised using the LSF system, before the segmentation is determined. This is opposed to the bottom-up mask estimation that is conventionally used, which typically uses low-level information extracted from the spectrogram.

6.2.4 Other tasks

In this thesis, the focus has been on the task of robust recognition of sound events in noisy and overlapping conditions. The LSF approach additionally takes account of

channel distortion for cases where the response time of the channel is short. However, this may not be suitable for distortion caused by reverberation due to a room impulse response, where the response time may be significantly longer than the analysis window length. In such cases, the effect on the spectrogram operates across both time and frequency, typically causing a blurring of the temporal information. It may be possible to reduce this problem by utilising techniques from image processing such as compensation for motion blur effects. An alternative is possibly to use a reverberation mask to enable matching between LSFs in clean and mismatched conditions.

Another future topic of study will also be the application of the approaches presented in this thesis to speech recognition. While they may not be initially suited to recognising connected speech, it may be possible to apply them for phoneme detection. Provided that phonemes can be detected with sufficient accuracy, the output can be fused with the output of a conventional ASR system. An alternative application could be in speaker verification, where a text dependent template of the given speaker could be modelled as a spectrogram image for classification.

References

[1] R. F. Lyon, "Machine Hearing: An Emerging Field," *IEEE Signal Processing Magazine*, vol. 27, pp. 131–139, Sept. 2010.

[2] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental Sound Recognition With Time-Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1142–1158, Aug. 2009.

[3] B. Ghoraani and S. Krishnan, "Time-Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[4] V. Zue, "Notes on spectrogram reading," *Mass. Inst. Tech. Course*, vol. 6, 1985.

[5] J. Dennis, H. D. Tran, and H. Li, "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *IEEE Signal Processing Letters*, vol. 18, pp. 130–133, Feb. 2011.

[6] J. Dennis, H. D. Tran, and E. S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 367–377, Feb. 2013.

[7] J. Dennis, H. D. Tran, and H. Li, "Image Representation of the Subband Power Distribution for Robust Sound Classification," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2437–2440, Aug. 2011.

[8] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised Hough transform," *Pattern Recognition Letters*, vol. 34, pp. 1085–1093, July 2013.

[9] J. Dennis, H. D. Tran, and E. S. Chng, "Overlapping Sound Event Recognition using Local Spectrogram Features with the Generalised Hough Transform," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Sept. 2012.

[10] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, "Temporal Coding of Local Spectrogram Features for Robust Sound Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[11] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang, "Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition," in *Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information*, pp. 293–298, IEEE, 2007.

[12] 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*, p. 77, World Scientific Pub Co Inc, 2004.

[13] 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.

[14] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[15] J. Godfrey, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[16] C. H. Chen, "Pattern recognition applications in underwater acoustics," *The Journal of the Acoustical Society of America*, vol. 75, p. S75, May 1984.

[17] C. H. Chen, "Recognition of underwater transient patterns," *Pattern Recognition*, vol. 18, pp. 485–490, Jan. 1985.

[18] A. Mohamed and H. Raafat, "Recognition of heart sounds and murmurs for cardiac diagnosis," in *Proceedings of the International Conference on Pattern Recognition*, pp. 1009–1011, IEEE Comput. Soc. Press, 1988.

[19] B. Pinkowski, "A template-based approach for recognition of intermittent sounds," in *Computing in the 90's* (N. A. Sherwani, E. Doncker, and J. A. Kapenga, eds.), vol. 507 of *Lecture Notes in Computer Science*, pp. 51–57, Berlin/Heidelberg: Springer-Verlag, 1991.

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

[21] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.

[22] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robots: Sound localization and separation," *Artificial Life and Robotics*, vol. 1, pp. 157–163, Dec. 1997.

[23] 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 and Language*, vol. 24, no. 1, pp. 30–44, 2010.

[24] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[25] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp. 737–747, June 2001.

[26] H. G. Kim, J. J. Burred, and T. Sikora, "How efficient is MPEG-7 for general sound recognition?," in *AES 25th International Conference on Metadata for Audio*, 2004.

[27] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems," *IV Jornadas en Tecnologia del Habla*, 2006.

[28] 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," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2342–2345, Sept. 2010.

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

[30] D. O'Shaughnessy, *Speech Communications: Human and Machine*. Wiley-IEEE Press, 2000.

[31] D. Gerhard, "Audio Signal Classification: History and Current Techniques," in *Technical Report TR-CS 2003-07*, pp. 1–38, Department of Computer Science, University of Regina, 2003.

[32] M. Cowling, *Non-Speech Environmental Sound Classification System for Autonomous Surveillance*. PhD thesis, Griffith University, Gold Coast Campus, 2004.

[33] S. J. Barry, A. D. Dane, A. H. Morice, and A. D. Walmsley, "The automatic recognition and counting of cough.," in *Cough (London, England)*, vol. 2, p. 8, Jan. 2006.

[34] Y. T. Peng, C. Y. Lin, M. T. Sun, and K. C. Tsai, "Healthcare audio event classification using Hidden Markov Models and Hierarchical Hidden Markov Models," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1218–1221, IEEE, June 2009.

[35] A. Temko and C. Nadeu, "Classification of Meeting-Room Acoustic Events with Support Vector Machines and Variable-Feature-Set Clustering," in *Proceedings of*

the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, pp. 505–508, IEEE, 2005.

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

[37] A. Temko, D. Macho, and C. Nadeu, "Fuzzy integral based information fusion for classification of highly confusable non-speech sounds," *Pattern Recognition*, vol. 41, pp. 1814–1823, May 2008.

[38] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, pp. 1281–1288, Oct. 2009.

[39] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, and J. R. Casas, "Acoustic Event Detection Based on Feature-Level Fusion of Audio and Video Modalities," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 11, 2011.

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

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

[42] P. Khunarsa, C. Lursinsap, and T. Raicharoen, "Impulsive Environment Sound Detection by Neural Classification of Spectrogram and Mel-Frequency Coefficient Images," in *Advances in Neural Network Research and Applications*, pp. 337–346, 2010.

[43] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 1267–1271, 2010.

[44] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, pp. 1543–1551, Sept. 2010.

[45] E. Tsau, S. Chachada, and C.-C. J. Kuo, "Content / Context-Adaptive Feature Selection for Environmental Sound Recognition," in *Proceedings of the Asia-Pacific Signal & Information Processing Association (APSIPA)*, 2012.

[46] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 333–336, IEEE, Mar. 2012.

[47] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, p. 4 pp., 2005.

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

[49] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 165–168, IEEE, Apr. 2009.

[50] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2007.

[51] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric Representations of Bird Sounds for Automatic Species Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 2252–2263, Nov. 2006.

[52] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. H. Tauchert, and K.-H. H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, pp. 1524–1534, Sept. 2010.

[53] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 337–340, IEEE, May 2011.

[54] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, pp. 1–22, July 2006.

[55] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 885–888, IEEE, July 2006.

[56] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 321–329, Jan. 2006.

[57] S. Ravindran and D. Anderson, "Audio Classification And Scene Recognition and for Hearing Aids," *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 860–863, 2005.

[58] C. Freeman, R. Dony, and S. Areibi, "Audio Environment Classication for Hearing Aids using Artificial Neural Networks with Windowed Input," in *Proceedings of the IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP)*, no. Ciisp, pp. 183–188, IEEE, Apr. 2007.

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

[60] F. Beritelli and R. Grasso, "A pattern recognition system for environmental sound classification based on MFCCs and neural networks," in *Proceedings of the International Conference on Signal Processing and Communication Systems*, pp. 1–4, IEEE, Dec. 2008.

[61] T. Kinnunen, R. Saeidi, J. Lepp, and J. P. Saarinen, "Audio Context Recognition in Variable Mobile Environments from Short Segments Using Speaker and Language Recognizers," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, no. 132129, 2012.

[62] G. Muhammad and K. Alghathbar, "Environment Recognition from Audio Using MPEG-7 Features," in *Proceedings of the International Conference on Embedded and Multimedia Computing*, pp. 1–6, IEEE, Dec. 2009.

[63] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, pp. 2965–2979, Oct. 2008.

[64] Y. Muthusamy, E. Barnard, and R. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 33–41, Oct. 1994.

[65] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4889–4892, 2012.

[66] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 4072–4075, IEEE, May 2002.

[67] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 953–956, IEEE, 2005.

[68] E. Mumolo, M. Nolich, and G. Vercelli, "Algorithms for acoustic localization based on microphone array in service robotics," *Robotics and Autonomous Systems*, vol. 42, pp. 69–88, Feb. 2003.

[69] J. Liu, H. Erwin, and S. Wermter, "Mobile robot broadband sound localisation using a biologically inspired spiking neural network," *Proceedings of*

the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2191–2196, Sept. 2008.

[70] S. Marchand and A. Vialard, "The Hough transform for binaural source localization," in *Proceedings of the Digital Audio Effects Conference (DAFx)*, pp. 252–259, 2009.

[71] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE transactions on Speech and Audio Processing*, vol. 10, pp. 504–516, Oct. 2002.

[72] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, "Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 688–707, Mar. 2010.

[73] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds.," *Perception & psychophysics*, vol. 69, pp. 839–55, Aug. 2007.

[74] R. K. Reddy, V. Ramachandra, N. Kumar, and N. C. Singh, "Categorization of environmental sounds," *Biological cybernetics*, vol. 100, pp. 299–306, Apr. 2009.

[75] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech communication*, vol. 49, no. 6, pp. 501–513, 2007.

[76] K. Kumatani, J. McDonough, and B. Raj, "Microphone Array Processing for Distant Speech Recognition: From Close-Talking Microphones to Far-Field Sensors," vol. 29, 2012.

[77] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 114 – 126, 2012.

[78] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust Speech Recognition and Understanding*, pp. 1–22, 2007.

[79] A. Temko, *Acoustic Event Detection and Classification*. PhD thesis, Universitat Politecnica de Catalunya, Barcelona, 2007.

[80] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[81] D. O'Shaughnessy, "Formant Estimation and Tracking," in *Springer Handbook of Speech Processing* (J. Benesty, M. Sondhi, and Y. Huang, eds.), pp. 213–228, Springer Berlin Heidelberg, 2008.

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

[83] F. Pachet and P. Roy, "Exploring Billions of Audio Features," in *Proceedings of the 2007 International Workshop on Content-Based Multimedia Indexing*, pp. 227–235, IEEE, June 2007.

[84] 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.

[85] H. D. Tran and H. Li, "Sound Event Recognition With Probabilistic Distance SVMs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1556–1568, Aug. 2011.

[86] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Transactions on Multimedia*, vol. 14, pp. 1684–1689, Dec. 2012.

[87] G. Tzanetakis and P. Cook, "Multifeature Audio Segmentation For Browsing And Annotation," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 103–106, 1999.

[88] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: Sound Object Localization and Retrieval in Complex Audio Environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 429–432, IEEE, 2005.

[89] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," in *CUIDADO I.S.T. Project Report*, pp. 1–25, 2004.

[90] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

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

[92] T. Butko, *Feature Selection for Multimodal Acoustic Event Detection*. PhD thesis, Universitat Politecnica de Catalunya, 2011.

[93] M. Cowling and R. Sitte, "Analysis of speech recognition techniques for use in a non-speech sound recognition system," *Proceedings of the International Symposium on Digital Signal Processing for Communication Systems*, pp. 16–20, 2002.

[94] K. P. Murphy, *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[95] 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.

[96] G. Fung, O. L. Mangasarian, and E. W. Wild, "Proximal support vector machine classifiers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, (New York, New York, USA), pp. 77–86, ACM Press, 2001.

[97] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, pp. 275–294, July 1991.

[98] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[99] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[100] 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, p. 212, 2003.

[101] X. Xiao, E. S. Chng, and H. Li, "Normalization of the Speech Modulation Spectra for Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1662–1674, Nov. 2008.

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

[103] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, June 2001.

[104] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, pp. 101–116, Sept. 2005.

[105] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[106] G. Hu and D. Wang, "Auditory Segmentation Based on Onset and Offset Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 396–405, Feb. 2007.

[107] S. Srinivasan and D. Wang, "A model for multitalker speech perception," *The Journal of the Acoustical Society of America*, vol. 124, pp. 3213–3224, 2008.

[108] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, vol. 24, pp. 77–93, Jan. 2010.

[109] B. Raj, M. L. Seltzer, and R. M. Stern, "Robust speech recognition: the case for restoring missing features," in *Proceedings of Eurospeech*, 2001.

[110] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[111] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, pp. 2895–2907, Nov. 2003.

[112] V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1941—-1944, 2002.

[113] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1151–1154, 2009.

[114] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, p. 10, Nov. 2009.

[115] J. G. Wilpon, C.-H. Lee, and L. R. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 349–352 vol.1, IEEE, 1991.

[116] 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.

[117] H. Hermansky and S. Sharma, "Traps-classifiers of temporal patterns," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1003–1006, Nov. 1998.

[118] G. Peeters and E. Deruty, "Sound indexing using morphological description," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 675–687, Mar. 2010.

[119] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal Feature Integration for Music Genre Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1654–1664, July 2007.

[120] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification With Application to Musical Instrument Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 174–186, Jan. 2009.

[121] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acta Acustica united with Acustica*, vol. 88, no. 3, pp. 416–422, 2002.

[122] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," in *Proceedings of Eurospeech*, pp. 1–4, 2003.

[123] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4733–4736, IEEE, Mar. 2008.

[124] G. Kovacs and L. Toth, "Localized spectro-temporal features for noise-robust speech recognition," in *Proceedings of the 2010 International Joint Conference on Computational Cybernetics and Technical Informatics*, pp. 481–485, IEEE, 2010.

[125] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, pp. 736–752, 2011.

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

[127] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition using MP-based features," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1–4, IEEE, Mar. 2008.

[128] V. T. Peltonen, *Computational Auditory Scene Recognition*. PhD thesis, 2001.

[129] 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.

[130] S. Strahl and A. Mertins, "Analysis and design of gammatone signal models.," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2379–89, Nov. 2009.

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

[132] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, Apr. 2002.

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

[134] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, no. 1, pp. 4625–4628, IEEE, Apr. 2009.

[135] Z. Tuske, P. Golik, R. Schluter, and F. R. Drepper, "Non-stationary feature extraction for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5204–5207, IEEE, May 2011.

[136] Y. R. Leng, H. D. Tran, N. Kitaoka, and H. Li, "Selective Gammatone Filterbank Feature for Robust Sound Event Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2246–2249, 2010.

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

[138] J. O. Pickles, *An introduction to the physiology of hearing*. Academic Press, 2008.

[139] T. C. Walters, *Auditory-Based Processing of Communication Sounds*. PhD thesis, University of Cambridge, 2011.

[140] R. Duda and P. Hart, "Experiments in scene analysis," *Proceedings of the First National Symposium on Industrial Robots*, Apr. 1970.

[141] M. Gainza, B. Lawlor, and E. Coyle, "Onset based audio segmentation for the irish tin whistle," in *Proceedings of the International Conference on Signal Processing (ICSP)*, vol. 1, pp. 594–597, IEEE, 2004.

[142] J. Barker, M. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, Jan. 2005.

[143] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Computer Speech and Language*, vol. 24, pp. 94–111, Jan. 2010.

[144] N. Ma, J. Barker, H. Christensen, and P. Green, "Combining Speech Fragment Decoding and Adaptive Noise Floor Modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 818–827, Mar. 2012.

[145] R. C. DeCharms and M. M. Merzenich, "Primary cortical representation of sounds by the coordination of action-potential timing," *Nature*, vol. 381, pp. 610–3, June 1996.

[146] C. M. Wessinger, M. H. Buonocore, C. L. Kussmaul, and G. R. Mangun, "Tonotopy in human auditory cortex examined with functional magnetic resonance imaging," *Human Brain Mapping*, vol. 5, no. 1, pp. 18–25, 1997.

[147] 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, pp. 1008–1021, Sept. 2004.

[148] C. Köppl and G. Yates, "Coding of sound pressure level in the barn owl's auditory nerve.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 19, pp. 9674–86, Nov. 1999.

[149] W. Smit and E. Botha, "Spiking neural networks for sound recognition," in *12th Annual Symposium of the Pattern Recognition Association of South Africa*, 2001.

[150] 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, p. 1282, Jan. 2001.

[151] R. Gütig and H. Sompolinsky, "Time-warp-invariant neuronal processing," *PLoS biology*, vol. 7, p. e1000141, July 2009.

[152] R. Pichevar, H. Najaf-Zadeh, L. Thibault, and H. Lahdili, "Auditory-inspired sparse representation of audio signals," *Speech Communication*, vol. 53, pp. 643–657, May 2011.

[153] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design.," *Journal of computational neuroscience*, vol. 9, no. 1, pp. 85–111, 2000.

[154] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex.," *Journal of neurophysiology*, vol. 85, pp. 1220–34, Mar. 2001.

[155] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds.," *The Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, Aug. 2005.

[156] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 920–930, May 2006.

[157] R. Gütig and H. Sompolinsky, "The tempotron: a neuron that learns spike timing-based decisions," *Nature Neuroscience*, vol. 9, pp. 420–428, Feb. 2006.

[158] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer Speech and Language*, vol. 21, pp. 443–457, July 2007.

[159] 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 the International Conference on Language Resources and Evaluation*, vol. 2, pp. 965–968, 2000.

[160] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[161] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book, version 3.4*, vol. 3. 2006.

[162] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, 1993.

[163] C. Heil and D. Walnut, "Continuous and discrete wavelet transforms," *SIAM review*, vol. 31, no. 4, pp. 628–666, 1989.

[164] R. C. Gonzalez and R. Woods, *Digital Image Processing*. Prentice Hall Upper Saddle River, NJ, 2002.

[165] K. Mikolajczyk, "Scale & Affine Invariant Interest Point Detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, Oct. 2004.

[166] Y. Amit, A. Koloydenko, and P. Niyogi, "Robust acoustic object detection," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, p. 2634, 2005.

[167] A. Defaux, *Detection and Recognition of Impulsive Sound Signals*. PhD thesis, University of Neuchatel, Switzerland, 2001.

[168] R. E. Turner, *Statistical models for natural sounds*. PhD thesis, University College London, 2010.

[169] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts," *International Journal of Computer Vision*, vol. 56, pp. 151–177, Feb. 2004.

[170] P. M. Roth and M. Winter, "Survey of appearance-based methods for object recognition," in *Inst. for Computer Graphics and Vision, Graz University of Technology, Austria, Technical Report ICGTR0108 (ICG-TR-01/08)*, 2008.

[171] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[172] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[173] K. T. Schutte, *Parts-based Models and Local Features for Automatic Speech Recognition*. PhD thesis, Massachusetts Institute of Technology, 2009.

[174] H. Deshpande, R. Singh, and U. Nam, "Classification of music signals in the visual domain," in *Proceedings of the COST-G6 Conference on Digital Audio Effects*, pp. 1–4, 2001.

[175] I. Paraskevas and E. Chilton, "Audio classification using acoustic images for retrieval from multimedia databases," in *Proceedings of the EURASIP Conference focused on Video/Image Processing and Multimedia Communications*, vol. 1, pp. 187–192, IEEE, Faculty of Electrical Eng. & Comput, 2003.

[176] S. Barnwal, K. Sahni, R. Singh, and B. Raj, "Spectrographic seam patterns for discriminative word spotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4725–4728, IEEE, Mar. 2012.

[177] R. Duda and P. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[178] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer Vision for Music Identification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 597–604, IEEE, 2005.

[179] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, May 2004.

[180] S. Baluja and M. Covell, "Audio Fingerprinting: Combining Computer Vision & Data Stream Processing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 213–216, IEEE, 2007.

[181] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," in *Proceedings of the International Society on Music Information Retrieval (ISMIR)*, 2011.

[182] T. Muroi, T. Takiguchi, and Y. Ariki, "Gradient-based acoustic features for speech recognition," in *Proceedings of the 2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, no. Ispacs, pp. 445–448, Ieee, Dec. 2009.

[183] Y.-c. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognition Letters*, vol. 26, pp. 1327–1336, July 2005.

[184] T. Matsui, M. Goto, J. P. Vert, and Y. Uchiyama, "Gradient-based musical feature extraction based on scale-invariant feature transform," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 724–728, 2011.

[185] K. Behún, "Image features in music style recognition," in *Proceedings of the Central European Seminar on Computer Graphics (CESCG)*, 2012.

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

[187] K. T. Schutte and J. R. Glass, "Speech recognition with localized time-frequency pattern detectors," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pp. 341–346, IEEE, 2007.

[188] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proceedings of the Workshop on Statistical And Perceptual Audition (SAPA)*, 2008.

[189] A. L.-C. Wang, "An industrial strength audio search algorithm," in *Proceedings of the International Society on Music Information Retrieval (ISMIR)*, vol. 2, 2003.

[190] J. P. Ogle and D. P. W. Ellis, "Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 233–236, IEEE, 2007.

[191] A. Ashbrook and N. Thacker, "Tutorial: Algorithms For 2-Dimensional Object Recognition," in *Tina Memo No. 1996-003*, no. 1996, pp. 1–15, 1998.

[192] S. Uchida and H. Sakoe, "A Survey of Elastic Matching Techniques for Handwritten Character Recognition," *IEICE Transactions on Information and Systems*, vol. E88-D, pp. 1781–1790, Aug. 2005.

[193] J. Mundy, "Object recognition in the geometric era: A retrospective," in *Toward category-level object recognition*, pp. 3–28, 2006.

[194] J. Shih and L. Chen, "Colour image retrieval based on primitives of colour moments," in *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 149, p. 370, IET, 2002.

[195] F. A. Cheikh, "MUVIS: A System for Content-Based Image Retrieval," tech. rep., 2004.

[196] Y.-C. Cheng and S.-Y. Chen, "Image classification using color, texture and regions," *Image and Vision Computing*, vol. 21, pp. 759–776, Sept. 2003.

[197] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[198] L. Cieplinski, "MPEG-7 color descriptors and their applications," in *Computer Analysis of Images and Patterns*, pp. 11–20, Springer, 2001.

[199] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582–96, Sept. 2010.

[200] H. Tamura, S. Mori, and T. Yamawaki, "Textural Features Corresponding to Visual Perception," vol. 8, 1978.

[201] H. Yu, M. Li, H.-J. Zhang, and J. Feng, "Color texture moments for content-based image retrieval," in *Proceedings of the International Conference on Image Processing*, vol. 1, pp. 929–932, IEEE, 2002.

[202] W. Zou, Z. Chi, and K. C. Lo, "Improvement of image classification using wavelet coefficients with structured-based neural network.," *International Journal of Neural Systems*, vol. 18, pp. 195–205, June 2008.

[203] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[204] M. Bober, "MPEG-7 visual shape descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 716–719, 2001.

[205] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026–1038, Aug. 2002.

[206] B. Leibe, A. Leonardis, and B. Schiele, "Robust Object Detection with Interleaved Categorization and Segmentation," *International Journal of Computer Vision*, vol. 77, pp. 259–289, Nov. 2008.

[207] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, IEEE, 1999.

[208] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–16, 2004.

[209] F.-F. Li and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 524–531, IEEE, 2005.

[210] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2169–2178, IEEE, 2006.

[211] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1794–1801, IEEE, June 2009.

[212] S.-H. Peng, D.-H. Kim, S.-L. Lee, and C.-W. Chung, "A visual shape descriptor using sectors and shape context of contour lines," *Information Sciences*, vol. 180, pp. 2925–2939, Aug. 2010.

[213] L. Juan and O. Gwun, "A comparison of sift, pca-sift and surf," *International Journal of Image Processing (IJIP)*, no. 4, pp. 143–152, 2009.

[214] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE, June 2008.

[215] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[216] A. Opelt, A. Pinz, and A. Zisserman, "A boundary-fragment-model for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 575–588, 2006.

[217] V. Ferrari, F. Jurie, and C. Schmid, "Accurate Object Detection with Deformable Shape Models Learnt from Images," in *Proceedings of the IEEE Computer Soci-*

ety Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8, IEEE, June 2007.

[218] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 17–32, May 2004.

[219] S. Maji and J. Malik, "Object detection using a max-margin Hough transform," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1038–1045, IEEE, June 2009.

[220] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE, June 2008.

[221] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 762–769, IEEE, 2004.

[222] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 886–893, IEEE, 2005.

[223] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, pp. 1475–90, Nov. 2004.

[224] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1270–81, July 2008.

[225] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Ieee, 2006.

[226] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1458–1465, Oct. 2005.

[227] D. Grangier, F. Monay, and S. Bengio, "A discriminative approach for the retrieval of images from text queries," *Proceedings of the European Conference on Machine Learning (ECML)*, pp. 162–173, 2006.

[228] O. Boiman, E. Shechtman, and M. Irani, "In defense of Nearest-Neighbor based image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE, June 2008.

[229] F. Samaria and S. Young, "HMM-based architecture for face identification," *Image and Vision Computing*, vol. 12, pp. 537–543, Oct. 1994.

[230] A. Nefian and M. Hayes, "Hidden Markov models for face recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 2721–2724, IEEE, 1998.

[231] H. Le and H. Li, "Simple 1D Discrete Hidden Markov Models for Face Recognition," in *Visual Content Processing and Representation*, pp. 41–49, 2003.

[232] S. Eickeler, S. Müller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image and Vision Computing*, vol. 18, pp. 279–287, Mar. 2000.

[233] V. Bevilacqua, L. Cariello, G. Carro, D. Daleno, and G. Mastronardi, "A face recognition system based on Pseudo 2D HMM applied to neural network coefficients," *Soft Computing*, vol. 12, pp. 615–621, Oct. 2007.

[234] M. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[235] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–591, IEEE, Mar. 1991.

[236] M. A. Turk and A. P. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, Jan. 1991.

[237] N. N. Shankar and K. R. Ramakrishnan, "Parts based representation for pedestrian using NMF with robustness to partial occlusion," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–4, IEEE, July 2010.

[238] D. Skocaj and A. Leonardis, "Robust recognition and pose determination of 3-D objects using range images in eigenspace approach," in *Proceedings of the International Conference on 3-D Digital Imaging and Modeling*, pp. 171–178, IEEE Comput. Soc, 2001.

[239] G. Casalino, N. Del Buono, and M. Minervini, "Nonnegative Matrix Factorizations Performing Object Detection and Localization," *Applied Computational Intelligence and Soft Computing*, vol. 2012, pp. 1–19, 2012.

[240] C. Au, J. Legare, and R. Shaikh, "Face Recognition: Robustness of the 'Eigenface' Approach," in *McGill Report*, pp. 1–8, 2005.

[241] A. Leonardis and H. Bischof, "Dealing with occlusions in the eigenspace approach," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 453–458, IEEE Comput. Soc. Press, 1996.

[242] J. Midgley, *Probabilistic eigenspace object recognition in the presence of occlusion.* PhD thesis, University of Toronto, 2001.

[243] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, Aug. 2004.

[244] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, Dec. 1989.

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

[246] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Mining and Knowledge Discovery*, vol. 24, pp. 136–158, June 2011.

[247] J. Barker, M. Cooke, and D. P. W. Ellis, "The RESPITE CASA Toolkit v1.3.5." http://staffwww.dcs.shef.ac.uk/people/J.Barker/ctk.html, 2002.

[248] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 845–848, IEEE, 1990.

[249] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine learning*, vol. 29, no. 2, pp. 245–273, 1997.

[250] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications.," *Neural networks: the official journal of the International Neural Network Society*, vol. 13, no. 4-5, pp. 411–30, 2000.

[251] Y. Qi, P. S. Krishnaprasad, and S. A. Shamma, "The subband-based independent component analysis," in *Proceedings of the Workshop on Independant Component Analysis*, pp. 199–204, 2000.

[252] B. Arons, "A Review of The Cocktail Party Effect The Separation of Speech Channels Early Work," *Journal of the American Voice I/O Society*, vol. 12, no. 7, pp. 35–50, 1992.

[253] S. Ntalampiras, I. Potamitis, N. Fakotakis, and S. Kouzoupis, "Automatic Recognition of an Unknown and Time-Varying Number of Simultaneous Environmental Sound Sources," *World Academy of Science, Engineering and Technology*, vol. 59, pp. 2097–2101, 2011.

[254] N. Miyake, T. Takiguchi, and Y. Ariki, "Noise Detection and Classification in Speech Signals with Boosting," in *Proceedings of the IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 778–782, IEEE, Aug. 2007.

[255] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proceedings of the International Society on Music Information Retrieval (ISMIR)*, 2007.

[256] J. D. Krijnders, M. E. Niessen, and T. C. Andringa, "Sound event recognition through expectancy-based evaluation of signal-driven hypotheses," *Pattern Recognition Letters*, vol. 31, pp. 1552–1559, Sept. 2010.

[257] H. D. Tran and H. Li, "Jump Function Kolmogorov for overlapping audio event classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, no. 1, pp. 3696–3699, IEEE, May 2011.

[258] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound Event Detection in Multisource Environments Using Source Separation," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, pp. 36–40, 2011.

[259] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry* (F. Nielsen and R. Bhatia, eds.), pp. 341–371, Springer, 2012.

[260] A. Lehmann, B. Leibe, and L. Van Gool, "Fast prism: Branch and bound hough transform for object class detection," *International journal of computer vision*, vol. 94, no. 2, pp. 175–197, 2011.

[261] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[262] J. Allen, "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[263] P. M. Baggenstoss, "Statistical modeling using Gaussian mixtures and HMMs with Matlab," in *Tech. rep., Naval Undersea Warfare Center, Newport, RI*, 2002.

[264] M. Carreira-Perpinan, "Mode-finding for mixtures of Gaussian distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1318–1323, 2000.

[265] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of Eurospeech*, vol. 7, (Geneva), pp. 1009–1012, 2003.