Sound event recognition in home environments

Ng, Terence Wen Zheng

2014

Ng, T. W. Z. (2014). Sound event recognition in home environments. Master's thesis,
Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/61720>

<https://doi.org/10.32657/10356/61720>

NANYANG
TECHNOLOGICAL
UNIVERSITY

SOUND EVENT RECOGNITION IN HOME ENVIRONMENTS

NG WEN ZHENG TERENCE
SCHOOL OF COMPUTER ENGINEERING
2014

SOUND EVENT RECOGNITION IN HOME ENVIRONMENTS

NG WEN ZHENG TERENCE

School of Computer Engineering

A thesis submitted to the School of Computer Engineering
in in fulfilment of the degree of the Master of Engineering

2014

Acknowledgements

Foremost, I would like to express my gratitude to my supervisors, Dr. Chng Eng Siong (Nanyang Technological University, NTU) and Dr. Tran Huy Dat (Institute for Infocomm Research, I²R) for their patience, support and knowledge, which guided me throughout the research. I appreciate their expertises in the area of my research which provided me with many insights to many problems, and also their assistance in the scientific writings.

A special thanks goes out to my team-mate, Jonathan Dennis, at work in Human Language Technology department of I²R. Working together under the same supervisors has allowed us to discuss many related topics which helped me along the way. Also, he is a role model for me in various aspects of life outside work.

I would also like to thank my family for their continual support throughout my entire life. Finally, I would like to thank my wife and best friend, Chen Yi-Jie, for her love and endless encouragement.

Contents

Acknowledgements	1
Abstract	6
List of Publications	8
List of Figures	10
List of Abbreviations	12
1 Introduction	13
1.1 Background	13
1.2 Problems and Motivations	14
1.3 Contributions	17
1.4 Organisation of Thesis	18
2 Literature Review	19
2.1 Overview of Sound Event Recognition System	19
2.1.1 Detection	21
2.1.2 Feature Extraction	23
2.1.3 Classification	24
2.2 Noise Robustness for Background Noise	26
2.2.1 Signal level	27
2.2.2 Feature level	28
2.2.3 Model level	29
2.3 Noise Cancellation for Interference noise	29
2.3.1 Beamforming	30
2.3.2 Adaptive Filtering	30
2.4 Summary	35

3	Proposed System for SER under TV Noise	36
3.1	Overview	36
3.2	Regression-based Noise Cancellation	38
3.2.1	Linear regression mapping	40
3.3	Spectral GMM Dectector	42
3.4	Improved Subband Power Distribution Image Feature Classification Frame- work	44
3.4.1	Improved Subband Power Distribution Image Transformation	46
3.4.2	Image Feature Extraction	50
3.4.3	Missing Feature System	51
3.5	Experiments	54
3.5.1	Setup	54
3.5.2	Results and discussions	56
3.6	Summary	59
4	Class-based Compensation for SVM with Limited Training Data	60
4.1	Overview	60
4.2	Class-based Compensation for SVM	63
4.2.1	Class-based Compensation for binary SVM	64
4.2.2	Integrating Class-based Compensation into multi-class SVM	68
4.3	Experiments	69
4.3.1	Setup	69
4.3.2	Results and discussions	70
4.4	Summary	72
5	Conclusions and Future Work	73
5.1	Conclusions	73
5.2	Future Work	74
5.3	Publications	75
	References	76

Abstract

The sounds that we hear in everyday environments contain a wide variety of acoustic information that assists us in accomplishing many daily tasks. Sound event recognition (SER) aims to automatically detect and classify these sounds to provide more information about the surroundings. This research focuses on sound events found in home environments, which has a wide range of potential applications. However the audio signal received in the home environment is completely unstructured and there exists a range of challenges faced by these applications. In this thesis, we focus on two of the commonly faced problems while performing SER in the home environments: (1) the presence of interference noise and (2) limited training data.

The problem of interference noise refers to noise which is highly non-stationary and may be regarded as a signal itself. An example of an interference noise commonly occurring in the home environment is the audio signal produced by the television (TV). Conventional noise robustness methods that aim to improve the results of sound recognition under background noise typically assume that the noise is stationary or slowly changing. These assumptions do not hold for interference noise and will not be effective if used in practice. To reduce the interference noise, many existing methods assume the use of an additional reference microphone to receive the TV signal. With the knowledge of TV reference signal, the problem is simplified to estimating the room impulse response using adaptive filtering. Instead of adaptive filtering, the approach taken in this thesis is based on a regression mapping in the frequency domain. This is called the regressive noise cancellation (RNC), which finds a global minimum for the error function instead of iteratively minimising the error function. While this is shown to improve the cancellation compared to the previous techniques, some noise remains in the form of residual noise. To address the residual noise, an existing subband power distribution image feature (SPD-IF) classification framework is employed to localize

the noise and signal into separate regions, followed by a missing feature classification performed on the reliable parts. An enhancement to the SPD-IF is proposed where the subband power distributions are estimated by utilising the temporal information across the subband. From the experimental results, the proposed RNC cancellation, together with the improved SPD-IF, outperforms several combinations of conventional cancellation and classification methods.

The second problem faced in the home environments is that of limited training data. This often worsens the performance of any recognition systems significantly. Collecting large databases has always been a big challenge as labelling is time consuming and expensive. One common way to overcome this problem is to use Semi-Supervised Learning (SSL), which utilizes an initial model trained from a small initial training data set to classify the unlabelled data. The most reliable samples are first selected and subsequently used to improve the initial model. While most research that deals with limited training data focuses on improving the performance of SSL methods, there is less attention to directly improve the feature for the case of limited training data. To this end, the approach taken in this thesis is to make the features more discriminative and for this, a class-based compensation (CBC) method is proposed. The idea of CBC is to learn a set of filters for each of the binary classes of the SVM classifier to enhance the discriminative capability of the features for classification. To enable this, CBC employs Fisher Linear Discriminant (FLD) analysis on the power spectrum distribution between the class pairs to assign higher weights to the frequency components which best discriminate the class information. Experimental results show that the compensation is able to perform well with the constraint of limited training samples. Moreover, the compensation method further improves the accuracy when used with in conjunction with previous noise robustness techniques in noisy environments. Together, the approaches presented in this thesis can form the basis of a robust SER system that performs well in home environments.

List of Publications

- (i) W. Z. T. Ng, H. D. Tran, J. Dennis, and E. S. Chng, “Robust sound event recognition under tv playing conditions,” *IEEE China Summit and International Conference on Signal and Information Processing*, 2013
- (ii) W. Z. T. Ng, H. D. Tran, J. Dennis, and E. S. Chng, “A robust sound event recognition framework under tv playing conditions,” *Asia Pacific Signal and Information Processing Association*, 2013
- (iii) W. Z. T. Ng, H. D. Tran, T. H. Huynh, and E. S. Chng, “Adaptive semi-supervised tree svm for sound event recognition in home environments,” *Asia Pacific Signal and Information Processing Association*, 2013

List of Figures

1.1	SER in the presence of TV noise using an additional reference microphone for noise cancellation	16
2.1	The structure of a typical SER system consisting of three main processes: Detection, feature extraction and classification.	20
2.2	Comparison between energy and spectral features of a bell sound both in clean conditions and corrupted by white noise at 5dB SNR	22
2.3	An example of a GMM model consisting of two mixtures to model the one-dimensional features marked on the x-axis.	25
2.4	A SVM generates a hyperplane separating the feature vectors from two classes (cross vs circles)	25
2.5	A block diagram illustrating at the various stages of noise robust techniques categories within the SER system. The three categories are signal, feature and model levels and the different techniques are optional which can be applied individually or simultaneously.	26
2.6	A block diagram illustrating adaptive filtering using an adaptation algorithm to update the filter weights from one iteration to the next, minimising the error.	31
2.7	A block diagram illustrating frequency domain adaptive filtering (FDAF): The time domain signals are transformed to frequency domain using FFT before the adaptation process. The capitalised symbols represent the signal in the frequency domain.	34
3.1	A block diagram illustrating the modules in the proposed SER system which includes the novel RNC and iSPD-IF	37

3.2	A block diagram of the steps used in RNC to remove the noise and reconstruct the final output signal.	41
3.3	A block diagram of SPD-IF/iSPD-IF classification framework consisting of three main modules: SPD image transformation, SIF image feature extraction and missing feature system. The proposed iSPD-IF improves on the SPD image transformation module.	44
3.4	Examples to illustrate the SPD representation of various sounds. Top row represents the spectrogram surface and bottom row represents the SPD representation	47
3.5	Example of PDF estimation using histogram and NP-Windows. Input B (bottom row) which is a reverberated version of signal A (top row). The NP-Windows method produces less mismatch between the clean and reverberated input as compared to histogram.	49
3.6	Partitioning of the SPD image, and extraction of distribution statistics to generate the image feature where $L_i j, color$ is the RGB color distribution of the i^{th} row, j^{th} column sub-block.	51
3.7	SPD of bell sound masked in noise at 0dB SNR(left)/high SNR conditions(right) where the yellow dotted area from (a) and (b) indicating the reliable regions for classification	52
3.8	A block diagram describing the steps of the missing feature system which are applied on the extracted image feature.	54
3.9	Bird's eye view of the recording room indicating positions of the TV, reference/operating microphones and the possible locations of the sound events	55
4.1	$N(N - 1)/2$ discriminative filters are constructed for each of the $N(N - 1)/2$ pairs of the N class SVM.	63
4.2	An illustration of the separability index for different pairs of distributions. The separability index decreases as the two distributions get more overlapped and vice versa.	65
4.3	An illustration of the proposed filter generated between two sound classes.	67

List of Abbreviations

ARMA	Auto-regressive Moving Average
AFE	Advanced Front-End
ASR	Automatic Speech Recognition
CBC	Class-based Compensation
DCT	Discrete Cosine Transform
FDAF	Frequency Domain Adaptive Filtering
FFT	Fast Fourier Transform
FLD	Fisher's Linear Discriminant
EM	Expectation-Maximisation
GMM	Gaussian Mixture Model
IFFT	Inverse Fast Fourier Transform
iSPD-IF	Improved Spectral Power Distribution Image Feature
kNN	k -Nearest Neighbours
LMS	Least Mean Squares
MAP	Maximum A-posteriori Probability
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
MMC	Maximum Marginal Clustering
MSE	Mean-Square Error
MVN	Mean Variance Normalisation
NP-Windows	Non Parametric Windows
NTU	Nanyang Technological University

OAA	One-against-all
OAo	One-against-one
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
RGB	Red, Green, and Blue
RNC	Regressive Noise Cancellation
RWCP	Real Word Computing Partnership
SER	Sound Event Recognition
SIF	Spectrogram Image Feature
SIR	Signal-to-Interference Ratio
SNR	Signal-to-Noise Ratio
SPD	Subband Power Distribution
SPD-IF	Spectral Power Distribution Image Feature
SSL	Semi-Supervised Learning
STFT	Short Time Fourier Transform
SVM	Support Vector Machines
TV	Television
VAD	Voice Activity Detection

Chapter 1

Introduction

1.1 Background

The sounds that we hear in everyday environments contain a wide variety of acoustic information which assists us in the accomplishment of many daily tasks. Among these sounds, speech is arguably the most important as we use it extensively for communication. However, beyond speech, we are capable of producing non-speech sounds like crying, coughing or clapping which convey our emotion. In addition, there are more general sounds produced by objects, for example door closing or a phone ringing, that provides us with more information and context about the environment. The research topic on sound event recognition (SER) aims to automatically detect and classify non-speech sounds to provide more information about the surroundings. The term "sound event" may refer to either a single sound class (e.g. phone ringing, bell, door knocking) or more contextual information (e.g. sports, news, crime). In this thesis, we focus on the techniques on a single sound class.

This thesis focuses on the use of SER in home environments, where there are a wide range of applications. From the literature, it is clear that these applications typically fall into three main categories:

Safety Surveillance Safety surveillance systems analyse sound events to recognise distress situations so that appropriate help can be obtained immediately. One type of such safety surveillance system aims to detect general sound activities from the home environment to track user's behaviour pattern [4, 5, 6, 7] for

elderly care applications. Also, more specific sounds from applications like scream detection [8], fall detection [9] and bathroom monitoring [10] have also been proposed.

Home Automation Home automation consists of aids that are activated by certain sound events. Home automation using audio started gaining its popularity specifically with speech [11, 12], where speech commands are used to control or interact with various appliances, such as washing machines, television and fans. Recently, sound event recognition was also proposed by Wang *et al.* [13] to trigger predefined home automation services associated with sounds such as a doorbell, glass breaking, door knocking, telephone ringing and coughing.

Social Television In addition to these well established applications, there exists a relatively new area of SER applications for home environments: Social television. The idea of social television using SER was proposed by [14, 15] to enhance television viewing with the co-presence of remote participants. The system captures engagement data using sound events which then trigger pre-defined audio cues such as laughter to be played to the remote participants.

The wide range of applications using SER in home environments affirms its practicality and usefulness. However, there exists a range of problems faced by these applications in the unstructured home environment. The objective of this research is to develop novel techniques to address these problems and robustness issues that compromise the sound event recognition task.

1.2 Problems and Motivations

Two commonly faced problems while performing SER in the home environment are the presence of interference noise and limited training data:

Interference Noise Interference noise here refers to noise which is highly non-stationary and can be regarded as a signal itself. An example of an interference noise which commonly occurs in the home environment is the signal produced by the TV. The signal emitted by the TV is altered by the room acoustics depending on the position of the microphone in the room and the signal received by the microphone,

$x(t)$, for the SER task can be expressed as:

$$x(t) = h(t) * r(t) + s(t) + n(t) \quad (1.1)$$

where $h(t)$ is the unknown room impulse response, $r(t)$ is the interference signal from the TV, $s(t)$ is the signal from target sound events, $n(t)$ is the background noise and $*$ represents the convolution operator. To reduce the interference noise, Vacher *et al.* [16] proposed using a reference microphone placed at the TV to obtain $r(t)$. An example of such a two-microphone cancellation scheme is illustrated in Figure 1.1. With the knowledge of $r(t)$, the problem is simplified to estimating $h(t)$, the room impulse response and subsequently using the estimated $h(t)$ to subtract $r(t)$ from $x(t)$.

Another solution by [14, 15] employs a hardware solution which uses a directional microphone to enhance the sound event signal for the SER tasks. However, a directional microphone is only effective for a small region of the room and does not work throughout all locations of the room. Apart from these solutions, other SER noise robustness techniques e.g. [13, 8] focuses mainly on background noise and hence is not effective for interference noise such as TV noise. With the lack of research to deal with interference noise, this motivates us to find a solution using techniques apart from the conventional methods. One good direction to proceed is to adapt the framework in [16] who utilize a reference microphone to capture the interference signal at its source. This interference signal can then enhance the received signal provided the room impulse response is well estimated. In this work, we aim to improve the estimate of $h(t)$ and investigate better feature-classifier combinations for the SER task.

Figure 1.1: SER in the presence of TV noise using an additional reference microphone for noise cancellation

Limited Training Data Assessing or collecting large databases has always been a big challenge. Although large amounts of data can be cheaply and automatically collected, subsequent labelling is very time consuming and expensive to obtain. With limited training data, it is well known that the performance of any recognition tasks significantly worsens.

One common way to overcome the problem of limited training data is to use Semi-Supervised Learning (SSL) [17] which utilizes an initial model trained from the small training data set to classify the unlabelled data. During classification of the unlabelled data, a confidence measure is used to identify the reliable samples that can be used to improve the initial model. By using the reliable samples to further train the initial model, it is possible to improve the representation of the model. The Support Vector Machine (SVM) classifier is commonly used to represent the SSL models as (1) it usually requires a smaller set of training data to estimate the initial model, (2) the distance from the test vector to the SVM hyperplanes is a good confidence measure to identify the reliable samples.

SSL can be regarded as an add-on method to improve any existing feature-classifier combination that addresses the problem of limited training data. While most research that deals with limited training data focuses on improving the per-

formance of SSL methods, there is less attention to directly improve the feature for the limited training data. This work aims to improve the features by making them more discriminative.

1.3 Contributions

The author has contributed the following original and novel research:

- Regressive noise cancellation(RNC): a preprocessing method which utilises an additional reference microphone to effectively cancel the TV noise. The RNC uses an empirical mapping between the spectral powers of reference microphone and the operating SER microphone for cancellation. Specifically, the linear regression mapping is chosen in this work and a closed formed solution is obtained. This is published in [\[1\]](#) and will be presented in Chapter [3.2](#).

- A novel method to enhance the classification of audio signals. We propose an improved subband power distribution image feature (iSPD-IF) which is an extension of an existing classification framework proposed in [18]. In the feature extraction process of [18], the probability density function(PDF) of the sub-band spectral powers are calculated using conventional histogram. In this work, we propose to improve the estimate of the PDF using the temporal information across the sub-bands. This is published in [2] and will be presented in Chapter 3.4.1.
- A class-based compensation for SVM: we propose a filter which can be used on spectral based features to increase the discriminability between two classes. The filters enhance the discriminative capability between two classes of a binary SVM by giving higher weights to the frequency components which best discriminates the classes and vice versa. The class-based compensated feature is subsequently integrated into a multi-class SVM to improve the SER performance in situations where the training data is limited. This is published in [3] and will be presented in Chapter 4.2.

1.4 Organisation of Thesis

The rest of this thesis is organised as follows: Chapter 2 gives an overview of the typical SER framework and an in-depth survey of the commonly used modules like detection, feature extraction and classification is included. In addition, different types of noise robust techniques including those designed for background and interference noise are introduced. Chapter 3 then proposes a system solution for the first problem of performing SER in the presence of interference noise. The proposed system includes (1) a novel preprocessing method called as regressive noise cancellation (RNC) which utilizes a reference microphone for noise cancellation and (2) an improved subband power distribution image feature(iSPD-IF) classification framework for noise robust SER. Chapter 4 introduces an class-based compensation feature for SVM to increase the discriminability between the classes which improves the classification performance despite the limited training data. Chapter 5 concludes the thesis and proposes future work.

Chapter 2

Literature Review

In this chapter, a literature review of relevant topics is discussed. In the first section, an overview of typical sound event recognition system and its various modules is presented. The subsequent section describes existing noise robust techniques to improve the SER accuracy in the presence of stationary or slowly varying background noise. In the last section, noise cancellation techniques are presented that specifically deal with interference noise found in home environments.

2.1 Overview of Sound Event Recognition System

The aim of a sound event recognition (SER) system is to correctly detect and classify an unknown audio segment into a predefined sound class. A SER system must therefore be trained prior to being used for testing. The training process learns a model from a training data set, which consists of audio clips of the predefined classes and their class labels. Using the trained model, a testing process evaluates an unknown testing sample to predict the class label. Both processes typically consist of the three main modules shown in Figure 2.1: detection, feature extraction, classification. This structure has been implemented by various authors such as in [19, 20, 21].

The detection module is the first step for both training and testing processes. In an automatic speech recognition (ASR) system, the detection module is the voice activity detector (VAD) which aims to detect speech from non-speech [22]. Similarly, in a SER system, the detector attempts to segment the audio stream into isolated sound events.

Figure 2.1: The structure of a typical SER system consisting of three main processes: Detection, feature extraction and classification.

Isolated events are required for the classification step as the classifier only associates each segment with a single class label.

After the audio signals are segmented, a feature extraction is utilised for both training and testing process. The feature extraction step transforms the audio segment into a suitable domain which compactly and discriminatively retains the most useful information for sound event classification. Most of the methods used for audio feature extraction were originally proposed for ASR. These features were subsequently used for SER and also achieved good performance [23, 24].

In the last step of the system, the classification module operates on both the training and testing process, but with some variation. In the training process, the features extracted from the training signals, together with their respective ground truth class labels are used to train a model. The role of the trained model is to capture the distribution of each sound class's feature space. In the testing process, an unknown test feature is classified using the various trained models to predict the class label for the sound event. Various cost functions such as minimum classification error (MCE) [25], maximum likelihood (ML) [26], Maximum A posteriori Probability (MAP) [27] have been proposed to train the models as well as performing classification.

In the rest of this section, a review of the conventional approaches to these modules is provided, where the merits and limitations are presented.

2.1.1 Detection

The detection module is the first important step of a SER system. The function of the detection module is to segment the input into isolated sound events without any silence period. A good segmentation prevents noise or multiple events from being included in the classification, as these will deteriorate the performance of the classifier. Multiple events including overlapped sound events which requires other specialised methods is not considered in this work. The detection module typically consists of two subtasks: Extraction of acoustic features and the decision rules.

The features used for any sound activity detection should discriminate between sound events and noise. The most intuitive feature for detection is the energy of the signal, which physically represents the loudness of the sound. Humans naturally associate a sound which is louder than the background noise as an event of interest. The energy feature is shown to be effective for detecting sound events in an audio-based surveillance system in [28]. One drawback of the energy feature is that it fails to work in situations when the background noise loudness is of similar or higher levels compared to the sound events. Figure. 2.2a and Figure. 2.2b illustrates the energy levels of a bell in clean conditions and corrupted by white noise at 5dB SNR. It is observed that in such noisy conditions, it is hard to detect the signal as it is heavily masked by noise.

A robust method to overcome high energy noise is to use spectral-domain features, where the frequency content of the signal over time is utilized. An efficient approach to transform the signal to the spectral domain is by Fast Fourier Transform (FFT) [29]. Recall that the energy feature fails to discriminate an event when the noise which has equal or higher energy levels. However in the spectral domain, the frequency content of the sound events may not be completely overlapped by the noise and as such certain frequency bins that belong to the sound event of interest will have higher power. The spectrograms of the signals belonging to 2.2a and Figure. 2.2b are illustrated in 2.2c and Figure. 2.2d respectively. It can be seen in 2.2d that the the bell's spectrum is easily distinguished from the noise. Nonetheless, the ability to robustly detect is highly dependent on the amount of overlap between the noise and the sound event in the spectral domain.

Another commonly used feature for detection is based on the harmonicity of speech

Figure 2.2: Comparison between energy and spectral features of a bell sound both in clean conditions and corrupted by white noise at 5dB SNR

signals [30], particularly used for voice activity detection (VAD), since human speech has a unique structure which contains multiple harmonics of the fundamental frequency F_0 . One advantage of this structure is that the harmonic shape is retained even in noisy conditions, which remains robust for detection [31]. However, this method assumes the existence of harmonics in sound events, which does not apply to all types of sound events.

After a set of detection features has been extracted from the signal, a decision rule is used to decide if the signal is an event or noise. The simplest way is to use a simple distance measure and a threshold, where the decision boundary is estimated empirically based on the distribution of the extracted features in the training data set.

A test feature vector $\mathbf{y} \in \mathbb{R}^n$ is decided as an event if it exceeds the threshold, and noise otherwise [32]. Besides using a threshold, machine learning techniques can also be used to provide decision rules for the detection. Popular methods include Support Vector Machine(SVM) and Maximum Marginal Clustering(MMC) have been shown to improve performance over using thresholds [33, 34].

2.1.2 Feature Extraction

The purpose of feature extraction is to express an audio signal into a new representation which better characterises the relevant sound event information. The new representation should provide large discrimination between different sound classes while having a small variation within class.

A common approach to extract audio features is by using a frame-based technique, where short frames of around 16-32ms are characterised by the spectral shape [35]. The most popular frame based features are Mel-Frequency Cepstral Coefficients (MFCCs) [36] and Perceptual Linear Prediction (PLP) [37]. Both representations generate a description of the short-time spectral shape of the input signal that is based on auditory-like signal processing. The performance of several spectral features has been compared in [38] and MFCCs generally outperform PLP and other frame-based features. While these features were originally designed for ASR systems, they are also applicable for SER as they capture useful spectral information of sounds. Particularly, the previous work in [39] showed that MFCCs can be successfully used for SER.

To extract MFCCs, each windowed frame of the audio signal is first transformed to the spectral domain using the Fast Fourier Transform (FFT). A Mel-scaled triangular filterbank is then applied to produce a smoother spectral representation that is less sensitive to pitch harmonics. This operation mimics the human frequency warping of the cochlea. The result is a feature that models the sensitivity of the human ear. The features are then decorrelated by taking the Discrete Cosine Transform (DCT) and the temporal dynamics of the feature vectors, commonly known as the deltas and delta-deltas, are generated by appending its first and second derivatives [40].

2.1.3 Classification

The classification module is the final step of the SER system. After the feature vectors are extracted from the audio segments, the classification module matches the testing feature vectors against training features to predict the correct label. An effective way to perform the matching is by training a model that represents the training data's feature space, and using that model to find the best match. This is the idea behind many popular techniques [41] including Gaussian Mixture Models (GMM) and Support Vector Machines (SVM) which are briefly introduced below.

Gaussian Mixture Model This approach has been widely used for audio processing classification problems such as speaker identification [42], language identification [43] and audio fingerprinting [44]. This method models the feature space using Gaussian density components. The simplest form of this method is to use a Gaussian distribution with only one mean and variance parameter, however, this basic form does not always fit well with the data. For a better estimate of the distribution of the features, the GMM uses a mixture of Gaussian density component which consist of more pairs of mean and variance parameters as illustrated in Figure. 2.3. The mean and variance parameters can be estimated iteratively using the Expectation-Maximisation (EM) algorithm [45] on the training feature vectors to maximise a criterion such as maximum likelihood (ML). After the GMM model for each class is trained, the likelihood score for a test feature vector can be evaluated for each of the class models. The model that generates the highest score is assigned as the predicted class.

Support Vector Machine This has been widely used to solve signal processing related classification problems like speaker identification [46], face recognition [47] and audio retrieval [48]. SVM is a binary classifier that separates feature vectors belonging to two classes using a hyperplane in a high-dimensional space [49, 50]. SVM is designed with an optimisation criterion that will generate a hyper-plane that best separates the two clusters, i.e. the margin between the closest points the two classes will be maximum. An example of a classification problem with 2-dimensional feature vectors is visualized in Fig 2.4. Given the separating hyperplane, a test feature vector is classified into one of the class depending on which side the feature is located.

Figure 2.3: An example of a GMM model consisting of two mixtures to model the one-dimensional features marked on the x-axis.

Figure 2.4: A SVM generates a hyperplane separating the feature vectors from two classes (cross vs circles)

There are two methods to extend this binary SVM to a multi-class SVM, namely the one-against-one (OAO) [51] and one-against-all(OAA) method [52]. In the OAO method, $N(N - 1)/2$ binary SVM classifiers are constructed, using all the pair-wise combinations of the N classes. The resultant class is the class which gets the majority vote among the $N(N - 1)/2$ binary SVM classification decisions. In the OAA method, N binary SVM classifiers are constructed, where each of the binary SVM classifiers uses one of the N class as positive examples and all the remaining classes as the negative example. The resultant class is the class with the largest margin to the hyperplane.

2.2 Noise Robustness for Background Noise

In section 2.1, the sequence of modules used in a typical SER system was introduced. Most of these SER systems are trained with data recorded with a high signal-to-noise ratio (SNR). However, if the systems are tested in noisy conditions, previous studies have shown that decreasing SNR reduces the classification accuracy [53]. The reduction of performance is due to the mismatch between the training and testing feature vectors, which are distorted or masked by the noise, thus yielding mismatch models of the target class.

One simple and effective solution for background noise robustness is to perform multi-conditional training [54], where the SER models are trained on data with different types of noise conditions and SNRs. This process generates acoustic models with prior information of the target sound events for the given noise conditions. These generalized models often improve the classification accuracy when tested with noisy samples. However, multi-conditional training results in a ‘smoothed out’ version of the model which slightly reduces the classification accuracy for high SNR conditions [55].

Other conventional techniques for background noise robustness typically fall into three categories as shown in Figure 2.5: (1) enhancement at the signal level before feature extraction, (2) compensation at the feature level, and lastly, (3) adaptation at the model level in the classifier. The rest of this section reviews the most common techniques in each of these three areas.

Figure 2.5: A block diagram illustrating at the various stages of noise robust techniques categories within the SER system. The three categories are signal, feature and model levels and the different techniques are optional which can be applied individually or simultaneously.

2.2.1 Signal level

Noise robustness applied at the signal level attempts to enhance a noisy signal by reducing noise in the time domain before the feature extraction process as illustrated in the green box of Figure. 2.5. Common alternative descriptions of this method include ‘preprocessing’, ‘front-end processing’ or ‘signal enhancement’. The objective is to modify the signal at the temporal level to be as close to the clean training signal as possible.

An early method of signal enhancement is the spectral subtraction (SS) method, which proved to be very effective for ASR systems [56]. In this method, the noise power spectrum is estimated from noise only frames and this noise is subtracted from the noise corrupted signal in each frequency bands, f , with the following equation:

$$|\hat{X}_{SS}(f)| = \left(|Y(f)|^a - b \overline{|N(f)|^a} \right)^{\frac{1}{a}} \quad (2.1)$$

where $|\hat{X}_{SS}(f)|$ is the estimated clean signal’s spectral amplitude, $|Y(f)|$ is the spectral amplitude of noise corrupted signal, $|N(f)|$ is the spectral amplitude of noise, a and b are parameters and $\overline{[\cdot]}$ represents the time averages. One drawback of this method is that the non-linear operation often generates residual ‘musical’ noise, which requires further post processing like smoothing. In addition, a reliable noise detector to estimate the noise statistics is also crucial for this method.

A related method is Wiener filtering. In this approach, a linear filter is designed to minimise the expected squared error between the clean and filtered noisy speech signal. The Wiener filter [57], $W(f)$, in the frequency domain is expressed as:

$$W(f) = \frac{E[|Y(f)|^2] - E[|N(f)|^2]}{E[|Y(f)|^2]} \quad (2.2)$$

where $E[\cdot]$ denotes the expected value. With the formulated Wiener filter, the noise power spectrum is estimated as:

$$|\hat{X}_{Wiener}(f)|^2 = W(f)|Y(f)|^2 \quad (2.3)$$

The above equation is related to spectral subtraction as Equation 2.1 can be expressed

as:

$$|\hat{X}_{ss}(f)|^2 = \frac{|Y(f)|^2 - \overline{|N(f)|}^2}{|Y(f)|^2} |Y(f)|^2 \quad (2.4)$$

when $a = 2$ and $b = 1$. In this case, it means that in Wiener filtering, the ensemble averages are estimated for both the noise and noise signal spectrum. However in spectral subtraction, time averages are used to estimate the noise spectrum and the instantaneous noise spectrum, but the ensemble average is not used. An advanced version of signal enhancement involving a two stage Wiener filter is applied in the ETSI Advanced Front-End (AFE) [58] standard for noise reduction on ASR systems. The drawback of all these methods is that they become less effective when the characteristics of the corrupting noise becomes more non-stationary.

2.2.2 Feature level

Noise robustness at the feature level as illustrated in the purple box of Figure. 2.5 can be categorised into two methods. The first class of method is by using feature compensation, where feature vectors are normalised so that compensated feature vectors becomes less affected by noise. The most commonly used normalisation is the mean and variance normalisation (MVN) method [59]. MVN is a simple but effective method to reduce the mismatch between the training and testing feature vectors given different channels. Another popular method of feature compensation is auto-regressive moving average (ARMA) filtering [60], which has been shown to achieve good improvements when used with speech features for ASR. Apart from these popular methods, there are a host of other feature compensation methods which are summarized in [61].

The second method of feature level noise robustness is the missing feature approach. The missing feature approach splits the feature vector into reliable and unreliable components [62, 63]. The unreliable components are then ‘masked’ out to remove their effects. Classification is then performed by ‘imputation’ which restores the unreliable components using suitable replacements based on the clean models [64]. Another way to perform masking is ‘marginalisation’ which factors out the distribution of the unreliable components of the observed spectral information [62].

2.2.3 Model level

At the model level, noise robustness can be improved using model adaptation which transforms the distributions of the acoustic models trained on clean data to become closer to the distribution of the incoming noisy features as illustrated in the orange box of Figure. 2.5.

One popular model adaptation technique is Maximum Likelihood Linear Regression (MLLR) [26] which can be applied to GMM systems. In MLLR, the means of the Gaussians distributions are rotated and/or translated by an affine transformation to better fit the statistics of the noisy data. The parameters of this transformation are learned by an iterative EM-algorithm that maximises the likelihood of the adaptation data. Another popular method is the Maximum A posteriori Probability (MAP) adaptation [27, 65], which iteratively re-estimates the model parameters such that their a posteriori distribution is maximised for the noisy data. MAP generally requires larger amount of adaptation data to be effective as there are more parameters to be estimated [66] as compared to MLLR which is more effective for small amount of adaptation data. Apart from these popular methods, there are a host of other model adaptation methods which are summarized in [67].

2.3 Noise Cancellation for Interference noise

The noise robustness methods described in the previous section generally improve the results of sound recognition under background noise. However, these methods assume that the noise conditions during testing are stationary or slowly changing. These assumptions do not hold for interference noise such as sound playing from a radio or television (TV), which is highly non-stationary and directional. Such types of interference noise can be regarded as a target signal and therefore require specific approaches to noise cancellation to effectively handle the noise.

With the presence of signal-like interference, additional information is needed to overcome this problem. There are two main types of multi-channel techniques: (1) beamforming: which uses a microphone array to exploit the spatial information for enhancing the signal, (2) adaptive filtering: a method which actively cancels the noise by placing an additional reference microphone near the noise source. These two techniques

are introduced in the rest of this section.

2.3.1 Beamforming

Beamforming uses an array of microphones to form a response pattern with higher sensitivity in desired directions. To form a specific response pattern, the propagation of sound is modified by a spatial filter. Beamforming is widely used in applications like speaker diarization [68], speech enhancement [69] and hands-free telecommunication [70].

A conventional method of beamforming is the delay-and-sum beamformer, where all weights of spatial filter are equal and is steered to a specific direction by selecting appropriate phases. A more robust method is the Capon beamformer [71], a data dependant method which maximises the power in the output signal in addition to steering to a specific direction. The advantage of the Capon beamformer is that it has noise reduction capabilities which output of higher signal-to-noise ratio (SNR) signal compared to the conventional delay-and-sum beamformer.

For the problem of reducing the TV noise, prior knowledge of the position of the TV source is required. This is feasible in the home environments since the position of the TV is often fixed. In addition, prior knowledge of the location of the sound event source is also required for the best beamforming performance. However, the location of the sound event is usually not fixed and therefore not readily available. Also, the beamforming performance highly depends on the number of microphones to be effective [72]. The increased number of microphones then requires specialised multi-channel hardware, which can be costly. With these constraints, although beamforming is useful for many applications, it is not entirely suitable for the application of SER under TV playing conditions.

2.3.2 Adaptive Filtering

Adaptive filtering utilizes a filter that adjusts its weights according to an optimization algorithm driven by a feedback error signal. The block diagram of a typical system is illustrated in Figure 2.6. It is effective in many applications such as electrocardiogram (ECG) analysis [73], and also more generally for noise cancellation tasks [74]. From

the discussions of SER in home environments illustrated in Figure 1.1, the room impulse response is needed to be estimated for subtracting off the TV reference signal. One effective way to estimate the unknown room impulse response with an additional reference microphone is by using adaptive filtering.

Figure 2.6: A block diagram illustrating adaptive filtering using an adaptation algorithm to update the filter weights from one iteration to the next, minimising the error.

The estimation problem is solved by designing the filter in such a way that the difference between $d(n)$, the sample value of the desired signal at time n , and the filter output is made as “small” as possible in a statistical sense. The difference,

$$e(n) = d(n) - y(n) \quad (2.5)$$

is called the estimation error where the filter output $y(n)$ is the convolution sum of the filter input and estimated impulse response, i.e:

$$y(n) = \sum_{m=1}^M w_m u(n-m) = \mathbf{u}^T(n) \mathbf{w} \quad (2.6)$$

where $u(n)$ is the input signal at time sample n , $\mathbf{u}(n) = [u(n), u(n-1), \dots, u(n-M+1)]$ at time n , $\mathbf{w} = [w_1, w_2, \dots, w_M]$ are the filter tap weights of the estimated room impulse response and M is the filter length. In order to estimate the unknown room impulse response, the desired signal is assigned as the signal received by SER microphone and

the input signal is assigned as TV signal received by the reference microphone.

To make the error as “small” as possible in a statistical sense, the minimum mean squared error criterion is used to optimize the filter. Specifically, the filter tap weights, $w_1, w_2, \dots, w_M$ are chosen so as to minimize an index of performance, J , defined as the mean square error:

$$J(\mathbf{w}) = E[e(n)^2] = E[(d(n) - \mathbf{u}^\top(n)\mathbf{w})^2] \quad (2.7)$$

$$= E[d(n)^2] + E[\mathbf{w}^\top \mathbf{u}(n) \mathbf{u}^\top(n) \mathbf{w}] - E[d(n) \mathbf{u}^\top(n) \mathbf{w}] - E[\mathbf{w}^\top \mathbf{u}(n) d(n)] \quad (2.8)$$

$$= E[d(n)^2] + \mathbf{w}^\top E[\mathbf{u}(n) \mathbf{u}^\top(n)] \mathbf{w} - 2\mathbf{w}^\top E[\mathbf{u}(n) d(n)] \quad (2.9)$$

$$= \sigma_d^2 + \mathbf{w}^\top \mathbf{R} \mathbf{w} - 2\mathbf{w}^\top \mathbf{p} \quad (2.10)$$

where $E[\cdot]$ is the expected value, σ_d^2 is the variance of the desired response, $\mathbf{R}$ is the correlation matrix of the tap-input vector $\mathbf{u}(n)$, and $\mathbf{p}$ is the correlation vector between the tap-input $\mathbf{u}(n)$ and the desired response $\mathbf{d}(n)$. By minimizing J , the best or optimum filter is obtained in the minimum mean square sense. A closed form solution to minimise J is derived in [57] and the optimum weight $\mathbf{w}_o = \mathbf{R}^{-1}\mathbf{p}$. In practice however, prior knowledge of $\mathbf{R}$ and $\mathbf{p}$ are not readily available and therefore, an alternative way to obtain the optimal weights is by the method of steepest descent [75]. The method of steepest descent adapts the filter weights using the gradient of the quadratic performance surface in Equation 2.7. The iterative gradient descent algorithm updates the weights by taking a step in the direction of the negative gradient of the following objective function:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \frac{\mu}{2}[-\nabla(n)] \quad (2.11)$$

where $\mathbf{w}(n)$ is the filter weights at time n , μ is the step size parameter and $\nabla(n)$ is the gradient of the quadratic surface at time n . By updating the weights iteratively, the weights will eventually converge to the global solution which corresponds to the optimum weights. In the case where the input signals are stationary, the gradient of the mean-squared error cost function $J(w(n))$ is

$$\nabla(n) = \frac{\partial J(\mathbf{w}(n))}{\partial \mathbf{w}} = \frac{\partial E[e^2(n)]}{\partial \mathbf{w}} = 2E \left[\frac{\partial (d(n) - \mathbf{u}^\top(n)\mathbf{w})}{\partial \mathbf{w}} e(n) \right] = -2E[\mathbf{u}(n)e(n)] \quad (2.12)$$

To update the weights using steepest descent, the mean values of the product between the input and error signal needs to be estimated. As the true mean values of the gradient is not readily available, the instantaneous gradient is used in practice. The instantaneous gradient estimation technique used with steepest decent method is called the least mean-square (LMS) algorithm. The LMS algorithm [76, 77] is the most widely used adaptive filtering algorithm, being employed in several communications systems. It has gained popularity due to its simplicity, ease of computation and proven robustness. Instead of using short term averages to estimate the gradient derived in Equation 2.12, the LMS algorithm uses the instantaneous gradient to update the weights for the steepest descent method. A simple estimate of the instantaneous gradient is

$$\hat{\nabla}(n) = -2\mathbf{u}(n)e(n) \quad (2.13)$$

This instantaneous gradient, $\hat{\nabla}(n)$, is an unbiased estimator for $\nabla(n)$. Substituting the instantaneous gradient into Equation 2.11 for the weight update equation, the final LMS algorithm becomes:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu\mathbf{u}(n)e(n) \quad (2.14)$$

where μ is the step-size, a weight update parameter. The step-size, μ , controls the speed of convergence. In the derivation for the convergence criteria for the LMS algorithm, there is a fundamental assumption [57] such that each input signal, $u(n)$, is statistically independent of all previous sample vectors, i.e:

$$E[\mathbf{u}(n)\mathbf{u}^\top(k)] = 0, \quad k = 0, 1, \dots, n-1 \quad (2.15)$$

This clearly does not hold for signals recorded at the TV. Since this fundamental assumption is invalid, if the LMS algorithm is applied directly, convergence is not guaranteed. To make the assumption more valid, a possible way is to transform the input signal into a different domain such that the input signal becomes less correlated and hence more statistically independent.

One efficient and widely used method is the frequency-domain adaptive filters (FDAF) which transforms the signal representation from time domain to frequency domain. The idea behind frequency-domain adaptive filters (FDAF) [78, 79, 80] is

Figure 2.7: A block diagram illustrating frequency domain adaptive filtering (FDAF): The time domain signals are transformed to frequency domain using FFT before the adaptation process. The capitalised symbols represent the signal in the frequency domain.

similar to the conventional time domain adaptive filters. The main difference of FDAF is that before adapting the weights to minimise the error, the input channels are first transformed to the frequency domain using FFT [29] as illustrated in Figure 2.7. The advantage of transforming into the frequency domain representation is that the filter bank structures generate signals that are more uncorrelated [79]. The uncorrelated property decreases ‘eigen-spread’ of the input signal’s correlation matrix. Consequently, the convergence rate is improved as the speed of gradient descent method is proportional to the ‘eigen-spread’ [57]. Also, the excess mean-square error (MSE) which is inversely proportional to the ‘eigen-spread’ is reduced.

For our application of adapting interference signals like from TV, the uncorrelated property between the filter banks structures in the frequency domain makes the fundamental assumption more valid. This allows the weights to converge if the step-size is set appropriately. From our preliminary experiments on performing adaptive filtering on the TV signals, it was observed that the LMS algorithm does not converge using time domain representation, whereas it does converge in the frequency domain representation.

2.4 Summary

This chapter first presents an overview of a typical SER system consisting of the detection, feature extraction and classification modules. A review of the conventional approaches to these modules is provided, where the merits and limitations are presented. The next section introduces noise robustness for both background noise and interference noise. Techniques presented on noise robustness for background noise includes signal enhancement at the signal level, feature compensation at the feature level and adaptation at the acoustic models level. The last section introduces noise cancellation for interference noise like TV noise which includes multi-channel techniques like acoustic beamforming and adaptive filtering. In the following chapter, we propose a robust SER system which overcomes the limitations of the conventional methods for home environments.

Chapter 3

Proposed System for SER under TV Noise

This chapter presents a proposed system for the problem of performing SER in the presence of interference noise, specifically television (TV) noise. The first section gives an overview of the modules used in the proposed system and justifies the choices of modules used for each specific task. The modules include the proposed regressive noise cancellation (RNC) which is a preprocessing method which utilises an additional reference microphone to effectively cancel the TV noise. Also, the proposed improved subband power distribution image feature (iSPD-IF) is used to enhance the classification of audio signals where residual noise remains after cancellation. In the last section, experiments are carried out and compared with conventional methods and presented to validate our system.

3.1 Overview

This section presents an overview of the proposed system to perform SER in the presence of TV noise. Figure 3.1 shows the proposed system consisting of a sequence of modules similar to a typical SER structure as described in Section 2.1. One key difference in the proposed system is that an additional preprocessing step to address the TV noise before detection is performed. Typical noise robustness methods described in 2.2 assume the noise sources to have either a fixed or slowly changing statistics and

thus are not effective for such interference noise.

Figure 3.1: A block diagram illustrating the modules in the proposed SER system which includes the novel RNC and iSPD-IF

In the proposed system, we assume the use of additional microphones is possible to overcome this limitation. With this assumption, an alternative solution using beamforming with a microphone array is possible as previously described in 2.3.1. Although the beamforming method is able to reduce the interference noise by exploiting the spatial information, the performance highly depends on the number of microphones [72]. The increased number of microphones requires specialised hardware which can be costly and thus not practical. Instead, the proposed system is designed such that it simply requires an additional reference microphone in the preprocessing module, which is used to remove the interference signal. Specifically, a proposed regressive noise cancellation (RNC) module, motivated by the frequency domain adaptive filtering (FDAF) method [79] introduced in 2.3.2 where one microphone is used for reference and one microphone is used for the SER system. The advantage of our proposed RNC approach is that there is no assumption for the reference signal to be stationary like in conventional adaptive filtering. Instead, an empirical mapping is used for the cancellation, which improves both the cancelled signal and the SER accuracy when compared to conventional adaptive filtering. The proposed RNC will be discussed in section 3.2.

After preprocessing by RNC, the output signal is passed to the detection module for sound activity detection. Although in principle the RNC output should be close to

the clean conditions, it usually contains some residual noise. The residual noise causes spikes in the energy of the output signal, therefore conventional energy-based features would not be a reliable feature for detection. Instead, the full spectral feature is chosen to model the residual noise statistics using a Gaussian Mixture Model (GMM). The details of this spectral GMM detector module will be discussed in section 3.3. The detected audio segments are then further processed for feature extraction and classification.

To overcome the residual noise problem present in the RNC output for classification, the improved subband power distribution image feature (iSPD-IF) classification framework is proposed. The iSPD-IF is an extension of the subband power distribution image feature (SPD-IF) introduced in [18]. The basic idea of the SPD-IF is to transform the spectrogram into a new image representation, called the subband power distribution (SPD), representing the probability density function (PDF) of the power spectrum for each subband. With this new representation, the residual noise and signal are better localized and separable in each subband and this allows a missing feature classification is applied on the reliable parts using k -Nearest Neighbours [81]. The iSPD-IF improves the image feature extraction module over the original SPD-IF. We present the details of the proposed iSPD-IF classification module later in section 3.4. Experiments were performed using a standard dataset of sounds commonly found in home environments. The results show that the iSPD-IF classification framework gives a good improvement over methods using conventional MFCC features with several popular classifiers like SVM and GMM.

3.2 Regression-based Noise Cancellation

This section discusses the proposed regression-based noise cancellation (RNC) method as a preprocessing method to reduce interference noise. The setup of RNC is similar to adaptive filtering where one microphone is used as a noise reference which captures the noise source (i.e. TV, radio, CD player), and another microphone is used to capture the desired sound event as previously illustrated in 1.1. The idea of RNC is to estimate a mapping function between the reference and operating microphone in the absence of sound event during a calibration period. This mapping function basically captures the room impulse response, $h(t)$, between the TV noise to the received microphone. The

learnt function is then used to actively cancel the interference noise from the corrupted signal of the operating microphone.

To formulate a relationship between the microphones for the mapping, we first observe that the physical model [82] between the operating and reference microphones are related by:

$$x(t) = \begin{cases} h(t) * r(t) + s(t) + n(t) & \text{with sound event} \\ h(t) * r(t) + n(t) & \text{without sound event} \end{cases} \quad (3.1)$$

where $x(t)$ is the signal received by the SER operating microphone, $r(t)$ is the TV signal received by the reference microphone, $s(t)$ is the signal from target sound events, $n(t)$ is the background noise in the room, $h(t)$ is the unknown room impulse response and $*$ represents the convolution operator.

Assuming an initial period that is free from any sound events, we use this as a calibration to find the empirical mapping for RNC. The first step is to estimate the delay d between the microphones using the autocorrelation function:

$$d = \arg \max_{\tau} E[x(t)r(t + \tau)] \quad (3.2)$$

After removing the delay between the two inputs, the inputs are converted into their power spectrum domain by taking the windowed STFT and can be expressed as:

$$P_X[t, k] \approx \begin{cases} |H[t, k]|^2 P_R[t, k] + P_S[t, k] + P_N[t, k] & \text{with sound event} \\ |H[t, k]|^2 P_R[t, k] + P_N[t, k] & \text{without sound event} \end{cases} \quad (3.3)$$

where H is the unknown frequency response characterizing the transfer function (acoustic and channel) between the reference and SER operating microphone. P_R and P_X denotes the spectral powers of the reference and SER operating microphone respectively. P_S and P_N denotes the spectral powers of an sound event and background noise respectively. Symbols t and k denote the frame index and frequency bin, respectively. Due to the windowing effect and room attenuation, $H[t, k]$ is not a constant but more like a random variable distributed around its mean value.

In the power spectrum domain, since the unknown frequency response, H , is multiplicative instead of convolutive, this allows us to estimate H separately in each of

the respective frequency bins when there is no sound event. For each fixed frequency bin, k , the aim is to find a function G under the “without sound event” condition, such that:

$$G_k = \arg \min_g \int_t |P_X[t, k] - g(P_R[t, k])|^2 dt \quad (3.4)$$

The function G is an empirical function to estimate the unknown frequency response. With this function, the reference signal can be mapped and cancelled from the operating SER microphone, leaving only the target sound event.

3.2.1 Linear regression mapping

In general, the function G can be any one-to-one function of any form, analytic or non-analytic. One simple choice of a function for G is the linear model, i.e:

$$G(P_R[t, k]) = c_1[k]P_R[t, k] + c_2[k] \quad (3.5)$$

where c_1 and c_2 are the mapping coefficients. Since Equation 3.3 is an approximation due to the STFT windowing and also room attenuation, therefore c_2 can be regarded as a bias to compensate these effects.

By utilising a short calibration period, where only the interference source is active, the mapping coefficients c_1 and c_2 in each frequency bins can be learnt. The power spectrum of both the reference, $\bar{P}_R = P_R[1, k], P_R[2, k] \dots P_R[M, k]$ and the operating microphone, $\bar{P}_X = P_X[1, k], P_X[2, k] \dots P_X[M, k]$ are accumulated over the initial frames respectively for each frequency bins. Using the frames from the calibration phase, a closed form solution [83] for each frequency bin k can be derived as follows:

$$[c_1(k) \quad c_2(k)]^\top = (\bar{R}^\top \bar{R})^{-1} \bar{R}^\top \bar{P}_X^\top \quad (3.6)$$

where $\bar{R} = [\text{ones}(M, 1)^\top \bar{P}_R^\top]^\top$ and M is the number of frames. A more reliable estimate of c_1 and c_2 can be formulated by weighting the samples. Since the samples with higher power are usually more reliable, we can give additional weights on these samples and modify the above equation as follows:

$$[c_1(k) \quad c_2(k)]^\top = (\bar{R}^\top W \bar{R})^{-1} \bar{R}^\top W \bar{X} \quad (3.7)$$

Figure 3.2: A block diagram of the steps used in RNC to remove the noise and reconstruct the final output signal.

where W is the weighting function based on the power of the reference, i.e. $W = \bar{P}_R$. By rearranging Equation 3.3 in the presence of sound event result in the following equation:

$$P_S[t, k] \approx P_X[t, k] - (|H[t, k]|^2 P_R[t, k] + P_N[t, k]) \quad (3.8)$$

Therefore by utilizing the learnt mapping function, the desired output signal power, S , can be obtained by cancelling the mapped interference from the noisy signal at the operating microphone:

$$P_S(t, k) = \max(P_X[t, k] - \{c_1(k)P_R[t, k] + c_2(k)\}, 0) \quad (3.9)$$

A floor value of zero for the signal power is implemented to prevent over subtraction.

Now, in order to reconstruct the estimated output signal in time domain, the phase information from the operating microphone is required. However it is found that this does not suffer as significantly compared to the magnitude in the presence of noise [84]. Therefore, the estimated output signal can be reconstructed in time domain using the phase of the noisy observed signal with an inverse FFT(IFFT) by:

$$s(t) = IFFT(P_S(t, k) * \varphi[X(t, k)]) \quad (3.10)$$

where $\varphi[X(t, k)]$ is the phase of the observed signal. The steps to reconstruct the output signal are summarised in Figure. 3.2.

There are several key differences and similarities between the proposed RNC with conventional adaptive filtering methods like FDAF [79]. One of the similarities is that both uses a reference microphone to minimise an error function between the reference and operating microphone in the spectral domain. Also, both methods requires an initialisation period with an absence of sound event to operate, which means FDAF does not have an advantage over RNC despite being an iterative method. Lastly, both FDAF and RNC produces residual or musical noise caused by the spikes in random frequency bins.

The key differences between the two methods are that, firstly, FDAF iteratively minimises the error function whereas RNC uses a calibration period to find the global minimum. This means that FDAF has a disadvantage that it might not be convergent, and even if it does converge, the error will be at best equal to the global solution. This does however mean that the SER operating microphone in FDAF can be physically moved around after calibration and the FDAF will slowly converge unlike RNC. This limitation does not seriously affect the scenario of performing SER in a home environment, where the SER operating microphone can be assumed to be at a fixed location in the room. Another main difference is that FDAF uses a linear filter is used to find the relation between the two microphones, whereas an empirical mapping function is used in RNC. In real life scenarios where room attenuation exists, a linear filter is not sufficient to model the relation, and empirical mapping has an advantage.

3.3 Spectral GMM Dectector

The aim of the spectral GMM detector is to detect the sound event segments from the output of RNC. The RNC introduced in the previous section significantly reduces the interference noise found in the home environment. However such noise reduction algorithms can never completely remove all the interference, hence there may still exist spikes in random frequency bins. These spikes can sometimes combine to produce a high enough energy such that the energy feature would not be a reliable feature for detection. Instead, spectral-domain features are a better choice since the residual noise is sparser within the spectral domain.

One commonly used spectral based feature, the log power spectrum $S(k)$ is extracted by taking the discrete Fourier transform (DFT) of a windowed frame $s[n]$ from

the output signal followed by compressing the dynamic range using the log function as follows:

$$S(k) = \log_{10} \left| \sum_{n=0}^{N-1} s[n] w[n] e^{-i2\pi \frac{kn}{N}} \right|, \quad k = 0, 1, \dots, N-1 \quad (3.11)$$

where N is the number of samples per frame and $w[.]$ is the Hamming window.

An initial segment of the RNC output signal without any sound event is extracted and used to train a Gaussian Mixture Model(GMM). The output from the RNC pre-processing should be free from the TV interference, however residual noise from the TV often exists. Thus the model should be trained with the TV playing to capture the noise statistics of the residual noise together with any ambient noise from the room. This model is a probabilistic model which represents the likelihood of any feature compared against the trained model. The distance between an observed feature S and the model $f(.)$ is then as follows:

$$f(S) = \sum_{m=1}^M P(m) \mathcal{N}(S; \mu_m, \Sigma_m), \quad (3.12)$$

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:

$$\mathcal{N}(S; \mu, \Sigma) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (S - \mu)' \Sigma^{-1} (S - \mu) \right). \quad (3.13)$$

A threshold for detection is automatically sets based on the 95% confidence level of the likelihood of an initial segment without any sound event. A segment is then detected if the likelihood of the test segment exceeds the threshold.

3.4 Improved Subband Power Distribution Image Feature Classification Framework

The audio segments which are detected using the approach in the previous section are then further processed for feature extraction and classification under an improved subband power distribution image feature (iSPD-IF) classification framework. The iSPD-IF is an extension of the subband power distribution image feature (SPD-IF) classification framework introduced in [18]. In this section, the SPD-IF is recapped and the improvements to the existing method will be presented. A summary of the SPD-IF can seen in Figure 3.3.

Figure 3.3: A block diagram of SPD-IF/iSPD-IF classification framework consisting of three main modules: SPD image transformation, SIF image feature extraction and missing feature system. The proposed iSPD-IF improves on the SPD image transformation module.

The first step of the SPD-IF framework [18] is to transform an audio segment into its subband power distribution (SPD) image representation. This is done by transforming the power spectrogram of the segment into an image representation characterized by the histograms taken from the power distributions in each subbands. This SPD image is a two-dimensional sound event image that captures the distribution of the log-spectral power over time in each subband. The advantage of the SPD image is that the noise and signal are better localized and separable in each subband. This property is the key to separate the residual noise produced by the output RNC signal and the target sound event signal. The SPD image transformation will be recapped in more details in Section 3.4.1.

The proposed iSPD-IF builds on the SPD-IF framework, where the key difference lies in the method for SPD transformation. In the iSPD-IF framework, an alternative method of estimating the power distributions within the SPD image transformation is proposed. Experiments will be carried out using the structure of Figure 3.3 to compare iSPD-IF and SPD-IF. The alternative method involves using a non-parametric windows (NP-Windows) estimation [85] instead of the conventional histogram. The NP-Windows is able to give a better estimation of the power distributions by utilising the temporal information across the sub-bands.

Once the SPD image is formed, an image feature is extracted for classification using a method introduced in [86] for the spectrogram image feature(SIF). The image feature is based on the visual signature of the image which, first maps an image into a higher dimensional space by quantising the dynamic range into different regions, and then extracts an image feature inspired by the colour layout feature from image processing. Classification is performed using a missing feature system, which is used to take advantage of the fact that it is a simple task to separate the reliable and unreliable regions of the SPD. The missing feature system applies a noise mask on the SPD-IF, leaving only the reliable, high-power regions of the sound event and subsequently uses a k -Nearest Neighbours classifier [81] to get the final result. The missing feature system will be discussed in more detail in Section 3.4.3.

3.4.1 Improved Subband Power Distribution Image Transformation

In this subsection, an improvement to the SPD image transformation originally introduced in [18] is discussed. The basic idea of the SPD is to transform the spectrogram image into a new image representation which captures the distribution of spectral power in each frequency subband over time. The advantage of this representation is that the signal and noise are better localized and separable. Also, this new representation is not sensitive to onset and offset detection as compared to the conventional spectrogram image.

To extract the SPD image, an input signal is first transformed into its time-frequency spectrogram representation. The gammatone filterbank transformation [87] is chosen for the time-frequency spectrogram representation, as there is no tradeoff between time and frequency resolution, which is a common drawback of the conventional short-time Fourier transform (STFT) representation. The gammatone spectrogram, $S_g(k, n)$, is subsequently normalised into a grey-scale image in the range $[0, 1]$ as follows:

$$G(k, n) = \frac{\log S_g(k, n) - \min(\log S_g(k, n))}{\max(\log S_g(k, n))} \quad (3.14)$$

where k represents the centre frequencies of the filters, n is the time index, and log-power is used to compress the dynamic range of the spectrogram to enhance the high-power elements.

Next, for each subband, k , the probability density function (PDF) of the power across time is estimated. As the upper and lower bounds of the distribution are fixed by the normalisation of the spectral power to a grey-scale intensity, the raw distribution can be simply estimated using a non-parametric approach based on the histogram:

$$H(k, b) = \frac{1}{N} \sum_{n=1}^N \mathbf{1}_b(G(k, n)) \quad (3.15)$$

where N is the number of time samples in the segment and $\mathbf{1}_b$ is the indicator function, which equals one for the b^{th} bin if $G(k, n)$ lies within the range of the bin and is zero otherwise. The resulting $H(k, b)$ matrix therefore represents the raw probability distribution information for each frequency subband over time, which is constrained

to lie in the range $0 \leq H(k, b) \leq 1$. Figure 3.4 shows examples to illustrate the SPD representation of various sounds, bell sound, noise from air-con and bell sound masked in air-con noise. The top row represents the spectrogram surface and its probability densities of the power across time in each subband are transformed to produce the SPD representation in the bottom row.

Figure 3.4: Examples to illustrate the SPD representation of various sounds. Top row represents the spectrogram surface and bottom row represents the SPD representation

However, the histogram estimation assumes that the observations to be discrete independent and identically distributed (i.i.d) samples. For sound events, due to the modulation effects, the signal powers in each subband are often strongly correlated across time. This means there is useful temporal information for the PDF estimation that can be utilised. Another main disadvantage of the histogram is that the requirement to predefine the number of bins, such that the arbitrary bin boundaries and the block like nature of the resulting PDF estimate are not suitable in many applications. Parzen windowing [88] avoids arbitrary bin assignments and leads to smoother PDFs, however, a suitable kernel shape and size must be chosen.

An effective way to overcome these problems is to use the non-parametric windows (NP-Windows) estimation [85]. The motivation of NP-Windows estimation comes

from estimating the PDF of discrete points in an image. By assuming the continuity of the images and apply up-sampling, a more accurate PDF can be obtained with the additional interpolated points. Many interpolation schemes used for up-sampling proceed by fitting piecewise functions to the signal samples and re-sampling these at the required points. However, since the aim is to estimate the PDF, the re-sampling step can be avoided and the PDF of each piecewise section can be calculated directly in closed-form. Such an approach is therefore more accurate and more efficient to implement than the up-sampling method. In fact, its accuracy is dependant only on the accuracy of the piecewise representation, not the number of samples nor the number of bins. Calculating the PDF using NP-Windows [85] consists of three main steps:

1. Calculate the polynomial coefficients for the signal samples
2. Calculate the PDF for each piecewise section, such that the signal is considered a function of a uniform random variable representing its domain
3. Populate the appropriate bins for each piecewise section.

For our application, the powers of the adjacent bins will be treated as piecewise continuous as previously explained. The choice of polynomial can be of any degree, however from our preliminary findings, the linear polynomial provides equal or better performance compared to those with higher degrees. Therefore we estimate the distribution in each subband by first connecting each adjacent pair of data input with a straight line in the form $l_{k,i}(x) = a_{k,i}x + b_{k,i}$ where i represents the piecewise index and k represents the subband. For each piecewise straight line, a PDF, g , is assigned scaled to its gradient:

$$g_{k,i}(z) = \begin{cases} \frac{1}{|a_{k,i}|} & b_{k,i} \leq z \leq a_{k,i} + b_{k,i} \\ 0 & otherwise \end{cases} \quad (3.16)$$

Then, an arbitrary number of bins can be chosen for the output histogram by summing up all the PDF that lies within the interval. Equation 3.15 therefore becomes:

$$H(k, b) = \frac{1}{N-1} \sum_{i=0}^{N-1} \left(\int_{l_b}^{r_b} g_{k,i}(z) dz \right) \quad (3.17)$$

where l_b and r_b are the left and right edges of a particular histogram bin, b and N is the number of frames in the segment.

With this formulation of estimating the PDF, the proposed ‘improved’ SPD replaces the original SPD for the iSPD-IF framework. Figure 3.5 illustrates an example of different PDF estimation of a clean input A and its reverberated version, input B. It can be seen that the histogram of signal A is mismatched with input B with several spikes. However by using NP-windows, the output distribution of both input A and B is smoothed and is less mismatch between the clean and reverberated input.

Figure 3.5: Example of PDF estimation using histogram and NP-Windows. Input B (bottom row) which is a reverberated version of signal A (top row). The NP-Windows method produces less mismatch between the clean and reverberated input as compared to histogram.

3.4.2 Image Feature Extraction

After the iSPD image is formed, an image feature based on the visual signature is extracted using spectrogram image feature (SIF) [86]. The SIF method was initially introduced specifically to apply on the spectrogram image. However in [18], the SPD image was used in place of the spectrogram image. The motivation for using the SPD image instead was because the SPD image has a more robust representation where the signal and noise are better localised and separable. Experiments showed that the SPD image consistently outperformed the spectrogram image in noisy conditions. Thus in this thesis, the SIF method is used to extract an image feature from the SPD image and the proposed iSPD.

It was found that most of the information is contained within a small region of the dynamic range and it is desirable to enhance the information of the raw SPD within this small region that better represents the important signal information for classification. A method used in image processing referred to as “contrast stretching” [89], can be performed here as follows:

$$I(k, b) = \begin{cases} H(k, b) \times h, & \text{if } H(k, b) < \frac{1}{h} \\ 1, & \text{otherwise} \end{cases} \quad (3.18)$$

where I is the enhanced image, h is an appropriate constant, and x, y represent the image indices for image feature extraction. To extract the final feature, the enhanced image is partitioned into 10x10 local sub-blocks, and then compute the image pixel distribution statistics. The image pixel distribution is inspired by the color layout which is described further in [90]. The final image feature is a 600 (2x3x10x10)-dimensional vector, using red, green and blue (RGB) quantisation regions with the mean and variance to capture the distribution statistics as illustrated in Figure 3.6

Figure 3.6: Partitioning of the SPD image, and extraction of distribution statistics to generate the image feature where $L_{ij,color}$ is the RGB color distribution of the i^{th} row, j^{th} column sub-block.

3.4.3 Missing Feature System

Since the horizontal axis of the iSPD is the normalised spectral power, it is clear that the reliable signal area of the iSPD image is separated from the noise by a one-dimensional line. This is because the noise, unlike the sound events, is usually spread throughout entire frequency spectrum, hence leaving sparse peaks belonging to the sound event. This allows the noise and signal to be separable even at very low SNRs like $0dB$. Here it can be seen that by choosing only a region containing the reliable sub-blocks (indicated by yellow dotted area in Figure 3.7a), it matches the same region of a bell sound in high SNR condition from training as illustrated in Figure 3.7b. This provides motivation to find the boundaries of the noise mask, in order to perform classification using only the reliable sub-blocks.

(a) The SPD is divided into two regions, noise and signal, by a white line.

(b) High SNR SPD template from training.

Figure 3.7: SPD of bell sound masked in noise at $0dB$ SNR(left)/high SNR conditions(right) where the yellow dotted area from (a) and (b) indicating the reliable regions for classification

To find the boundaries of the noise mask, a non-stationary noise estimation approach based on the SPD representation is proposed in [18]. This idea is motivated by the observation that despite changes in the non-stationary noise intensity, the characteristics of the noise distribution remain the same over time. In the SPD representation, this change in intensity can be approximated as a shift in normalised spectral magnitude of the noise distribution. Hence, if we extract an SPD from a segment containing only noise, $I_N(k, z)$, we can assume that the noise in the SPD containing both noise and signal is represented as $I_N(k, z + a)$. The problem therefore simplifies to estimating a , the change in the non-stationary noise intensity.

To estimate a , we take the cross-correlation ($\star$) between $I_N(k, z)$ and $I(k, z)$ to find the intensity difference, a_{max} with the highest correlation. Since $I(k, z)$ is a mixture of the noise and signal distributions, we perform the cross-correlation separately on each SPD-IF subband, k , such that the highest correlation should occur between two noise-dominated subbands:

$$a_{max} = \max_a [I(k, z) \star I_N(k, z + a)] \quad , \quad \forall k \quad (3.19)$$

The final SPD-IF noise estimate, $n(k)$ is then simply:

$$n(k) = n_{max}(k) + a_{max} \quad (3.20)$$

where $n_{max}(k)$ is the maximum occupied bin for each frequency subband given by:

$$n_{max}(k) = \arg \max_z (I_N(k, z) > 0) \quad (3.21)$$

After the noise estimate is found, the SPD is divided into the reliable region, $I_r(k, z)$, and the unreliable region, $I_u(k, z)$, as:

$$H(k, z) \rightarrow \begin{cases} I_r(k, z), & \text{if } z > n(f). \\ I_u(k, z), & \text{otherwise.} \end{cases} \quad (3.22)$$

In Figure 3.7a, $n(f)$ represents the white solid line, $I_r(k, z)$ represents the reliable region inside the yellow dotted area, $I_u(k, z)$ represents the unreliable region outside the yellow dotted area. The missing features are compensated by masking the unreliable SPD-IF regions and using only the reliable regions for classification with k -Nearest Neighbours [81]. Here, the Hellinger distance [91] is used to measure the distribution distance between image features, as follows:

$$d_H(x, x_T) = \sum_{k=1}^{N_r} \left(1 - \sqrt{\frac{2\sigma_k\sigma_{T,k}}{\sigma_k^2 + \sigma_{T,k}^2}} e^{-\frac{1}{4} \frac{(\mu_k - \mu_{T,k})^2}{\sigma_k^2 + \sigma_{T,k}^2}} \right)^{\frac{1}{2}} \quad (3.23)$$

where N_r is the number of reliable dimensions, x is a test vector and x_T is a sample from the training data, μ and σ^2 are the distribution means and variance of the image pixel distribution. Finally, the sample from the training database which obtains the minimal distance against a test segment is assigned as the predicted label. Figure 3.8 summarises the steps of missing feature system which are applied on the extracted image feature.

Figure 3.8: A block diagram describing the steps of the missing feature system which are applied on the extracted image feature.

3.5 Experiments

In this section, experiments are being carried out to demonstrate the performance of our proposed system on a sound recognition task. The main aim is to evaluate the effectiveness of the two proposed methods, RNC and iSPD-IF classification framework against the conventional methods. The method used for evaluation is the accuracy of the classification on sound events that occurs in home environment, in the presence of an interference noise.

3.5.1 Setup

Sound Database For the sound events, the following ten sound classes related to the home environment setting from the Real Word Computing Partnership (RWCP) Sound Scene Database [92] are selected: horn, bells5, bottle1, phone4, whistle1, whistle3, clock2, ring, doorlock and trashbox. All the sound files have a high signal-to-noise ratio (SNR), and each contains an isolated sound, with some silence before and after the sound. Each class has 100 samples and the ten classes give a total of 1000 samples.

For the interference noise, a half hour news segment from a local TV channel is utilized. It includes many segments of speech, music and random sounds, which serves as a good test for the effectiveness of the system. The news segment was recorded using a direct audio-in cable from the TV, and it will be used for playback in our repeated experiments of different signal-to-interference ratio (SIR).

Recordings To simulate sound events that occur in home environment in the presence of an interference noise, both the clips from the sound event database and the

Figure 3.9: Bird's eye view of the recording room indicating positions of the TV, reference/operating microphones and the possible locations of the sound events

news segment were played back from two speakers. The noise levels of the news segment was varied during recording to evaluate the performance at different SIR of 5, 0 and -5 dB. Both the sound classes and news segment are played back and recorded in a medium-sized room ($10\text{m} \times 4\text{m} \times 3\text{m}$) and the positions of the equipment is seen in Figure 3.9. Speakers were used to simulate a TV in a fixed position, while the speaker used to playback the sound events was placed in ten various positions around the room. The operating microphone used for SER is situated in the centre of the room; the reference microphone is situated beside the TV. The sampling rate of the recordings is 16000hz in 16bits resolution, recorded with high quality omni directional microphones (Shure-MX184).

Experimental Methods The following preprocessing methods are evaluated:

1. Proposed RNC: Regressive Noise Cancellation using subband power regression mapping, the linear regression is chosen as described in Equation 3.5. A initial segment of 3s is used for the learning the mapping coefficients.
2. Baseline FDAF: a frequency domain adaptive filtering method introduced in [79]. In section 2.3.2, time domain adaptive filtering may fail to converge and preliminary experiments verified the claim. Thus, FDAF is chosen as the method of comparison.

and the following classification methods are evaluated:

1. SPD-IF, a robust sound classification framework introduced in [18]. The default parameters is the same as author's chosen parameters.
2. Proposed iSPD-IF, a extension of SPD-IF, using NP-Windows to improve the estimation of power densities.
3. MFCC-SVM: MFCC features modelled with one-against-one (OAO) SVM.
4. MFCC-GMM: MFCC features modelled with a GMM model
5. MFCC-GMM-MVA: based on MFCC-GMM with mean, variance and arma normalisation (MVA).
6. MFCC-GMM-MVA-Multi: based on MFCC-GMM-MVA with multi-conditional training using additive noise and also convoluting with random room impulse responses. The noise are added from random segments of the TV signal at various noise levels.

For all various preprocessing and classification methods, frame lengths of 16ms with frame shifts of 8ms were used throughout. All MFCC features include deltas and delta-deltas, without the 0th coefficient and log energy to reduce mismatch due to loudness, giving a total of 36 dimensions. All methods using GMM are generated with 10 mixtures. This number of mixture was chosen empirically across the entire database. For training, only the original clean samples from the CD are used in each classification method. For testing, each sound event is segmented using the ground truth time label for fair comparison, since the focus of the evaluation is solely on the classification and not the detection accuracy. Both training and testing are coded and evaluated using Matlab.

3.5.2 Results and discussions

The classification accuracy is investigated in two parts, firstly in the absence of TV playing, to serve as a baseline performance for the classification methods. Secondly, sound events with the news segment played back in the background are used to evaluate the effectiveness of the preprocessing methods on the various classification methods.

Without interference

An experiment is conducted to evaluate the performance in the absence of TV playing as it is an important step to find an upper bound performance for each of the classification method when compared to playing in the presence of TV. The experimental results presented in Table 3.1 indicate the baseline performance for various classification methods.

	MFCC-SVM	MFCC-GMM	MFCC-GMM-MVA	MFCC-GMM-MVA-Multi	SPD-IF	iSPD-IF
Accuracy	44.60	48.20	67.40	95.12	97.84	98.52

Table 3.1: Classification results without interference (%).

Even though no interference is present, the sound events are generated from various positions in the room and hence the recorded test samples will be mismatched with the clean training data due to the channel response. This is reflected in the result for MFCC-SVM and MFCC-GMM which performs badly at 44.60% and 48.20% respectively.

Using MFCC-GMM-MVA, i.e. MFCC-GMM with mean, variance and ARMA normalisation (MVA), this system performed better at 67.40%, as it is compensated slightly for the mismatch. However, the performance without interference was expected to perform at least 90% for a good lower bound for the experiment with interference. A multi-conditional training is added to MFCC-GMM-MVA (MFCC-GMM-MVA-Multi) where the training uses additive noise and convoluting with random room impulse responses performs much better at 95.12%. The multi-conditional training is effective in reducing the mismatch caused by the room’s impulse response.

The proposed classifier iSPD-IF and the original SPD-IF both achieve a good baseline performance in the mismatched condition, with accuracies more than MFCC-GMM-MVA-Multi. Also, even though the SPD-IF was not tested on convolution noise in our previous work, from this result we find that both the SPD-IF (97.84%) and iSPD-IF (98.52%) perform well with the recordings from the distant microphone. With these findings, only the top three feature-classifier combination with scores over 95% are used to evaluate experiments in the presence of interference noise.

With interference

	MFCC-GMM-MVA-Multi			SPD-IF			iSPD-IF		
SIR	-5db	0db	5db	-5db	0db	5db	-5db	0db	5db
No Processing	41.78	43.78	47.24	51.18	62.94	69.96	51.32	63.74	70.88
FDAF	49.80	50.64	50.88	87.64	90.06	91.20	90.54	91.50	92.54
RNC	61.56	65.44	66.00	89.38	92.80	94.98	91.46	93.96	96.10

Table 3.2: Classification results with interference at different SIR levels (%).

This section discusses the performance when the sound classes are recorded with interference under different Signal-to-Interference (SIR) levels as shown in Table 3.2. It can be seen that the accuracy of each method reduces significantly when no pre-processing has been done before classification in the first row. However, we observe that both SPD-IF and iSPD-IF outperforms the conventional MFCC-GMM-MVA-Multi method at all SIR levels. This is because SPD-IF and iSPD-IF are able to mask out some of the non-stationary noise from the interference. However, it cannot cope completely with the highly non-stationary TV noise.

The last two rows of Table 3.2 show the classification results after pre-processing the signals with the FDAF and RNC methods. It can be clearly seen that using additional reference channel and applying pre-processing in the time domain prior to feature extraction improves the results in all cases. The most significant result is that our proposed RNC outperforms FDAF for all three classification methods at all SIR levels. Another important observation is that the improvements for MFCC-GMM-Multi method are much less than for both SPD-IF methods. This is because the residual noise left from the preprocessing are effectively masked off in the both SPD-IF methods. However for MFCC-GMM-MVA-Multi method, the MFCC features are sensitive to the slight changes caused by the residual noise. This results in mismatch, thus achieved less relative improvement.

Finally comparing our proposed iSPD-IF and the original SPD-IF in both Tables 3.1 and 3.2, we can see that our proposed iSPD-IF consistently outperforms our original SPD-IF, with between 1-4% improvement. This is due to the use of NP-Windows to estimate the SPD, which reduces mismatch under reverberant conditions.

Overall, our proposed RNC preprocessing combined with the iSPD-IF gives a very good accuracy of above 90% for all SIR cases. In particular at 5db SIR, the average

accuracy was 96.1% which is less than 2.5% difference to the clean baseline. In addition, a significant advantage of the SPD-IF methods, as compared to the conventional multi-conditional training, is the simplicity of using only the original clean signals for training which removes the possibility of room mismatch occurring, while maintaining a superior performance in the classification accuracy.

3.6 Summary

This chapter introduced a framework to perform SER in home environments in the presence of interferences such as TV, radio or music playing. The framework includes two novel methods: (1) the regression-based noise cancellation (RNC), a preprocessing to greatly reduce the non-stationary interference (2) improved subband power distribution image feature (iSPD-IF) which is an extension of the robust SPD-IF, a classification framework designed to transform the signal to a novel representation where the signal is separable from the residual noise.

Experiments have shown that RNC works effectively for preprocessing when paired with a range of classifiers. At the same time, experiments shown that both the SPD-IF and iSPD-IF outperform all the other feature/classifier combinations. The proposed iSPD-IF shows an improvement over the original SPD-IF by utilizing a better PDF estimation of the SPD feature, achieving more than 96% classification accuracy of ten sound classes under 5dB TV interference.

Chapter 4

Class-based Compensation for SVM with Limited Training Data

In the previous chapter, the problem of performing sound event recognition (SER) in interference noise was addressed, in particular with the presence of TV playing. In this chapter, we turn our attention to the problem of limited training data. The iSPD-IF which was proposed in the previous chapter although is a robust classification framework, however, it does not perform as well here as the k -Nearest Neighbours classifier usually requires a large dataset for training. The first section gives an overview of the existing solutions to limited training data. A novel compensation technique is then introduced in Section 4.2, where the focus is on making the features more discriminative to improve the classification performance. The proposed feature is integrated with multi-class SVM and experiments are carried out and compared with various methods to validate our approach.

4.1 Overview

In this chapter, the problem of performing SER with limited training data is examined. One common way to overcome the problem of limited training data is to use Semi-Supervised Learning (SSL) [17]. This utilizes an initial model trained from the small training data set, which is subsequently used to classify the unlabelled data. During classification of the unlabelled data, a confidence measure is used to identify the reliable

samples that can be used to improve the initial model. By using the reliable samples to further train the initial model, it is possible to improve the representation of the model. The Support Vector Machine (SVM) classifier is the most commonly used in SSL approaches as (1) it requires only a small set of training data to estimate the initial model, (2) the distance from the test vector to the SVM hyperplanes is a good confidence measure to identify the reliable samples [17]. SSL can be regarded as an improvement to address the problem of limited training data, in addition to any feature-classifier combination. While most research that deals with limited training data focuses on improving the performance of SSL classification methods, there is less attention to directly improve the feature for the case of limited training data. In this chapter, we propose to improve the features by making them more discriminative to improve the overall performance.

One common way to make features more discriminative is by using feature selection [93, 94]. The main idea of feature selection is to select a subset of the feature vector that leads to the smallest classification error. Various well-known feature selection methods have been proposed and can be found in [95]. However, one major drawback of feature selection is that it is not as effective for limited training samples [95, 96] and may even worsen the overall classification accuracy.

In this chapter, a class-based compensation for SVM is proposed to improve the feature to make the classification more discriminative. The idea of the compensation is to use a different feature for each of the binary SVM classes that best discriminates each class pair. To do this, a discriminative filter is constructed for each class pairs of the SVM. The filters' purpose is to enhance the discriminative capability between the binary classes by giving higher weights to the frequency components which best discriminate the class information. With this set of filters, the important frequency components will be emphasised, which will improve the classification accuracy. Even with limited training data, the different clips from the same sound class should contain similar dominating frequency components, such that a small training set is sufficient to capture these frequency components.

One advantage of the proposed feature compensation method is that it can be easily used in conjunction with existing noise robustness techniques for SER introduced in Chapter 2.2. This means that the noise which usually exists in the home environment is simultaneously addressed. In addition, SSL methods could also be used in conjunction

with the proposed feature compensation method to achieve a better overall result. Therefore, the proposed method does not replace either the noise robustness techniques or SSL methods. Instead, it is regarded as a method that can be used to further enhance the existing techniques when in combination.

4.2 Class-based Compensation for SVM

In this section, a novel feature compensation method for SVM is proposed to make the classes more discriminative for classification. Previously in Chapter 2.1.3, two popular methods were introduced for multi-class SVMs: the one-against-one (OAO) [51] and one-against-all (OAA) method [52]. A comparative study of various types of multi-class SVM in [97] shows that OAO-SVM generally gives better or equal performance as compared to the other methods. Besides OAO-SVM's advantage of having good performance, it is also relatively less complex and fast to train, therefore being the most popular choice among the multi-class SVMs, it will be the choice of SVM in this thesis.

In the OAO method, $N(N - 1)/2$ binary SVM classifiers are constructed, using all the pair-wise combinations of the N classes. The idea of the feature compensation method is to adapt the features for each of the $N(N - 1)/2$ binary SVM classifiers as seen in Figure 4.1, by using specific features designed for each of the class pairs. An effective way to adapt the features is by constructing a discriminative filter for each of the binary SVM classifiers. The discriminative filter can be applied to any spectral based feature by giving higher weights to the frequency components which best discriminate the classes and vice versa. This enhances the discriminative capability between any binary classes. Next, subsection 4.2.1 details the construction of the filter for each of the binary classes and subsection 4.2.2 introduces a framework to integrate the filters into a multi-class SVM.

Figure 4.1: $N(N - 1)/2$ discriminative filters are constructed for each of the $N(N - 1)/2$ pairs of the N class SVM.

4.2.1 Class-based Compensation for binary SVM

Exploiting the fact that between two classes, the characteristic, high-power frequency components, which best discriminates the classes, usually occurs in different bins for each class. The aim is to give higher weights to such frequency bins and lower weights to bins that are less discriminative.

Using clean training data from two classes i, j at each binary classification node, the spectral powers are first computed by taking the Short-Time Discrete Fourier Transform (ST-DFT) for each sound clip:

$$X_{class}(k, t) = \left| \sum_{n=0}^{N-1} x_{class,t}[n] w[n] e^{-i2\pi \frac{kn}{N}} \right|, \quad k = 0, 1, \dots, N-1 \quad (4.1)$$

where $class \in \{i, j\}$ is the class index, k is the frequency bin, t is the frame index, N is the number of samples per frame and $w[\cdot]$ is the Hamming window. Next, the spectral powers across each subband, k , are appended to the vectors $M_{class}(k)$, where $class \in \{i, j\}$. The spectral powers are appended for every sound clip into their respective vectors $M_i(k)$ and $M_j(k)$:

$$M_{class}(k) = [X_{class}(k, 1), X_{class}(k, 2), \dots, X_{class}(k, t_{n_{class}})] \quad (4.2)$$

where n_i and n_j is the number of samples for class i and j respectively. To create a filter which gives higher weights to the frequency bins of higher importance and vice versa, a suitable measure of the discriminability of the data must be used. Here, the Fisher's linear discriminant (FLD) analysis [98] formulated by Fisher *et. al* is utilised. Fisher defined the separation between the distributions of the two classes to be the ratio of the variance between the classes to the variance within the classes:

$$S_{i,j}(k, \vec{w}) = \frac{\Sigma_{\text{between}}}{\Sigma_{\text{within}}} = \frac{(\vec{w} \cdot \vec{\mu}_i^k - \vec{w} \cdot \vec{\mu}_j^k)^2}{\vec{w}^T \Sigma_i^k \vec{w} + \vec{w}^T \Sigma_j^k \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_i^k - \vec{\mu}_j^k))^2}{\vec{w}^T (\Sigma_i^k + \Sigma_j^k) \vec{w}} \quad (4.3)$$

where μ_i^k and μ_j^k represents the means of $M_i(k)$ and $M_j(k)$ respectively and Σ_i^k and Σ_j^k represents the variances of $M_i(k)$ and $M_j(k)$ respectively. From Equation 4.3, the

separability index $r_{ij}(k)$ between two classes can be defined as follows:

$$r_{ij}(k) = \max_{\vec{w}} S_{i,j}(k, \vec{w}) \quad , \quad k = 0, 1, \dots, N - 1 \quad (4.4)$$

The separability index is a measure of how closely related or overlapped two distributions are. A lower value of r_{ij} implies that the distributions are more overlapped and vice versa. The range of values of the separability index is between 0 and ∞ , where 0 occurs when the two distributions are completely identical. Figure 4.2 illustrates the separability index between two different pairs of distributions. Firstly, in Figure 4.2a, the two distributions in red and blue respectively have a separability index of 0.0067. In Figure 4.2b, the separability index decrease to 0.0017 as the two distributions are more overlapped with each other.

Figure 4.2: An illustration of the separability index for different pairs of distributions. The separability index decreases as the two distributions get more overlapped and vice versa.

The separability index, r_{ij} , increases if the two distributions are less overlapped and vice versa, and this is a useful property for constructing the filter. Recall that the characteristic, high-power frequency components, which best discriminate the classes, usually occur in different bins for each class. This means that the distribution of their frequency components do not overlap significantly. Therefore by assigning r_{ij} as the filter weight, the important frequency bins get assigned higher weights, which enhances the discriminative capability. The weights are normalised such that the overall energy

of the input is retained. The final weights are therefore as follows:

$$W_{ij}(k) = \frac{r_{ij}(k)}{\sum_{k=1}^N r_{ij}(k)} \quad , \quad k = 0, 1, \dots, N-1 \quad (4.5)$$

After the weights have been generated, they are subsequently used to adapt the features specifically for the binary class classification by multiplying the spectral powers by the weights:

$$\begin{cases} \hat{X}_i^{(ij)}(k, t) = X_i(k, t)W_{ij}(k) & , \quad k = 0, 1, \dots, N-1 \\ \hat{X}_j^{(ij)}(k, t) = X_j(k, t)W_{ij}(k) & , \quad k = 0, 1, \dots, N-1 \end{cases} \quad (4.6)$$

where t is any fixed frame.

Figure 4.3 illustrates the proposed filter generated between two sound classes, class 1 and class 2, and also the effects of the filters on the sound samples. Sample 1 is a sample clip from class 1 and its spectrogram is shown in Figure 4.3a. From sample 1's spectrogram, it is seen that most of the spectral energy is concentrated around frequency bin 75 as this corresponds to a harmonic at this given frequency. Sample 2, a sample clip from class 2 and its spectrogram is shown in Figure 4.3a and the spectral energy is more uniformly distributed across all frequency bins. The generated filter as shown in Figure 4.3b gives a much higher weight for the regions around frequency bin 75 to emphasis on the discriminative bins. The effects of the filter is illustrated in Figure 4.3d and 4.3e, where the spectral information of class 1 is greatly emphasised at frequency bin 75 to make the comparison more discriminative.

For sound events in general, if the different clips from the same sound class contain similar dominating frequency components, a small training set is sufficient to capture these frequency components. Algorithm 1 is a summary of the steps to generate the class-based compensated filter:

Algorithm 1 Designing a discriminative class-based compensated filter between two classes

Require: Clean training data from two classes $\{i; j\}$

- 1: Compute spectral powers, $X_{class}(k, t)$, of every sound clip.
 - 2: **for all** subbands k , such that $1 \leq k \leq N$ **do**
 - 3: Extract k -th subband $X_m(k, t)$ for every samples.
 - 4: Append $X_{class}(k, t)$ into respective vectors $M_i(k)$ or $M_j(k)$.
 - 5: Do FLD analysis, get separability index $r_{ij}(k)$.
 - 6: Assign filter weights $W_{ij}(k) = \frac{r_{ij}(k)}{\sum r_{ij}(k)}$ with normalisation
 - 7: **end for**
 - 8: **return** W_{ij}
-

(a) Sample 1: spectrogram of a sample clip from Class 1

(b) Sample 2: spectrogram of a sample clip from Class 2

(c) Adapter filter with weights W_{ij} generated for class 1 and 2.

(d) Sample 1 after applying the class-based compensated filter

(e) Sample 2 after applying the class-based compensated filter

4.2.2 Integrating Class-based Compensation into multi-class SVM

After constructing a filter designed to increase the discriminability between each binary combination, a framework is introduced for extending this into a multi-class OAO-SVM. To implement the class-based compensated filter into an N-class OAO-SVM, $N(N - 1)/2$ sets of filters will be learnt for each pair-wise combination and Algorithm 2 details the full steps required in this process. One advantage of the proposed feature is that since the weights are applied directly to the spectral powers, there is no restriction on the choice of features extracted. For comparative purposes in our later experiments, the popular MFCC feature is chosen as a common feature across each of the algorithm.

Algorithm 2 Class-based Compensated MFCC integration with multi-class OAO-SVM

Require: Clean training data from all the N classes

- 1: Training Process:
 - 2: **for all** $N(N - 1)/2$ pair-wise combinations **do**
 - 3: Construct class-based compensated filter, W_{ij} for classes i and j according to Algorithm 1
 - 4: Apply filter, W_{ij} on each spectrogram for train samples according to Equation 4.6
 - 5: Extract features using adapted spectrograms
 - 6: Train binary SVM for class i and j using new features
 - 7: **end for**
 - 8: Testing Process:
 - 9: **for all** $N(N - 1)/2$ pair-wise combinations **do**
 - 10: Using W_{ij} from training process, and apply on spectrogram of test sample according to Equation 4.6
 - 11: Extract features using adapted spectrogram
 - 12: Evaluate features on the binary SVM and the winner class gets a vote
 - 13: **end for**
 - 14: **return** Class with majority votes
-

4.3 Experiments

In this section, experiments are carried out to demonstrate the performance of the proposed class-based compensation method integrated with a SVM classifier on a sound recognition task. The main aim is to evaluate the effectiveness of the proposed method with limited number of training samples and noisy conditions that are commonly found in home environments. The method is compared against several other techniques and the accuracy of the classification on sound events is used as a measure for the evaluation.

4.3.1 Setup

Sound Database For the sound events, the same ten sound classes used in the previous chapter are selected from the RWCP Sound Scene Database [92]: horn, bells5, bottle1, phone4, whistle1, whistle3, clock2, ring, doorlock and trashbox. Each class has 100 samples and the ten classes give a total of 1000 samples. For each class, 10 samples are used for training and 90 samples used for testing. Note the very low count of 10 training samples is used to simulate the limited data problem. —

To evaluate the performance in a realistic home environment, the performance of the proposed method is evaluated in the presence of noise will also be evaluated. The noise clip was chosen as a home related background noise, which is the sound generated from a washing machine, taken from Audio Pro European SFX Library [99]. The noise was added at different signal-to-noise ratio: clean, $20dB$, $15dB$, $10dB$ and $5dB$.

Experimental Methods Firstly, three methods are chosen as the baseline without applying the proposed class-based compensation. Two of the three methods are trained with the one-against-one SVM (OAO-SVM) classifier. Among the two methods trained with the OAO-SVM, the first feature is the conventional MFCC and second feature is the MFCC with state-of-the art noise reduction ETSI Advanced Front-End (AFE) to deal with the background noise. The last baseline method is the improved subband power distribution image feature (iSPD-IF) technique from the previous chapter which uses the k -Nearest Neighbours clas-

sifier. The last method is included to compare the performance of the different classifiers with limited training samples. In summary, the following baseline features and classifiers combinations are used for evaluation:

1. MFCC-SVM: MFCC feature and OAO-SVM classifier
2. AFE-MFCC-SVM : MFCC feature with AFE noise reduction and OAO-SVM classifier
3. iSPD-IF-kNN: Improved subband power distribution image feature with k -Nearest Neighbours classifier

Next, the proposed class-based compensation applied on the baseline methods to validate its effectiveness. Among the three baseline methods, the class-based compensation is applied on only the two which uses the OAO-SVM classifier. In summary, the following features and classifiers combinations with class-based compensation are used for evaluation:

4. CBC-MFCC-SVM: Proposed class-based compensated MFCC feature and OAO-SVM classifier
5. CBC-AFE-MFCC-SVM: Proposed class-based compensated MFCC feature with AFE noise reduction and OAO-SVM classifier

All MFCCs used are set up with standard configuration with 36-dimensions including deltas. The discriminative filters for the class-based compensated MFCCs are trained using the 10 training samples from each class. The OAO-SVMs are setup with the linear kernel [100].

4.3.2 Results and discussions

Table 4.1 summarizes the empirical results of our experiments on the RWCP sound event database with limited training data of 10 training samples per class. We first evaluate the performance of the baseline methods. Firstly, comparing the iSPD-IF-kNN and MFCC-SVM, it is observed that MFCC-SVM outperforms the iSPD-IF-kNN by a significant margin in clean conditions. This is as expected as the k -Nearest Neighbours classifier usually requires a large number of training samples to be effective. However,

Method	Clean	20dB	15dB	10dB	5dB	Average
iSPD-IF-kNN	91.0	90.2	87.6	84.8	82.6	87.2
MFCC-SVM	94.8	89.6	85.8	80.8	77.8	85.8
CBC-MFCC-SVM	97.8	92.4	90.0	84.2	82.2	87.3
AFE-MFCC-SVM	94.8	93.4	91.0	85.0	83.0	87.4
CBC-AFE-MFCC-SVM	97.8	94.0	92.6	86.2	84.0	90.9

Table 4.1: *Recognition results (classification accuracy in percentage %).*

in the noisy conditions, iSPD-IF-kNN remains more robust against noise and outperforms the MFCC-SVM which has no noise robustness capability. Next, comparing AFE-MFCC-SVM against MFCC-SVM, it can be seen that the additional noise reduction improves the performance of the MFCC-SVM across all noise levels except in clean conditions where the performance remained the same. With the additional noise reduction, the AFE-MFCC-SVM now outperforms the iSPD-IF-kNN in noisy conditions as the iSPD-IF-kNN suffers from a poor clean baseline.

Now changing the attention to the performance of the proposed class-based compensation by first comparing CBC-MFCC-SVM against MFCC-SVM. It is observed that the class-based compensation improved the accuracy by 3 – 4% across all SNRs in noisy condition as well in clean conditions. This affirms the effectiveness of the discriminative filters applied onto the MFCC features. It is also noted that the CBC-MFCC-SVM performed close to AFE-MFCC-SVM in noisy conditions even though the CBC-MFCC-SVM has no noise robustness applied. Lastly, observing the performance of the proposed class-based compensation applied on AFE-MFCC-SVM, the compensation also provided a boost in performance to the already well performing AFE-MFCC-SVM. The improvements ranged from 1 – 3% across the different noise conditions. This affirms that the class-based compensation is able to provide an additional improvement when used with noise robust features. Overall, the CBC-AFE-MFCC-SVM achieved the best results for our recognition task with limited training data across all conditions.

4.4 Summary

In this chapter, a novel class-based compensation is proposed, which learns a set of filters for each of the binary classes of the OAO-SVM classifier to enhance the discriminative capability of the system with limited training data. The use of FLD analysis on the power spectrum distribution between the class pairs, assigns higher weights to the frequency components which best discriminates the classes and vice versa. The experimental results have shown the compensation is able to perform well with the constraint of limited training samples. Moreover, the compensation method further improves the accuracy when used with in conjunction with noise robustness techniques in noisy environments. In summary, the class-based compensation is an effective way to enhance the discriminability with limited training data when applied independently or with noise robustness techniques.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis addressed two commonly faced problems that occur when performing SER in the home environments: (1) the presence of interference noise and (2) limited training data. For the problem of interference noise, a novel preprocessing method, regressive noise cancellation (RNC) which utilises an additional reference microphone to enable effective cancellation of the television noise is proposed. Unlike conventional adaptive filtering, which iteratively minimises the error function to find the unknown room impulse response, RNC is based on a regression mapping in the frequency domain to find a global minimum for the error function. The proposed RNC method has shown an improvement over conventional adaptive methods when using the classification accuracy of baseline SER systems as a performance measure. Similar to other filtering methods, some residual noise remains after RNC processing. To improve performance, the subband power distribution image feature (SPD-IF) classification framework which can localize the noise and signal into separate regions is applied to extract the reliable parts for feature generation. The author also extended the SPD-IF method by improving the feature extraction step under the improved subband power distribution image feature (iSPD-IF) framework utilizing the temporal information across sub-bands. From the experimental results, both proposed iSPD-IF and original SPD-IF outperformed methods such as SVM and GMM using conventional MFCC features. In addition, the proposed iSPD-IF gave a consistent improvement over the

original SPD-IF, enabling the proposed system to achieve more than 96% classification accuracy in a challenging signal-to-interference ratio of 5dB TV noise conditions.

In Chapter 4, a class-based compensation (CBC) method for the problem of limited training data is proposed. The idea of CBC is to learn a set of filters for each of the binary classes of the OAO-SVM classifier to enhance the discriminative capability of the features for classification. To enable this, CBC employs FLD analysis on the power spectrum distribution between the class pairs to assign higher weights to frequency components which best discriminate the class. Experimental results show that the compensation is able to perform well with the constraint of limited training samples. Moreover, the compensation method further improves the accuracy when used with in conjunction with noise robustness techniques in noisy environments.

In summary, the author has contributed the following novel works in this thesis:

- Regressive noise cancellation (RNC): a preprocessing method which utilises an additional reference microphone to effectively cancel the TV noise. The RNC uses an empirical mapping learnt between the spectral powers of reference microphone and the operating SER microphone for cancellation. This is published in [1].
- The improved subband power distribution image feature (iSPD-IF): an extension of an existing classification framework originally proposed in [18] to enhance the classification of audio signals where residual noise remains after cancellation. This is published in [2].
- Class-based Compensation for SVM, a filter which can be used on spectral-based features to increase the discriminability between two classes. The filters enhance the discriminative capability by modifying the spectral weighting between two classes of a binary SVM. This is published in [3].

5.2 Future Work

Future work of this thesis can be carried out in several directions:

- For RNC method introduced in Chapter 3.2, several other empirical mapping functions can be explored for more effective cancellation: e.g. Empirical MMSE,

non-linear regression, etc. Also, the cost function in Equation 3.4 using L_2 Euclidean norm can be extended using other measures for example L_1 norm. Finally, it may be possible to design a post processing for RNC that can reduce the effects of residual noise.

- For the iSPD-IF introduced in Chapter 3.4, the experimental results demonstrated that the method is effective in a reverberant conditions. However, the evaluation for reverberant conditions was not fully comprehensive as it was not the main emphasis in this work. In the future work, more comprehensive experiments should be carried out by varying the reverberation decay parameter, T_{60} .
- In the experimental setup to validate the proposed class-based compensation in Chapter 4.3, the MFCC is chosen as the baseline spectral-based feature. The proposed class-based compensation has shown to perform well with MFCC. In the future work, we plan to test the class-based compensation with other spectral-based features for a more comprehensive evaluation.

5.3 Publications

The following papers have been published during the course of this research:

- **Ng Wen Zheng Terence**, Tran Huy Dat, Jonathan Dennis, Chng Eng Siong, “Robust Sound Event Recognition Under TV Playing Conditions” in IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) 2013
- **Ng Wen Zheng Terence**, Tran Huy Dat, Jonathan Dennis, Chng Eng Siong, “A Robust Sound Event Recognition Framework Under TV Playing Conditions” in Asia Pacific Signal and Information Processing Association (APSIPA) 2013
- **Ng Wen Zheng Terence**, Tran Huy Dat, Huynh Thai Hoa, Chng Eng Siong, “Adaptive Semi-supervised Tree SVM for Sound Event Recognition in Home Environments” in Asia Pacific Signal and Information Processing Association (APSIPA) 2013

References

- [1] W. Z. T. Ng, H. D. Tran, J. Dennis, and E. S. Chng, “Robust sound event recognition under tv playing conditions,” *IEEE China Summit and International Conference on Signal and Information Processing*, 2013.
- [2] W. Z. T. Ng, H. D. Tran, J. Dennis, and E. S. Chng, “A robust sound event recognition framework under tv playing conditions,” *Asia Pacific Signal and Information Processing Association*, 2013.
- [3] W. Z. T. Ng, H. D. Tran, T. H. Huynh, and E. S. Chng, “Adaptive semi-supervised tree svm for sound event recognition in home environments,” *Asia Pacific Signal and Information Processing Association*, 2013.
- [4] G. Virone, D. Istrate, M. Vacher, N. Noury, J. Serignat, and J. Demongeot, “First steps in data fusion between a multichannel audio acquisition and an information system for home healthcare,” in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 2, pp. 1364–1367, IEEE, 2003.
- [5] P. van Hengel and J. Anemüller, “Audio event detection for in-home care,” in *Int. Conf. on Acoustics (NAG/DAGA)*, 2009.
- [6] H. Lozano, I. Hernáez, A. Picón, J. Camarena, and E. Navas, “Audio classification techniques in home environments for elderly/dependant people,” in *Computers Helping People with Special Needs*, pp. 320–323, Springer, 2010.
- [7] M. Vacher, D. Istrate, F. Portet, T. Joubert, T. Chevalier, S. Smidtas, B. Meillon, B. Lecouteux, M. Sehili, P. Chahuara, *et al.*, “The sweet-home project: Audio technology in smart homes to improve well-being and reliance,” in *Engineering*

- in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 5291–5294, IEEE, 2011.
- [8] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, “Scream detection for home applications,” in *Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on*, pp. 2115–2120, IEEE, 2010.
- [9] B. U. Töreyn, Y. Dedeoğlu, and A. E. Çetin, “Hmm based falling person detection using both audio and video,” in *Computer Vision in Human-Computer Interaction*, pp. 211–220, Springer, 2005.
- [10] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, “Bathroom activity monitoring based on sound,” in *Pervasive Computing*, pp. 47–61, Springer, 2005.
- [11] A. Gárate, N. Herrasti, and A. López, “Genio: an ambient intelligence application in home automation and entertainment environment,” in *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pp. 241–245, ACM, 2005.
- [12] B. Yuksekkaya, A. A. Kayalar, M. B. Tosun, M. K. Ozcan, and A. Z. Alkar, “A gsm, internet and speech controlled wireless interactive home automation system,” *Consumer Electronics, IEEE Transactions on*, vol. 52, no. 3, pp. 837–843, 2006.
- [13] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, “Robust environmental sound recognition for home automation,” *Automation Science and Engineering, IEEE Transactions on*, vol. 5, no. 1, pp. 25–31, 2008.
- [14] A. Colaco, I.-J. Kim, and C. Schmandt, “Back talk: Engagement sensing for enhancing video viewing,” in *Consumer Electronics (ICCE), 2011 IEEE International Conference on*, pp. 101–102, IEEE, 2011.
- [15] A. Colaço, I.-J. Kim, and C. Schmandt, “Back talk: An auditory environment for sociable television viewing,” in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, pp. 352–356, IEEE, 2011.

- [16] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, N. Noury, *et al.*, “Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living,” *New Developments in Biomedical Engineering*, 2010.
- [17] O. Chapelle, B. Schölkopf, A. Zien, *et al.*, *Semi-supervised learning*, vol. 2. MIT press Cambridge, 2006.
- [18] J. Dennis, T. H. Dat, and H. Li, “Image representation of the subband power distribution for robust sound classification.,” in *INTERSPEECH*, pp. 2437–2440, 2011.
- [19] L. Lu, H.-J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 504–516, 2002.
- [20] L. Ma, B. Milner, and D. Smith, “Acoustic environment classification,” *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, no. 2, pp. 1–22, 2006.
- [21] S. Chu, S. Narayanan, and C.-C. Kuo, “Environmental sound recognition with time–frequency audio features,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1142–1158, 2009.
- [22] J. Ramirez, J. M. Górriz, and J. C. Segura, “Voice activity detection. fundamentals and speech recognition system robustness,” *Robust Speech Recognition and Understanding*, vol. 6, no. 9, 2007.
- [23] 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, IEEE, July 2009.
- [24] 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, IEEE, Dec. 2008.

- [25] B.-H. Juang, W. Hou, and C.-H. Lee, “Minimum classification error rate methods for speech recognition,” *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 257–265, 1997.
- [26] M. J. F. Gales, *Model-based techniques for noise robust speech recognition*. Ph.D. thesis, University of Cambridge, Cambridge, UK, 1995.
- [27] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *Speech and audio processing, IEEE transactions on*, vol. 2, no. 2, pp. 291–298, 1994.
- [28] C. Clavel, T. Ehrette, and G. Richard, “Events detection for an audio-based surveillance system,” in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 1306–1309, IEEE, 2005.
- [29] P. Renevey and A. Drygajlo, “Entropy based voice activity detection in very noisy conditions,” *threshold*, vol. 5, no. 5.5, p. 6, 2001.
- [30] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, vol. 19. IET, 1979.
- [31] T. Kristjansson, S. Deligne, and P. Olsen, “Voicing features for robust speech detection,” *Entropy*, vol. 2, no. 2.5, p. 3, 2005.
- [32] J. Haigh and J. Mason, “Robust voice activity detection using cepstral features,” in *TENCON’93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on*, pp. 321–324, IEEE, 1993.
- [33] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, “Applying support vector machines to voice activity detection,” in *Signal Processing, 2002 6th International Conference on*, vol. 2, pp. 1124–1127, IEEE, 2002.
- [34] M. Farsinejad and M. Analoui, “A new robust voice activity detection method based on genetic algorithm,” in *Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian*, pp. 80–84, IEEE, 2008.

- [35] 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.
- [36] 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.
- [37] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, 1990.
- [38] C.-P. Chen, J. Bilmes, and D. P. Ellis, “Speech feature smoothing for robust asr,” in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings: March 18-23, 2005, Pennsylvania Convention Center/Marriott Hotel, Philadelphia, Pennsylvania, USA*, pp. 525–528, IEEE, 2005.
- [39] M. Cowling and R. Sitte, “Comparison of techniques for environmental sound recognition,” *Pattern Recognition Letters*, vol. 24, pp. 2895–2907, Nov. 2003.
- [40] J. W. Picone, “Signal modeling techniques in speech recognition,” *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.
- [41] 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.
- [42] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.
- [43] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, “Language identification using gaussian mixture model tokenization,” in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–757, IEEE, 2002.

- [44] A. Ramalingam and S. Krishnan, “Gaussian mixture modeling of short-time fourier transform features for audio fingerprinting,” *Information Forensics and Security, IEEE Transactions on*, vol. 1, no. 4, pp. 457–463, 2006.
- [45] T. K. Moon, “The expectation-maximization algorithm,” *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.
- [46] A. Stolcke, S. Kajarekar, and L. Ferrer, “Nonparametric feature normalization for svm-based speaker verification,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1577–1580, IEEE, 2008.
- [47] B. Heisele, P. Ho, and T. Poggio, “Face recognition with support vector machines: Global versus component-based approach,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, pp. 688–694, IEEE, 2001.
- [48] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 209–215, 2003.
- [49] 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.
- [50] 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, Citeseer, ACM Press, 2001.
- [51] S. Knerr, L. Personnaz, and G. Dreyfus, “Single-layer learning revisited: a step-wise procedure for building and training a neural network,” in *Neurocomputing*, pp. 41–50, Springer, 1990.
- [52] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard, *et al.*, “Comparison of classifier methods: a case study in handwritten digit recognition,” in *Pattern Recognition, 1994*.

- Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, vol. 2, pp. 77–82, IEEE, 1994.
- [53] B. H. Juang, “Speech recognition in adverse environments,” *Computer Speech and Language*, vol. 5, pp. 275–294, July 1991.
- [54] R. Lippmann, E. Martin, and D. Paul, “Multi-style training for robust isolated-word speech recognition,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP’87.*, vol. 12, pp. 705–708, IEEE, 1987.
- [55] H.-G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.
- [56] 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.
- [57] S. S. Haykin, *Adaptive Filter Theory*, 4/e. Pearson Education India, 2005.
- [58] 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.
- [59] 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.
- [60] C.-P. Chen and J. A. Bilmes, “Mva processing of speech features,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 257–270, 2007.
- [61] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, *et al.*, “Automatic speech recognition and speech variability: A review,” *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

- [62] 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.
- [63] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, pp. 101–116, Sept. 2005.
- [64] B. Raj, M. L. Seltzer, and R. M. Stern, "Robust speech recognition: the case for restoring missing features," in *Proceedings of Eurospeech*, 2001.
- [65] Q. Huo, C. Chan, and C. H. Lee, "Bayesian adaptive learning of the parameters of hidden markov model for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 334–345, 1995.
- [66] I. Ding, "Improvement of mllr speaker adaptation using a novel method," *International Journal of Information Technology*, vol. 5, no. 1, pp. 12–17, 2009.
- [67] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech communication*, vol. 42, no. 1, pp. 93–108, 2004.
- [68] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.
- [69] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *Signal Processing, IEEE Transactions on*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [70] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3, pp. 229–240, 1996.
- [71] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.
- [72] M. Brandstein and D. Ward, "bsgnai processing techniques and application," 2001.

- [73] N. V. Thakor and Y.-S. Zhu, “Applications of adaptive filtering to ecg analysis: noise cancellation and arrhythmia detection,” *Biomedical Engineering, IEEE Transactions on*, vol. 38, no. 8, pp. 785–794, 1991.
- [74] B. Widrow, J. R. Glover Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong Jr, and R. C. Goodlin, “Adaptive noise cancelling: Principles and applications,” *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.
- [75] R. Fletcher and M. J. Powell, “A rapidly convergent descent method for minimization,” *The Computer Journal*, vol. 6, no. 2, pp. 163–168, 1963.
- [76] P. S. R. Diniz, *Adaptive filtering: algorithms and practical implementation*. Springer, 2013.
- [77] B. WIDROW, M. E. HOFF, *et al.*, “Adaptive switching circuits,,” 1960.
- [78] S. S. Narayan and A. Peterson, “Frequency domain least-mean-square algorithm,” *Proceedings of the IEEE*, vol. 69, no. 1, pp. 124–126, 1981.
- [79] J. J. Shynk, “Frequency-domain and multirate adaptive filtering,” *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.
- [80] M. Dentino, J. McCool, and B. Widrow, “Adaptive filtering in the frequency domain,” *Proceedings of the IEEE*, vol. 66, no. 12, pp. 1658–1659, 1978.
- [81] D. T. Larose, “k-nearest neighbor algorithm,” *Discovering Knowledge in Data: An Introduction to Data Mining*, pp. 90–106, 2005.
- [82] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*, vol. 201. Springer, 1993.
- [83] S. M. Kay, “Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1),” 1993.
- [84] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

- [85] T. Kadir and M. Brady, “Non-parametric estimation of probability distributions from sampled signals,” tech. rep., Technical report, OUEL, 2005.
- [86] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions,” *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011.
- [87] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, 1987.
- [88] E. Parzen, “On estimation of a probability density function and mode,” *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [89] R. C. Gonzalez and R. Woods, *Digital Image Processing*. Prentice Hall Upper Saddle River, NJ, 2002.
- [90] L. Cieplinski, “MPEG-7 color descriptors and their applications,” in *Computer Analysis of Images and Patterns*, pp. 11–20, Springer, 2001.
- [91] M. Basseville, “Distance measures for signal processing and pattern recognition,” *Signal processing*, vol. 18, no. 4, pp. 349–369, 1989.
- [92] S. Nakamura, et al., “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” in *Proc. ICLRE*, pp. 965–968, 2000.
- [93] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in *ICML*, vol. 97, pp. 412–420, 1997.
- [94] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [95] A. Jain and D. Zongker, “Feature selection: Evaluation, application, and small sample performance,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 2, pp. 153–158, 1997.

- [96] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 1, pp. 4–37, 2000.
- [97] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.
- [98] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- [99] “Audio Pro European SFX Library <http://www.soundideas.com/>.”
- [100] G. Fung and O. L. Mangasarian, “Proximal support vector machine classifiers,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 77–86, ACM, 2001.