ACOUSTIC EVENT DETECTION IN POLYPHONIC ENVIRONMENTS USING ARTIFICIAL NEURAL NETWORKS

Jayawardhana Pathiranage Manesh Mihiranga

(188016R)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2021

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/ dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis/ dissertation under my supervision.

Name of the Supervisor: Dr. Sulochana Sooriyaarachchi

Signature of the Supervisor: Date:

ACKNOWLEDGEMENTS

It would never be possible to finish my dissertation without the encouragement, support, and supervision of various personalities, including my supervisor, family, friends and colleagues. At the end of this thesis, I would like to thank all those people who made this achievable and memorable experience for me.

First and foremost, I would like to thank my supervisor Dr. Sulochana Sooriyaarachchi for the continuous support and guidance I received in every aspect while completing this research.

I would also like to thank my progress review committee, Dr. Charith Chitraranjan and Dr. Ranga Rodrigo for their valuable insights and guidance. Their advice helped me to improve the state of my research work.

This work was funded by a Senate Research Committee (SRC) Grant of the University of Moratuwa. Further, I would like to thank for the committee for the financial support.

Finally, I would like to express my sincere gratitude to all of my friends and colleagues for motivating me all the time to do my best not limited to academic work. I thank my parents who have given me a fortunate life and always believing, trusting and supporting me.

ABSTRACT

Our environment is a mixture of hundreds of sounds that are emitted by different sound sources. These sounds are overlapped in both time and frequency domains in an unstructured manner composing a polyphonic environment. Identification of acoustic events in a polyphonic environment has become an emerging topic with many applications such as surveillance, context-aware computing, automatic audio indexing, health care monitoring and bioacoustics monitoring.

Polyphonic acoustic event detection is a challenging task aimed at detecting the presence of multiple sound events that are overlapped at a particular time instance and labeling. It requires a large amount of training data with a complex machine learning architecture thus making it a highly resource-consuming task. Hence, the accuracy of this research area is still not at a satisfactory level.

This study presents a neural networks-based classifier architecture with data augmentation and post-processing methods to improve accuracy. Two neural network architectures as a multi-label and combined single label are implemented and compared in the study. Previous literature reveals that Mel frequency cepstral coefficients and log Mel-band energies are the widely used features in the state of the art research in the area. Different data augmentation methods were used to ensure that the neural networks are trained for even the slight variations of the environmental sounds. A novel binarization method based on the signal energy is proposed to calculate the threshold value for binarizing the source presence predictions. Finally, the median filter based post processing was implemented to smoothen the detection results. The experimental results show that the proposed binarizing method improved the detection accuracy and recorded a maximum of 62.5% combined with the data augmentation and post-processing.

Keywords: Polyphonic Acoustic Event Detection, Dynamic Threshold Binarization, Deep Neural Networks

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACF	Autocorrelation Function
AED	Acoustic Event Detection
ANN	Artificial Neural Networks
BER	Band Energy Ratio
CNN	Convolutional Neural Network
CSL	Combined Single Label
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transformation
DNN	Deep Neural Network
DTB	Dynamic Threshold Binarization
DWTC	Discrete Wavelet Transform Coefficients
FFT	Fast Fourier Transformation
FTB	Fixed Threshold Binarization
GMM	Gaussian Mixture Models
GPU	Graphics Processing Unit
HMM	Hidden Markov Models
k-NN	k-Nearest Neighbor
LPC	Linear Prediction Coefficients
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
ML	Multi Label
NMF	Non-negative Matrix Factorization
PC	Personal Computer
PLP	Perceptual Linear Prediction
RMSE	Root Mean Square Energy
RNN	Recurrent Neural Network
SED	Sound Event Detection

SOM	Self Organizing Maps
STE	Short-Time Energy
SVM	Support Vector Machine
ZCR	Zero-Crossing Rate

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Acoustic Event Detection and Analysis

We can hear a variety of sounds that are emitted by one or more sound sources in our day to day life. The natural acoustic environment is made of several sound events that are mixed in both temporal and spectral domains. A segment of audio that a human listener can consistently label and distinguish in an acoustic environment is defined as an acoustic event [1]. Sound Event Detection (SED) or Acoustic Event Detection (AED) is the process of recognizing and distinguishing the presence of particular sound events with the onset and offset times and associate a label for the events. AED is related to many practical applications such as medical diagnosis, health care monitoring, criminal investigation, acoustic surveillance, context-aware computing, automatic audio indexing, multimedia event detection and bio-acoustic monitoring. Acoustic events are good descriptors for auditory scenes, as they help to understand and describe human and social activities. For example, the sound of traffic will be a feature of the urban environment, but the sound of a gunshot will be abnormal for the urban environment. If a system can automatically differentiate and identify such sound events, it can be used as an acoustic surveillance tool. On the other hand, if we can record the sound of our body organs from outside or inside of the body and automatically analyze the sounds to identify variations from a healthy body, we can use it as a supporting instrument for doctors. In this research, acoustic event detection is considered as the activity of detecting, identifying and labeling the sound events that are present at a time frame.

Until recently, most of AED tasks were manual processes that required significant time and expert knowledge. However, manual processes are not practical in continuous applications such as acoustic surveillance and continuous bio-acoustic

monitoring. Furthermore, it might also be challenging to fully automate applications which require high accuracy and precision. For example, medical diagnosis and criminal investigation are highly sensitive applications that may require human intervention. Fortunately, automatic AED has become a reality with the recent developments of machine learning and signal processing techniques and there are two main approaches for acoustic analysis, namely;

1. Monophonic acoustic event detection and analysis
2. Polyphonic acoustic event detection and analysis

1.1.1 Monophonic acoustic event detection and analysis

An acoustic environment which consists of one sound event at a particular time can be identified as a monophonic environment. As shown in figure 1.1, there can be more than one sound source. But only one sound event is present at a given time (i.e. no overlapping sound events). However, there are no ideal examples of a monophonic environment that we can observe in a natural environment. An acoustic environment in a sound isolated studio where one instrument is played at a time can be given as an example for a monophonic environment.

Acoustic event detection in a monophonic environment is a single label classification problem where only one label is assigned to an identified sound event at a particular time. According to the previous literature, monophonic acoustic event identification considered the most prominent sound event that is present at a particular time frame.

1.1.2 Polyphonic acoustic event detection and analysis

A polyphonic acoustic environment can be defined as an environment where multiple sound sources that simultaneously emit sounds. The additive combination of these sounds makes the environment polyphonic in which it is more complex and challenging to identify and distinguish simultaneous acoustic events. Figure 1.2 shows a sample representation of a polyphonic environment where multiple

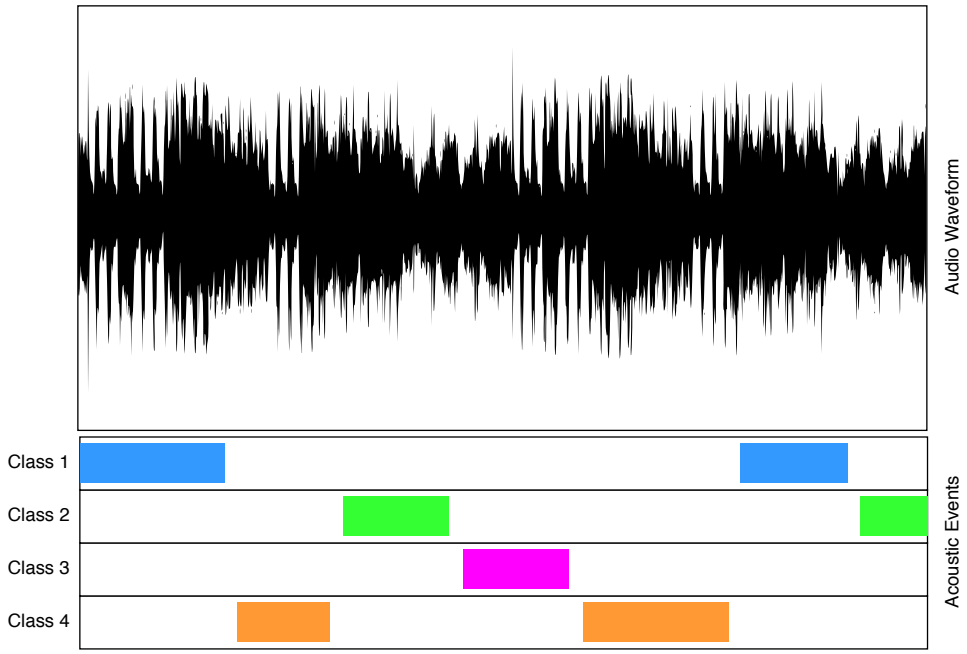

Figure 1.1: Representation of acoustic events in a monophonic environment

sound events can be identified at a particular time. The number of sound events that present at a particular time is called polyphony level [2]. For example, if we can distinguish five sound events at a time, the polyphony level is 5. AED becomes more challenging when the polyphony level getting higher. The natural environment is an ideal example of a polyphonic environment. As far as urban acoustic environments are considered, there are hundreds of sound sources including people, vehicles, birds, environmental phenomena such as rain and wind, machines and street music. There are multiple overlapped sound events that can be identified at a particular time in such an environment.

Acoustic event detection in a polyphonic environment is a multi-label classification problem where the classifier should be able to assign more than one label to a particular time frame.

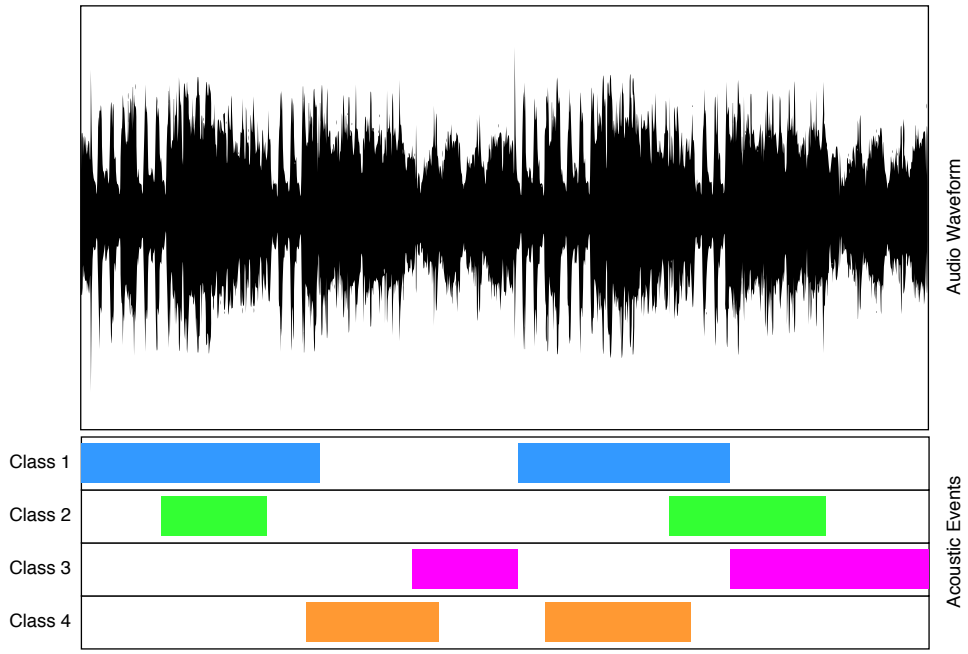

Figure 1.2: Representation of acoustic events in a polyphonic environment

1.2 Motivation

Automatic acoustic processing, detection and classification comprise a wide research area in which polyphonic AED is one of the most challenging areas because of the following reasons. The overlapped sound events can be mixed in both temporal and spectral domains. Therefore, it is hard to classify by decomposing the sound events. The unstructured pattern of occurrence of acoustic events in real-life makes the AED more challenging. There can be several sound sources that emit sounds belonging to the same class but with different characteristics. For example, the car horn is produced by several models of cars have different characteristics. It requires a large dataset to train a robust and highly accurate classifier in this complex scenario. It is also challenging to derive an effective feature set for better classification accuracy. There is no well-accepted robust feature set for this task although different feature combinations have been proposed for AED.

A few work has focused on pre-processing and post-processing techniques to

improve accuracy While most of the authors focused on features and classifier architecture. Data augmentation methods that have been used in speech recognition applications have not been tried for this polyphonic AED yet. As claimed in [3], the threshold value used for binarization varies based on the polyphony level. Implementing a dynamic threshold calculation method will further increase the classification accuracy. Therefore, this research aims to compare the performance of widely used feed forward neural network based classifiers along with state of the art feature sets with data augmentation, low computational binarization, and post-processing methods.

1.3 Problem

Problem addressed in this research is to find suitable data pre-processing and post-processing methods to improve the accuracy of the feed-forward neural networks for identification of acoustic events in polyphonic environments.

1.4 Research Objectives

The objectives of this research are

- to analyze the problem of acoustic events identification in a polyphonic environment
- to analyze suitable methods for identifying acoustic events in polyphonic environments
- to analyze possible methods to improve the accuracy of the existing classifier architectures
- to implement pre-processing and post-processing methods to improve the performance of the existing classifiers
- to evaluate the performance of the developed methods using urban acoustic datasets as a case study

1.5 Thesis Outline

The rest of the thesis is organized as follows: In chapter 2, previous literature related to AED is critically evaluated. Available datasets, features, and feature combinations and different classifier implementations in the previous literature are explained and compared in this chapter. Chapter 3 explains the methodology including selection of the dataset, data augmentation, feature extraction, training of the classifiers, and post-processing. Then chapter 4 summarizes how the experiments were conducted including hardware and software specifications. Chapter 5 presents the experimental results and the interpretations. Finally, chapter 6 concludes the thesis by summarizing the research and proposing future work.

Chapter 2

LITERATURE REVIEW

Automatic AED has been a popular research area during the last few decades. The problem of automatic AED is approached in previous literature under multiple aspects such as datasets, data augmentation, pre-processing, feature extraction, classification, and post-processing techniques. The following sections critically analyze previous literature under the above aspects.

2.1 Feature Extraction

Feature extraction and selection is a crucial step in every machine learning problem. In the audio processing domain, features extracted from the audio should encode a high dimensional audio input which is made up of thousands of samples, into a parametric representation with less number of values. That encoded representation should include prominent information that is useful to train a robust model for the detection and classification task. Many combinations of features have been mentioned in the previous literature. Some of these features are standard features used in signal processing and music information retrieval. Some have introduced novelty features needed for recognition. These features can be divided into six main classes.

1. Temporal (Time-based) features
2. Spectral (Frequency-based) features
3. Time-frequency-based features
4. Cepstrum-based features
5. Energy-based features
6. Biologically or perceptually driven features

Different representations have advantages and disadvantages depending on the sound type they are trying to represent.

2.1.1 Temporal Features

The temporal features are extracted directly from the representation in the time domain. There are severarl temporal features that have been used in previous literature on acoustic processing tasks.

Zero-Crossing Rate (ZCR)

ZCR is the number of times that a signal crosses the zero axis (or the signal sign changes) in a considered time frame. ZCR is a simple measure of the dominant frequency of a signal within the considered frame. ZCR can be defined by below equation.

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|w(n-m) \tag{2.1}$$

where

$$sgn[x(n)] = \begin{cases} 1 & ; \quad x(n) \geq 0 \\ -1 & ; \quad x(n) < 0 \end{cases}$$

and

$$w(n) = \begin{cases} \frac{1}{2N} & ; \quad 0 \leq n \leq N-1 \\ 0 & ; \quad otherwise \end{cases}$$

Pitch Range Features

Minimum and maximum values of the waveform in a considered time frame is one of the standards included in MPEG-7 audio standards. This gives two values for each time frame.

2.1.2 Spectral Features

Features extracted from the spectrum of the signal are called spectral of frequency domain features. In most cases, power modulus of the signal is used to calculate feature values. We can identify more than ten spectral features in the previous literature and those can be defined as below.

Fourier Coefficients

This is the baseline feature that can be identified under the spectral domain. Discrete Fourier Transformation (DFT) is applied to the signal and values are averaged in squared modulus over a pre-defined number of sub-bands. For the computational efficiency, Fast Fourier Transformation (FFT) which is the faster version of DFT is used by many people. The number of coefficients that we can get as features is the number of sub-bands.

Band Energy Ration (BER)

The ratio calculated by normalizing the energy of frequency sub-bands by the energy of the signal during the considered time frame is defined as BER. The size of the feature set is the number of sub-bands considered for the calculation.

Spectral Centroid

Spectral centroid represents the center of gravity of the spectrum. In detail, it gives out the frequency where the energy of the signal is concentrated. This can be defined as the first moment of the spectrum concerning the frequency.

$$SC = \frac{\sum_{n=0}^{M} n|X(n)|^2}{\sum_{n=0}^{M} |X(n)|^2} \tag{2.2}$$

Spectral Bandwidth/ Spectral Spread

The second central moment of the spectrum is defined as the spectral bandwidth. It represents how the energy is spread throughout the frequency spectrum.

Spectral Entropy

The entropy of the normalized spectral energies for a set of sub-frames.

Spectral Flatness

Spectral flatness (tonality coefficient) is a measure to quantify how much noise-like a sound is. A high spectral flatness (closer to 1.0) indicates the spectrum is similar to white noise. It is often converted to Decibels (dB).

Spectral Roll-off

Spectral roll-off measures the spectral skewness of the spectral shape of the signal. It gives out the frequency below which some percentile of the cumulative spectral power resides. Typically, a 95% percentile is used when calculating. This is mentioned as a useful feature to distinguish voiced and unvoiced speech.

$$SRF = max\left(SRF| \sum_{n=0}^{SRF} |X(n)|^2 < Th \sum_{n=0}^{M} |X(n)|^2 \right) \tag{2.3}$$

2.1.3 Time-frequency-based Features

These features extracted by a function that represents two-dimensional (2D) characteristics of both temporal and spectral domains (e.g.: spectrogram, wavelets). This kind of 2D representation is useful when the frequency variation against time is a significant factor. Moreover, spectrogram type representation (or other 2D representations) of the signal is used with the image analysis methods in the recent literature. Different time-frequency representations and parameters that are derived from those are mentioned in the previous literature related to the acoustic analysis.

Short-time Fourier Transform (Spectrogram)

A spectrogram is generated by applying the Fourier Transform considering a frame or a window over the time axis. These frames may or may not overlapped

(in most cases, a 50% overlap of the frames has been considered). Usage of the spectrogram was mentioned in many ways in the previous literature. Other than taking the spectrogram values directly, several features have been derived (e.g.: bandwidth, mean and standard deviation, peaks count in a time window) using a spectrogram as the intermediate representation.

Wavelet Coefficients

The wavelet transform gives a combined representation of frequency and time of the audio signal. This allows the time resolution to be dependant on the frequency. Therefore, start (onset) of a given frequency band can be located at a given time frame.

By using the wavelet as intermediate transformation, Discrete Wavelet Transform Coefficients (DWTC) is calculated. Here, inverse discrete wavelet transform is applied to the logarithm of the energies of the last wavelets.

Spectral Flux

Spectral Flux measures local changes in the spectrum. It gives quantitative measures on the difference in frequency distribution between two consecutive frames.

$$SF = \sum_{n=0}^{M} ||X_i(n)| - |X_{i+1}(n)|| \tag{2.4}$$

2.1.4 Cepstrum-based Features

These features are based on cepstrum, which is a nonlinear transformation of the spectrum that allows the spectrum envelope to be compactly represented by rejecting fine variations across nearby-frequency bins (for example, the exact location of the harmonics in a periodic signal). Mel Frequency Cepstral Coefficients (MFCC) and MFCC derivatives are the most used cepstrum-based features in the previous literature.

Mel Frequency Cepstral Coefficients (MFCC)

The Discrete Cosine Transform (DCT) of the logarithmic magnitude of the signal spectrum is the cepstrum of the MFCC. Mel-cepstrum is calculated on the outputs instead of directly on the spectrum of the Mel frequency filter bank. A Mel frequency filter bank, a non-linear frequency scale that corresponds to a perceptual linear pitch scale, is constructed according to the Mel scale and is characterized by a low-frequency linear part and a high-frequency logarithmic part.

Approximation for the Frequency to Mel transform can be defined as formula 2.5 and the inverse transform can be derived as formula 2.6. Mel scale, which is the graphical representation of the variation of Mel frequency against the frequency is shown in the figure 2.1.

$$m = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{2.6}$$

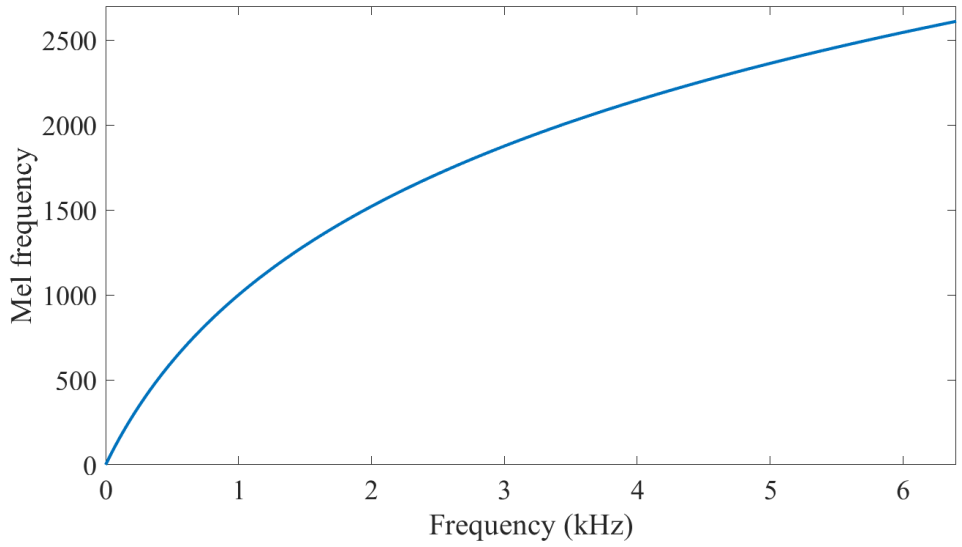

Figure 2.1: Mel scale

If we consider m_k equally spaced points, f_k frequency points can be found as

the perceptual distance is equal by the above formula. However, if we only take samples at f_k frequencies, we will lose all other information. Therefore, we take a weighted sum of the energies near the target frequency f_k as,

$$u_k = \sum_{h=f_{k-1}+1}^{f_{k+1}-1} w_{k,h}|x_h|^2 \tag{2.7}$$

where, $w_{h,k}$ is a weighting parameter.

Finally, MFCC values can be obtained by taking the DCT of the above u_k parameters. By taking the DCT at the end, it approximately decorelate the signal so that MFCC values are not correlated with each other.

Usually, the weighting parameters $w_{k,h}$ are chosen as a triangular function 2.8. As an example, the first filter bank will start at the first point, reach its peak at the second point, then return to zero at the third point. A graphical representation of the triangular Mel-filterbank is shown in figure 2.2.

$$w_{k,h} = \begin{cases} \frac{h-f_{k-1}}{f_k-f_{k-1}} & ; \quad f_{k-1} < h \leq f_k \\ \frac{f_{k+1}-h}{f_{k+1}-f_k} & ; \quad f_k < h \leq f_{k+1} \\ 0 & ; \quad otherwise \end{cases} \tag{2.8}$$

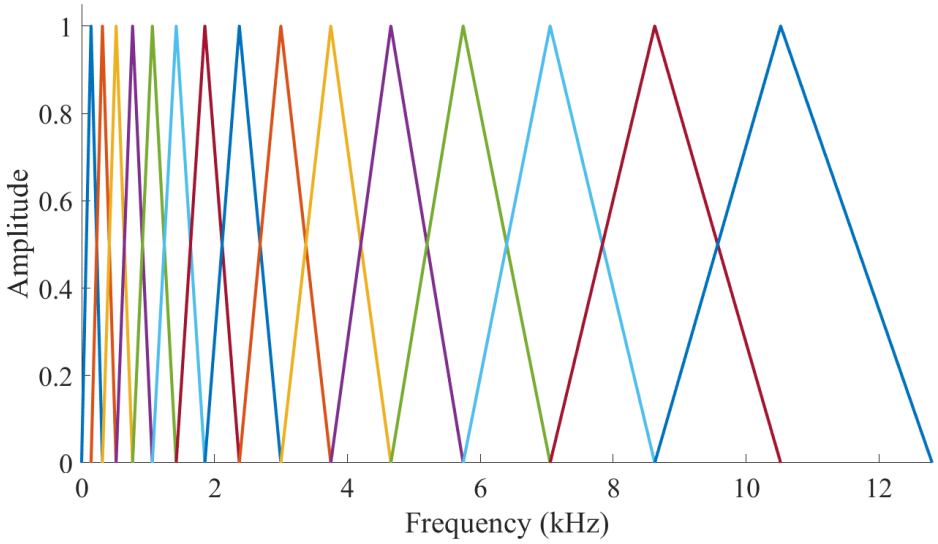

Figure 2.2: Triangular Mel-filterbank $w_{k,h}$

MFCC Derivatives

As a derived feature from MFCC, their derivatives are also used in previous literature. In most cases, first and second derivatives of MFCC through the time axis is calculated considering adjacent frames.

Linear Prediction Cepstral Coefficients (LPCC)

LPCC is a set of values when the Linear Prediction Coefficients (LPC) (see section 2.1.6) are represented in the cepstral domain.

2.1.5 Energy-based Features

Energy-based features can be calculated or in other words, energy can be extracted from any type of signal representation (i.e.: time-domain signal, spectrum or cepstrum). Therefore, those features can be listed under the respective representation as well. In the previous literature, these features have been mostly used in foreground extraction tasks. But some researchers used energy-based features for the acoustic classification tasks although there is a claim that the intra-class variation is significantly high in energy-based features.

Signal Energy

The sum of squares of the signal values within a frame is calculated and then normalized by the respective frame length.

$$E_s = \sum_{n=-\infty}^{\infty} |x(n)|^2 \tag{2.9}$$

Entropy of Energy

The entropy calculated with a set of normalized energy values of consecutive frames. It is identified as a useful measure of sudden changes in the signal.

Log Mel-band Energies

As mentioned in section 2.1.5, the cepstrum domain representation of the signal is used to calculate the energy of each bin. Then the calculated values are converted to the log scale. This is a derived feature from the intermediate step of the MFCC calculation.

2.1.6 Biologically/ Perceptually Driven Features

A class orthogonal to the previous ones is represented by biologically or perceptually motivated features. They can be based on representations of spectral, temporal, time-frequency, or cepstral, but are inspired by hearing psychophysiology and/or human voice. These features can be sub-divided into three subclasses,

1. Features that simulate the processing, in specific cochlear filtering, of the human auditory system.
2. Features that reproduce the psychological or emotional interpretation of auditory signals.
3. Features that are constructed according to the human vocal tract's physical behavior.

Linear Prediction Coefficients (LPC)

LPC is the filter coefficients in an all-pole model that approximates the acoustic signal. This was originally intended to model the human vocal tract response which is also known as Vocoder. Although LPC is widely used for speech vocal analysis applications, literature claims it is not suitable for impulsive ambient sounds.

Perceptual Linear Prediction Coefficients (PLP)

PLP are generated by mapping LPC values to the non-linear frequency scale as the human hearing. Information captured with PLP is somewhat similar to

MFCC. PLP derivatives are also mentioned in the literature related to speech recognition.

Log Frequency Coefficients

These coefficients are generated by a set of bandpass filters with logarithmically scaled bandwidths and central frequencies. According to the literature, the logarithmic scale approximately replicates the response of the human ear's filter bank.

Mel Frequency Coefficients

The calculation is similar to the log frequency coefficients. However, the central frequencies of bandpass filters are scaled as per the Mel scale so that it replicates the psychoacoustical perception of the pitch. The Mel scale is explained in the section 2.1.4 in detail.

Narrowband Autocorrelation Features

The Narrowband Autocorrelation Functions (ACF) are used to extract the features. The filter bank, where the central frequencies are spaced following the Mel scale, is used as a first step to filter the signal. Four features are extracted from the autocorrelation functions for every filter. Calculated four features are related to the four primary perceptual qualities of the sound. Those features and the qualities are as follows,

1. The intensity of the first positive peak to represent the loudness
2. The time delay between the first and second peaks to represent the pitch
3. The intensity of the second positive peak to represent the timbre
4. The duration of the envelope at -10 dB to represent the duration

2.1.7 Usage of the Features

When we analyze the usage of the above-mentioned features in the previous literature, different feature combinations have been tried based on the application.

Recognition of abnormal acoustic events for acoustic surveillance was presented in [4] with a combination of a wide range of features. Cepstral features like MFCC and MFCC derivatives, biologically driven features like LPC, spectral features like centroid, spread, flatness, skewness, roll-off, brightness and roughness, temporal features like ZCR were used to generate the feature vector. Feature extraction was done considering 25 ms frames with a 50% overlap. The Hamming window was used as the window function for feature extraction. 10 different combinations of these features were evaluated for the classification accuracy with Gaussian Mixture Models (GMM) as the classifier. The highest accuracy was given by the feature combination of 24 MFCC, 24 MFCC first derivatives, 12 LPC and spectral statistics listed above.

In [5], feature combination including temporal, spectral and cepstral domain features has been used for audio scene analysis with neural networks. ZCR, Short-Time Energyc (STE), BERs, bandwidth, frequency centroid, and MFCCs were used to generate the 20-D feature vector for each time-frame. 50 ms frames with the 20% overlapping were considered for feature extraction by the researchers. They could achieve a 0.83 F1 score for their experiment on the designed acoustic sensor network. Two-layer hybrid architecture with Hidden Markov Models (HMM) and neural networks was used for scene detection and classification.

Statistical values including the standard deviation and the mean of the spectrogram, MFCC, energy entropy, and spectral roll-off were used to generate a 12-D feature vector for the detection of violent scenes in film scenes [6]. Frame-based feature extraction was done with non-overlapping frames. According to the authors, feature selection and the statistics selection was a result of complex experimentation. Audio-based detection and classification results (F1 score) were 53% and 52.7% respectively which is 6.1% and 1.6% better than video-based results.

As the authors of [7], there is a significant difference in environmental sounds and speech. Therefore, we may need to consider features that can address the high non-stationarity of environmental sounds. They have used features derived from DFT, DCT, and DWT with other basic features. ZCR, short-time average energy, spectral centroid, spectral roll-off, MFCC, LPCC, PLP, and new wavelet-based features were proposed to address the problem of environmental sound recognition. They claim that the one-dimensional features like ZCR, short-time energy, spectral centroid, and roll-off fail to represent the data information. However, those features can improve the performance of the classifiers since they represent the information of two complementary domains. Moreover, the authors claim that DWTC did not perform well when used alone since DWTC tend to neglect high frequencies.

A relatively larger frame size which is 200 ms with a 50% overlap was used by the authors of [8] to detect abnormal sounds in a metro station. They evaluated the performance of MPEG-7 audio protocol descriptors and MFCC separately and combined. Spectrum flatness, waveform min-max, and audio fundamental frequency were the descriptors they employed.

Novel 2-D feature set based on the pitch range of audio and autocorrelation function was proposed in [9] for the classification of non-speech environmental sounds. Authors claim most of the environmental sounds are non-stationary because those sounds contain significant rapid changes in the spectrum. When calculating their novel features, the time delay between the first and the second positive peaks that are calculated from the ACF. The reciprocal of the time delay is defined as the pitch. Based on those values, two features are defined: 1) ratio of the maximum to minimum of the pitch range; 2) ratio of the standard deviation and the mean value of the pitch range. The performance of the novel features was tested with different classifiers and compared with MFCCs. Although the novel features did not perform well when they were used alone, a combination of MFCCs and proposed pitch range-based features could improve the accuracy of the classifiers up to 19%.

Authors of the [10] proposed a method of extracting the discriminative feature

set using a boosting approach from a large pool of features. 26 MFCCs and 26 log frequency filter bank parameters were used as two baseline feature sets. The first and second derivatives of each feature set were also calculated resulting in 78 feature components per set. AdaBoost which is a weak classifier selection method was used to select the features from the pool of the above feature components. They have shown that both detection and classification accuracies were higher for the feature set selected by boosting.

Although most of the above-discussed literature used omnichannel audio for feature extraction, recent works of literature have proposed features extracted from the multichannel audio as well. According to the authors of [11], multichannel features give a better representation of polyphonic audio. The reason they claim is different sound sources in the polyphonic audio show different intensities when we analyze as binaural channels. The authors evaluated three binaural features as binaural log Mel-band energies, time difference of arrival, and dominant frequencies. Feature extraction was done with 40 ms frames at 20 ms hop length. Extracted features both channels were layered one over the other to feed the Convolutional Recurrent Neural Network (CRNN) classifier. Proposed features gave a 6.1% improvement of the F_1 score over the monaural features for the same neural network architecture on the TUT-SED 2016 dataset.

The usage of image-based features is another approach that has been mentioned in the literature. Audio is converted to 2-D time-frequency representation (spectrogram in most cases) and image-based features are extracted. Authors of [12] used constant Q-transform (CQT) to represent the audio as a time-frequency image and extract image-based features. 30 s audio is converted to an image with 134 frequency bands and 500 frames. According to the authors, we can do feature learning in 3 ways: 1) taking individual frames; 2) taking a set of consecutive frames; 3) taking the full spectrogram image. The first method is not a good approach as it lacks significant temporal characteristics. The third method can take unreasonable time for computations which is also not a good option. Therefore, most of the literature used the second approach which is also known as context windowing. The authors used 2 s long context windows which

lead to 15 windows per 30 s recording. Average pooling along the time was done focusing on the matrix factorization techniques.

When we critically analyze the usage of the features and the feature extraction methods, most of the literature mentioned time-frequency features and/ or cepstral features as their main feature. Other features such as temporal, spectral, and energy-based features have been used with the time-frequency or cepstral features to improve the performance of the classifiers. Most of those features have been used in the multi-level classification architectures. Since the feature improvement was not the main objective of this research, widely used features such as MFCCs and Mel-band Energies were selected to proceed with the experiments.

2.2 Detection and Classification

After the feature extraction step, those features which are the parametric representation of the audio are used to train the classifier against the annotations. The terms acoustic events detection and acoustic events classification are used to refer to the same task in many works of literature. Typically, the AED categorized based on the classification method.

Based on the previous literature, we can divide classification methods into 3 main categories as,

1. Generative Classifiers
2. Discriminative Classifiers
3. Hybrid Classifiers

2.2.1 Generative Classifiers

Generative classifiers are based on the Bayesian concept where the posterior probability is assigned as the classification score of the test sample. In the classification process, multiple classifiers are used as one for each class and finally, the class is selected by the classifier which gives the highest posterior probability among

the evaluated classifiers. GMM and HMM are the widely used classifiers for the audio classification applications in the previous literature.

When we consider the advantages of the generative methods, the high interpretability of the models can be highlighted. Since these methods are defined in a Bayesian context, we can interpret how the classifier works during the training process. Other than that, the high generalization of the model and the ability to deal with the missing data are other advantages of these classifiers. However, the performance of the generative classifiers is slightly lower (moderate performance) than the discriminative classifiers in the previous literature.

2.2.2 Discriminative Classifiers

Discriminative models do the classification by a hyperplane which is calculated to separate the feature space as most training samples segregate to a correct class. The training phase of the discriminative models is more resource consuming compared with the generative models. Artificial Neural Networks (ANN) and Support Vector Machine (SVM) can be highlighted as widely used discriminative classifiers employed in the previous literature related to the acoustic classification domain. Other than ANN and SVM classifiers, Self Organizing Maps (SOM), k-Nearest Neighbor (k-NN) and Non-negative Matrix Factorization (NMF) have been mentioned in the literature.

When we analyze the previous literature, these methods have given better accuracy compared to generative methods. Therefore, the main advantage of the discriminative classifiers is the noticeable performance. However, the interpretability of these classification models is low.

2.2.3 Hybrid Classifiers

Researchers have tried complex classifier architectures which are a combination of the above two methods. A hierarchical classification approach with two or more classifier models is used in here. GMM with SVM, NMF with ANN, HMM with ANN are some combinations that have been mentioned in the previous literature.

Hybrid classifiers have shown improved accuracy in the literature. But the main disadvantage they mention is the complexity of the implementation and the significantly higher training time. We have to deal with many parameters to tune the model.

2.2.4 Usage of the Different Classifiers

The majority of the literature focused on monophonic AED where they assume that only one acoustic event is present at a considered time instance possibly overlapped with generic background noise. When we analyze the usage of classifiers in the past literature, the focus was moved towards neural networks based architectures recently. However, we can see the usage of different classifiers with different feature combinations for audio classification. Some applications of audio processing have experimented widely and have robust classifiers and feature combinations. For example, speech recognition and speaker recognition have shown acceptable performance in previous literature with a robust feature set.

When considering the audio data, we can notice that acoustic events with semantic meaning (e.g: human vocals) can be divided into a sequence of non-semantic atomic units. Those units have a well-defined and stable acoustic signature. Classification of these atomic units is easier in the perspective of machine learning approaches. But in most cases, those atomic units do not relate to meaningful classes from a human interpretative perspective. Also, since the overlapping event does not arise in a specified order or pattern, polyphonic acoustic events are more complex to classify.

Authors of the [13], directly translated the bag-of-words approach that was proposed in [14] into the bag-of-aural-word where the audio segments play the role of patches that allow one to deal with complex acoustic patterns made of several audio units where its order is unknown. The authors claim that the approach is robust in the presence of ambient noise because the proposed classifier can be trained to ignore the respective words related to the background noise. A limited set of classes including gunshot, screaming, and breaking glass that are mixed

with the background noise were used in the context of acoustic surveillance.

More complex implementation while considering the inner temporal structure of the acoustic segments is presented in the [15]. The authors modeled the audio units in an unsupervised approach with a set of HMMs and a random forest classifier. HMMs translate the acoustic signal into a set of acoustic units. The random forest classifier is fed with the histogram to get final classification results.

In [16], each segment of the audio signal is classified directly into a set of predefined classes by the LVQ algorithm. For each classification, the degree of confidence is calculated and the frames with a confidence level above a threshold value are taken to the final step. Voting between the high confidence frames is used to assign the final label. Unlike the earlier studies, each segment of the audio is associated with the same set of classes where the whole sequence of the audio is classified.

A fully unsupervised approach is presented in [17] for AED. First, a dictionary learning procedure like non-negative matrix factorization is used to extract the multi-dimensional sequence of symbols of the audio sequence. Deviation of Motif Discovery algorithm which is used for the analysis of DNA sequences is used to detect the recurring onsets of acoustic events.

Multi-style training for HMM is proposed by the authors of [7] for environmental sound identification. One model was trained using multi-conditional data that includes both clean and noisy data. Maximum A Posteriori and Maximum Likelihood Linear Regression algorithms were used to enhance the performance of the system further by reducing the environmental variability.

When addressing the polyphonic classification, an approach with two steps were proposed in [18] and [19]. First source separation techniques are used to unmix the overlapped sources and then classify the signals resulted after the unmixing. However, this approach is suboptimal because the signal is distorted to some extent at the unmixing step. Attractive options for this problem have been given by multi-label classification [20] based methods in recent literature.

Application of multi-label classification for the sound restricted domain such as music and bioacoustics can be found in the literature. In [21], multi-label

classification is applied for music emotion identification. The authors labeled 593 songs with 6 emotion class labels. 4 multi-label classification algorithms were compared including label powerset, binary relevance, random k-labelsets and, multi-label k-NN where the first 3 algorithms are problem transformation methods and the last one is an algorithm adaptation method. Based on the comparison results, random k-labelsets performed well over the other 3 methods.

Multi-instance multi-label (MIML) framework based method was proposed in [22] for the classification of the multiple simultaneous sounds of bird species. MIML bag generator which is an algorithm to transform input audio signal to bag of instances was used as the representation of the audio for the classifier. The authors achieved high accuracy of 96.1% for the classification of 13 bird species.

The performance of the multiple binary classifiers against a single multi-class classifier is compared in [23]. According to their claim, a hierarchical combination of multiple binary classifiers performs well if the number of classes in the output vector increases. In [4], a 3-level hierarchy was used to detect abnormal acoustic events. Sounds are subdivided into harmonic and non-harmonic categories at the top-level classifier. Then at the second level, voiced and non-voiced sound events are identified. Finally, at the third level, actual sound events are identified. Based on the results, this hierarchical implementation has shown a 1.33% improvement in the accuracy over single-level implementation.

Complex hybrid classifier architecture is proposed in [10] which is a combination of ANN, HMM, GMM and SVM. At the first stage, sequence modeling capabilities of HMM and context-dependent discriminative capabilities of ANN were used in cascade to identify Maximum A Posteriori probabilities and segment boundaries for each class. A GMM model was trained using the identified probabilities and finally, the parameters of the GMM model were fed to SVM. The authors called it the SVM-GMM super-vector approach while they claim 45% performance improvement.

Complex neural network implementations such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have been proposed in recent literature with better results. Multi-label Bi-directional Long Short Term Mem-

ory (BLSTM) RNN is proposed in [24] with an average F1 score of 65.5% and 64.7% for the identification of 1 s blocks and 50 ms frames respectively. The authors tested the model on a large dataset that has 61 sound event classes of real-life recordings. The reported accuracy is a reasonably good value since they have considered a significant number of classes.

CNN with 6 convolutional layers and 3 fully connected layers has been proposed for AED in [25]. This architecture was inspired by VGG Net [26] which is characterized by its simplicity. It replaces the large convolutional kernels by a 3x3 kernels stacked on top of each other without pooling between these layers. The authors adapted their CNN architecture for multiple instances learning as proposed in[27]. There the data is organized as bags where each bag contains a set of training instances. While the instance labels are known, only the bag level labels are provided. If one or more instances in the bag are positive, that bag is identified as a positive bag where negative bags means that all the instances are negative. The authors used a large augmented dataset with 28 sound events for their research and the proposed architecture reached 92.9% maximum accuracy.

Most of the previous work addressed the AED and speech recognition problem as a segmentation and classification task. The authors of the [28] addressed monophonic AED as a regression task where they proposed a random forest regression-based method in their research. 100 ms segments of the soundscape are defined as superframes and those superframes are then divided into 30 ms frames with 20 ms overlap. A feature set with 60 features were used to represent the short frames in the superframe. The architecture consists of 3 regression models to distinguish the foreground and background superframes, to recognize superframes between sound event categories, and to estimate the onset and offset of the events.

When we critically evaluate the usage of the different classifiers, the trend has been moved towards the discriminative methods such as neural networks, SVM, and hybrid methods with a hierarchical approach. However, when we consider the complexity of the implementation and the high computation requirements of the hybrid methods, neural networks based classifier architecture is selected for

this research. Since the main objective of the research is to improve the detection accuracy by improving the pre-processing, binarization, and post-processing, a deep feed-forward neural network is selected as the classifier. Multi-label and combined single label classier architectures will be compared so that future experiments can be done by taking the proposed method as a baseline.

2.3 Datasets

Urban sounds are an ideal polyphonic acoustic environment and this research studies urban sounds as the case study as previously mentioned in 1.4 There are several datasets of urban sound recordings mentioned in the previous literature such as TUT-SED Synthetic 2016 [3], UrbanSound8k [29], The URBAN-SED [30], CHiME-Home [31]. Datasets in [29, 31] are recorded in the real environment whereas datasets in [3, 30] are synthesized using randomly mixed and overlapped sound event samples.

2.3.1 TUT-SED Synthetic 2016

This is a synthesized dataset that was created by mixing different sound events under 16 sound event classes. Synthesizing process has used 994 isolated sound events and the total length of the synthesized soundscapes in the dataset is 566 minutes. According to [3], the polyphony level (defined in section 1.1.2) of the soundscapes varies from 0 to 5. This dataset comes with pre-categorized training, testing and validation sets with 60%, 20%, and 20% for each set respectively. Different instances of the sound events are used to synthesize the training, validation, and test partitions. Sound events were randomly selected inside the set for mixing. Selected events were segmented to a length of 3-15 seconds and separated by a random length of a silent region during the mixing process.

However, the authors did not add any background noise to the soundscapes when synthesizing. Therefore, it contains the noise in the isolated events and therefore the background noise depicts abrupt changes throughout the soundscape. Such noise variations are highly unlikely in real-life scenario and it is

identified as a main drawback of this dataset.

2.3.2 UrbanSound8k

UrbanSound8k is a dataset of short audio snippets containing 10 classes of isolated sound events, namely children playing, air conditioner, engine idling, car horn, gunshot, dog bark, siren, drilling, street music, and jackhammer. According to [29], "children playing" and "gunshot" are added for the variety of sound events. Besides that, all other classes are highly frequent in urban acoustic environments. There are 8,732 labeled isolated sound recordings with a total duration of 525 minutes. Although this dataset contains a significant amount of recordings, the dataset as it is does not represent a polyphonic environment considered in this research. It was required to manually mix and overlap the sound snippets to generate a polyphonic dataset.

2.3.3 The URBAN-SED

This dataset is synthesized using the clips from the UrbanSound8K dataset as the source material. The total duration of the soundscapes in the dataset is almost 30 hours and it includes 10,000 synthesized soundscapes. Among those 10,000 soundscapes, there are nearly 50,000 annotated sound events where each soundscape consists of 5 sound events on average. The soundscapes are saved in single channel, 44.1 kHz and 16-bit format. The length of a single soundscape is 10 seconds and there are 10 sound event classes as mentioned in section 2.3.2. Brownian noise is added for all the soundscapes as background noise when synthesizing the dataset. The authors in [30] claim that the added noise resembles the "hum" which is common in urban acoustic environments.

This dataset is also presorted into train, validation and test sets. UrbanSound8K dataset consists of 10 folds and URBAN-SED uses these folds as follows without mixing;

- Folds 1-6 were used to generate the train set with 6000 soundscapes
- Folds 7-8 were used to generate the validation set with 2000 soundscapes

- Folds 9-10 were used to generate the test set with 2000 soundscapes

The dataset contains annotations of the soundscapes in two formats, namely JAMS and TXT. JAMS format was used to give the full annotation while TXT contains the simplified version of annotations as tab-separated values. Both versions include at least the start and end times of every sound event with the label.

2.3.4 CHiME-Home

This dataset contains sound recordings of the domestic environment. The duration of the recordings is 6.8 hours containing 7 labels associated with the sources of the sound. Child speech, adult male speech, and adult female speech are human speaker events in the dataset. Sounds such as footsteps, bang, knock, and crash are labeled as percussive sounds that represent human activities. TV and video games are labeled under another class. The sound of the household appliances is labeled as broadband noise. And finally, there is another class for other identifiable sounds. Recordings were divided into 4-second non-overlapping chunks for the annotation process. Three human listeners individually annotated the segments resulting in 3 multi-label annotation sets. Jaccard indices between sets were used to evaluate the annotations and make the ground truth.

The main drawback of this dataset is it does not give detailed annotation with on-set and off-set of sound events. Therefore, we have to consider 4 s segments if we use this dataset. The precision of the classification would be very low in that case.

In this research, high polyphony levels are desirable for evaluating the performance of proposed binarization methods. For high accurate frame-based classification it is also required to have the exact onset and offset of the sound events. Although "TUT-SED Synthetic 2016" is suitable for polyphonic AED experiments the maximum polyphony level is 5 in the dataset. On the other hand, "UrbanSound8k" dataset is not a polyphonic sound dataset while "CHiME-Home" dataset does not provide the exact onset and offset of the sound events. In conclusion, "The URBAN-SED" dataset is the best option for this research

because the polyphony level goes up to 7 and the dataset comes with detailed annotations with the exact onset and offset thus reducing additional work required for training the classifier.

2.4 Summary

This chapter explored previous literature in the main areas of feature sets, classifiers, and datasets covering the data augmentation and post-processing methods that have been used as well. Different types of features have been mentioned in the literature that are categorized as temporal, spectral, time-frequency, cepstral, energy, and biologically driven features. Most of the authors used cepstral features like MFCC, energy-based features like log Mel-band energies, and biologically driven features like LPC as their main set of features while other features were used as secondary features.

Classifiers that have been used in the literature can be divided into 3 main categories as generative, discriminative, and hybrid. At the early stage, most of the authors used generative methods like HMM and GMM. However discriminative classifiers, mainly neural networks have experimented in the recent literature showing significant improvement compared to the generative methods. Some authors proposed hybrid methods with a hierarchical classification approach with a combination of both generative and discriminative methods which also shown good results. To improve the accuracy of the classification, different preprocessing and post-processing methods have also been implemented. But most of the polyphonic AED work focused on improving the features and the classifiers mainly.

There are several datasets available for the polyphonic AED task. Most of the datasets are manually synthesized using the isolated sound events while there are very few datasets with real-world recordings.

When the previous literature is critically evaluated, we can decide that the polyphonic AED problem is an emerging research area that is still to be improved.

Chapter 3

METHODOLOGY

This research methodology can be divided into 6 main steps as dataset selection, data augmentation, feature extraction, detection and identification, post-processing, and evaluation of results. The block diagram of figure 3.1 shows an overview of the methodology with the available methods that have been found in the literature.

This section describes the formation of the research methodology with the findings of the previous literature and how the experimental setup was finalized.

3.1 Baseline Study

Authors who proposed "The UrbanSED" dataset [30] have evaluated their dataset on the state-of-the-art model which is a CRNN proposed by Cakir et al. [3]. The proposed architecture consists of four main steps.

1. A context window of log Mel band energies is used as the features and those features are fed to convolutional layers with non-overlapping pooling over the frequency axis.

2. At the last convolutional layer, feature maps are stacked over the frequency axis and fed to recurrent layers.

3. The output of the recurrent layers is fed to a single feedforward layer with sigmoid activations.

4. Event presence probabilities given by the sigmoid activations are then binarized by a constant threshold value to obtain the prediction results.

The model was implemented in Keras. Adam optimizer with binary cross-entropy loss function was used when training the model. Performance of the

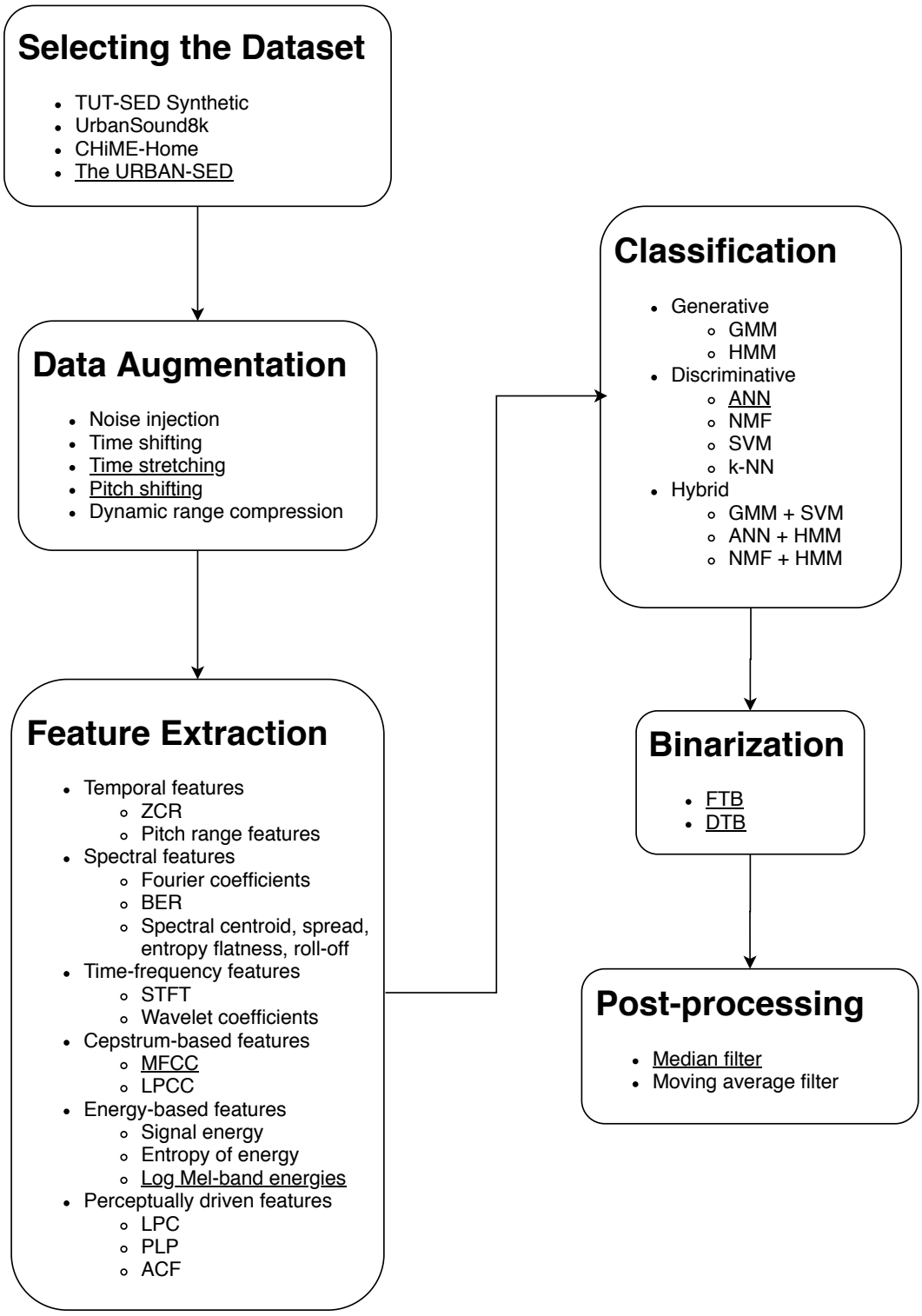

Figure 3.1: Overview of the methodology

implementation was presented as the 1 s segment-based F1 score. Based on the results, the authors claim an overall F1 score of 58.8% for the test of the dataset

containing 2000 soundscapes.

3.2 Formation of the Methodology

In the baseline study, authors have tried a complex neural network architecture with widely used log Mel band energies as the features. However, the authors have not tried out any data pre-processing and post-processing methods in their work. And most importantly they have used constant value to binarize the event presence probabilities. As claimed in [32], binarization threshold value should vary based on the polyphony level. Therefore, using a constant value for thresholding is not the optimum way.

After evaluating the findings of the previous literature and baseline study, research methodology is formed to increase the accuracy of the neural network-based classifiers by implementing preprocessing and post-processing methods with a dynamic threshold calculation method. Since CRNN architecture proposed in the baseline study requires significant processing power, this research is focused on implementing deep feedforward neural network with other supplementary modules to test the performance.

3.3 Selecting the Dataset

When the available datasets mentioned in Section 2.3 are compared, three of those have some drawbacks relating to this research. "TUT-SED Synthetic 2016" [3], is a potential dataset for polyphonic AED experiments. However, the maximum polyphony level is 5 in the dataset. Since a higher polyphony level is preferred to check the performance of the proposed binarization method, this dataset was not selected.

"UrbanSound8k" [29] dataset is not a polyphonic sound dataset. Another step is required to synthesize polyphonic sound data if the UrbanSound8k is used. Since it is not the main objective of the research, Therefore, it also does not match the requirements of the research.

The main drawback of the "CHiME-Home" [31] dataset was it does not pro-

vide the exact onset and offset of the sound events where it only annotated the presence or the absence of any sound event throughout the whole soundscape. Therefore, those annotations cannot be used for the frame-based classification.

When comparing all the above-mentioned datasets, the "The URBAN-SED" [30] dataset is the best option for this research. Its polyphony level goes up to 7 which is a better number that suits the current research. The dataset comes with detailed annotations with the onset and offset of each sound event which reduces the additional work required for preparing the annotations. The dataset is available free of charge under the terms of the Creative Commons Attribution 4.0 International License [33].

3.4 Data Augmentation

Machine learning models, especially deep neural networks significantly depend on the size of the training dataset and the quality of the data. The availability of adequate samples of training data that represents the possible variations of the real-world scenario leads the algorithm to learn a non-linear function from input to output which generalizes well. The correctly generalized model yields high classification accuracy on unseen data.

Although the selected dataset (The URBAN-SED) has 6000 soundscapes for training, all those soundscapes are synthesized using the isolated sound events. To represent the possible variations of the real-world environment, data augmentation is used for the training soundscapes. In the process of data augmentation, different types of deformations are applied to the annotated data. A better classification accuracy can be expected when the model is trained with such additional deformed data because the model would become more invariant to the possible deformation in a real-world scenario.

There are several widely used audio data augmentation methods mentioned in the literature.

1. Noise injection (adding background noise) [34, 35]

2. Time shifting (shifting audio to left or right along the time axis) [36]

3. Time stretching (changing speed) [34, 35]

4. Pitch shifting (changing pitch) [34, 35]

5. Dynamic range compression [35]

Since the Brownian noise was added when synthesizing the selected dataset, noise injection was already performed on the dataset. Time shifting is also not useful in my research problem since I did a frame-based classification. Therefore, shifting the audio forward or backward also shifts the frame-based features and annotations resulting in the same training set. Dynamic range compression is the process of compressing the dynamic range of samples with different parameters. For example, the audio can be compressed into film or music standards keeping a higher dynamic range while speech or radio quality compression has a lower dynamic range.

When applying time stretching, the audio is stretched by a given stretch factor. In a real-world acoustic environment, there could be slight variations among the same acoustic events as well. To populate the training set with the possible temporal variations, we can apply time stretching with a small stretch factor.

Pitch shifting is the process of shifting up or shifting down the pitch of the audio. This is also a possible scenario in the natural environment where sound events that are categorized under the same class can show variations of the pitch (ie. different car horns with different pitches). Pitch shifting is also selected as a data augmentation method in this research to ensure the test set represents possible variations of the pitch among the same sound event.

3.5 Feature Extraction

At the training phase, raw data cannot be fed to the classifier. Hence, a parametric representation of the data is used to represent the qualities of the data. Feature extraction is the process of preparing that parametric representation. As explained in Section 2.1 different types of features have been mentioned in

the previous literature. Some authors used them alone while some publications mention different feature combinations.

Temporal features are extracted from the time domain representation of the audio. Therefore those features miss the frequency domain characteristics of the audio. Spectral features that are extracted from the frequency domain representation also have the same drawback where those miss the time-domain characteristics. Because of that reason, both spectral and temporal features have not been used alone in the literature.

Time-frequency-based features represent the variations of the frequency parameters against the time. Therefore, different features derived from the time-frequency representation have been used in the literature for AED applications. However, the linear scale of the frequency axis might miss the prominent frequencies since it gives equal space for all the frequency bins. Cepstrum-based features are a viable solution for that problem since those are based on the nonlinear transformation of the spectrum. Biologically/ perceptually driven features such as LPC and PLP are widely used in speech recognition applications but not in environmental sound identification.

In the previous literature, MFCCs and log scald Mel band energies are widely used as primary features for polyphonic AED. As explained in the section 2.1.4, MFCCs are motivated by the non-linear frequency feature of the human auditory system and have shown good performance in audio processing and identification applications [37]. MFCCs represent selected energy in different frequency bands as the feature of the target. MFCCs has been one of the most used feature in audio processing and speaker recognition system because it is a good simulation of the human auditory system. Other than MFCCs, log scaled Mel band energies have also shown good results in previous literature for AED applications. Mel band energies are also a cepstrum domain enery-based feature that has some similarities to the MFCCs as well. These two feature sets have been selected to do the experiments of the research and results are compared.

3.6 Classification

As explained in the section 2.2, several classifiers have been mentioned in the previous literature. Artificial neural networks has been selected as the classifier since ANN-based architectures have shown significant improvement in recent research work. When we address the polyphonic AED, there can be more than one acoustics event present at a given time. To identify those simultaneous acoustic events, the classifier architecture should be able to assign more than one label for each instance (10 labels in this research).

The multi-label classification problem is addressed in two ways in this research.

1. Multi-label (ML) classifier
2. Combined single label (CSL) classifier

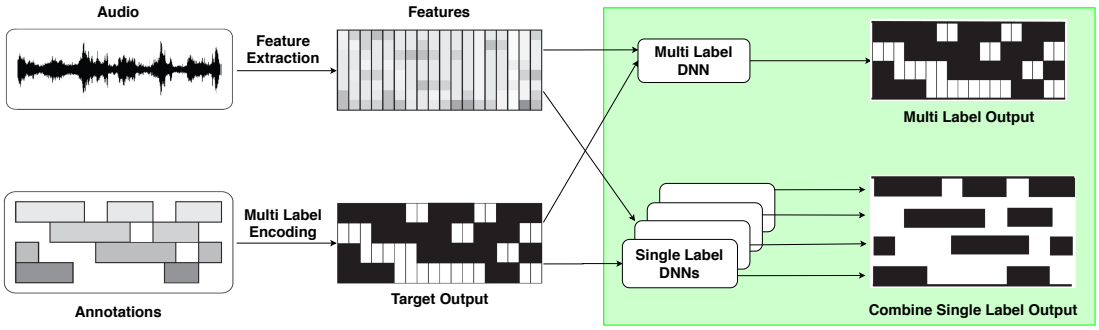

Figure 3.2: Difference between ML-DNN and CSL-DNN

Figure 3.2, graphically shows the high-level difference between the two methods.

The baseline work explained in section 3.1 proposed a complex neural networks based classifier architecture that is a combination of CNN, RNN, and FNN architectures. This is a complex neural network architecture that requires significant processing power as well. But this kind of complex implementation will not be applicable every time. Therefore, simple implementation of a neural network with deep feedforward neural network is proposed in this research.

Multi-label Classifier

A multi-label classifier calculates the output vector with the classification results of multiple classes. In my context, the size of the output vector is 10 which represents the 10 acoustic event classes in the dataset. In a multi-label neural network, the number of nodes in the output layer equals to the number of classes. In this approach, we need to train a single model to identify all the events. Therefore, the complexity of the implementation is lower compared to the CSL classifier.

Combined Single Label Classifier

A combined single label classifier is a set of classifiers that are trained to detect each class individually. The multi-label classification result is obtained by combining the outputs of each single label classifier. In this case, I trained 10 neural networks to identify each sound event class, and then the results of those classifiers are combined to achieve the polyphonic AED. Implementation of the CSL classifier is complex compared to the ML classifier since we need to train a model for each class separately and also we need to do additional work to combine the results of those models to get the final results. However, the CSL classifier has several advantages although the implementation is complex. If we need to add a new class, we can train a model for a particular class and add it to the existing model set. Further, we can customize the classifier by choosing only the necessary classes based on the application.

Since the main objective of the research was to improve the accuracy by the data augmentation and post-processing, basic deep feed forward neural networks were selected as the architecture.

3.7 Binarization

The activation function of the output layer of all the above mentioned neural network architectures is Sigmod function. As mentioned in previous sections sigmoid function bounds the output value between 0 and 1. Since the source

presence predictions are Boolean results, we need a way to map the probability given by the sigmoid function to either 0 or 1 where 0 represents the negative classification and 1 represents the positive classification. This is achieved by a threshold value which divides the output probabilities to the positive and negative results. Two types of binarization methods were tried in this research.

1. Fixed threshold binarization (FTB)

2. Dynamic threshold binarization (DTB)

Both methods were tested on the same classifiers and results are compared to evaluate the performance of the novel DTB method.

3.8 Post-processing

After the binarization process, the output vector gives the labeled classification results for ten classes. A significant number of false positives and false negatives were identified when the results were evaluated. Most of those were spike type of results where consecutive frames that are preceding and succeeding gives the opposite result for the same class. In other words, the classifier identifies sudden sound events that last only 50ms or sudden 50ms silent regions of possible sound events for a particular class.

However as mentioned in [38], a sound event should last long for 200 ms – 300 ms for a human listener to clearly identify it. Based on that finding, we implemented a post-processing method to filter out the false classification results.

3.9 Summary

This chapter evaluated the methods and the findings of the chapter 2 to form the methodology of the study. As the first step, "The Urban-SED" dataset was selected as the most suitable dataset for the task. Then different data augmentation methods are evaluated and time stretching and pitch shifting were selected as the data augmentation methods for the research. Although there is a wide range

of features available for the AED applications, the most used features namely MFCC and log Mel-band energies were selected to train the classifiers and compare the results. Deep feedforward neural networks were selected as the classifier since neural network-based implementations have shown better performance in the recent literature. Polyphonic AED is identified as a multi-class, multi-label classification problem, and two approaches are proposed to address it as ML DNN and CSL DNN. A threshold value is used to binarize the source presence prediction probabilities of the neural network outputs. Novel dynamic threshold binarization method is proposed along with the fixed threshold binarization that has been used in the previous literature. Finally, the post-processing method is proposed to smooth the detection results by filtering the exceptionally short positives or negatives.

Chapter 4

SYSTEM AND EXPERIMENTAL DESIGN

As per the methodology in Chapter 3 the selected methods were implemented and tested for the results. Figure 4.1 shows an overview of the implementation with the selected methods at each step. The green color arrows show the data flow of the training process and the red color arrow shows the dataflow of the testing process. This chapter describes the process of each step and the problems faced in detail.

4.1 Data Augmentation

As the first step, data augmentation was done for the soundscapes in the training set of the dataset. Among the methods mentioned in the section 3.4, time stretching and pitch shifting were used as augmentation methods. Librosa python library [39] was used for the task and it was done on a laptop with intel core i7 processor, 12GB of RAM, and M2 solid-state drive. Fold 5 and 6 of the training set were used for the augmentation and it took less than 30 minutes to do all the augmentations. Augmented soundscapes were saved separately with a suffix appended to the original file name.

4.1.1 Time Stretching

This process changes the duration/ speed of the audio without affecting the pitch. I used time stretching to make sure the training set represents the slight temporal variations of the sound events. This ensures the trained model to be more robust for the real-world scenario where we can observe variations in the same sound class.

I used two stretch factors as x1.2 and x0.8 for time stretching. Fold 5 of the training set was used for the time stretching. x1.2 was applied to the first 500 soundscapes and x0.8 was applied for the other 500 soundscapes.

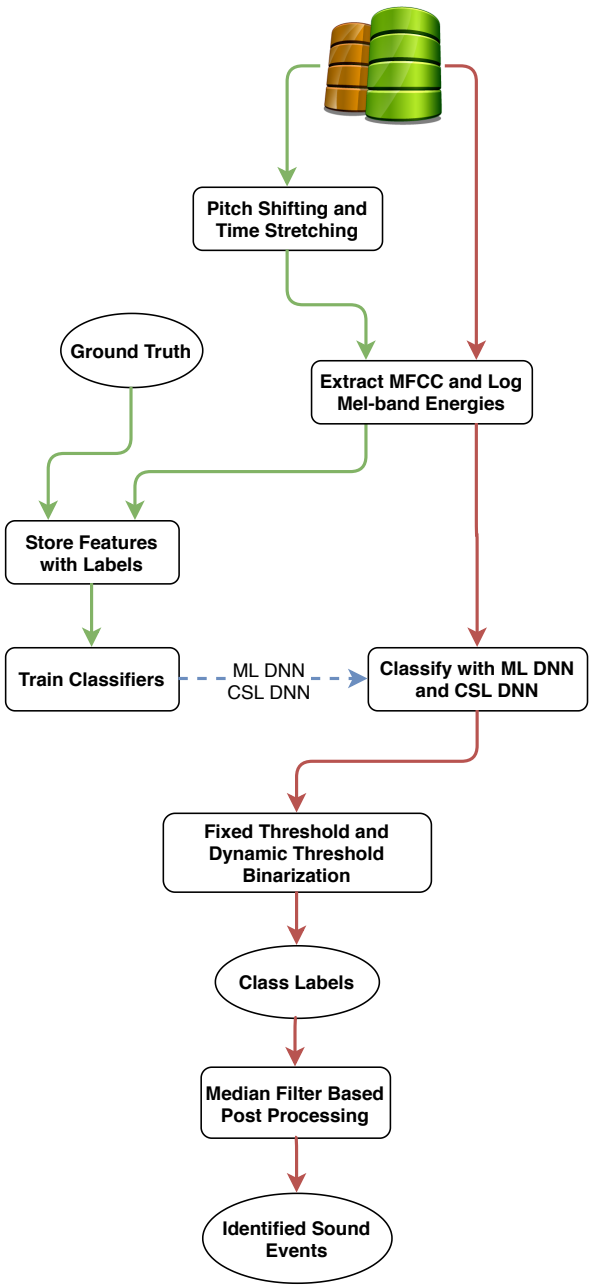

Figure 4.1: Experimental design

As shown in the figure 4.2 we can observe a clear variation of the duration of the audio after applying time stretching. 10 s long original soundscape stretched to 12 s when the stretch factor is 0.8 which slows down the speed of the audio. When the stretch factor is 1.2 audio is fast-forwarded reducing the duration of the soundscape to 8 s.

Since the duration of the soundscape is changed after the augmentation, it

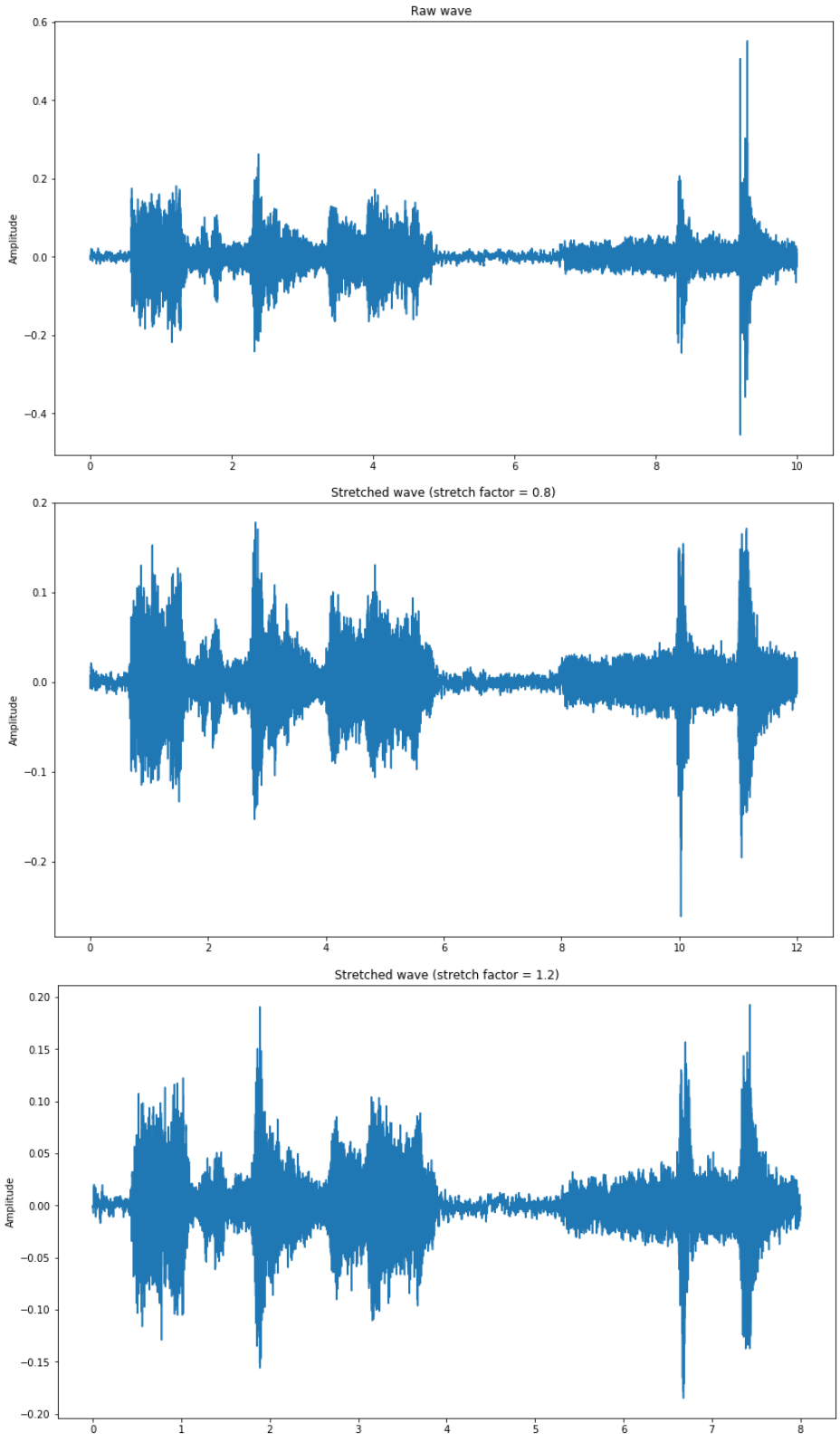

Figure 4.2: Waveforms of the original soundscape (top), stretch factor = 0.8 (middle) and stretch factor = 1.2 (bottom)

affects the later steps including feature extraction and classification. Therefore, I had to bring the augmented audio back to the 10 s length before those steps to make sure the number of samples and frames in each soundscape is constant.

For the fast-forwarded samples where the stretch factor is 1.2, Brownian noise which is the background noise of the dataset was padded to the end of the soundscape. This ensures the consistency throughout the 10 s recording better than zero-padding.

For the slow-downed samples where the stretch factor is 0.8, I had to crop the output to fit the 10 s length. The extra length of the soundscape after the 10 s limit was cut down.

4.1.2 Pitch Shifting

Pitch shifting changes the pitch of the audio without affecting the speed/ duration. This is an implementation of the pitch scaling used in musical instruments. As time stretching represents temporal variations, pitch shifting covers the spectral variations that can be observed in the same sound event class.

The training set was augmented by shifting up and shifting down the pitch by two semitones/ half steps. Same as in the time stretching, fold 6 of the training set was used to apply pitch shifting. The first half was used for shift up and the latter half was used for shift down.

As shown in figure 4.3, we can observe the signal is shrunk along the frequency axis without changing the duration when we shift down by two semi-tones/ half steps. And also the signal is more spread along the frequency axis when we shift up by two semi-tones/ half steps.

4.2 Feature Extraction

As mentioned in section 3.5, two 2-D feature vectors were prepared in the feature extraction process representing MFCCs and log Mel-band energies. Librosa python package was used to extract the features. It provides almost all the necessary implementation for the music and audio information retrieval systems.

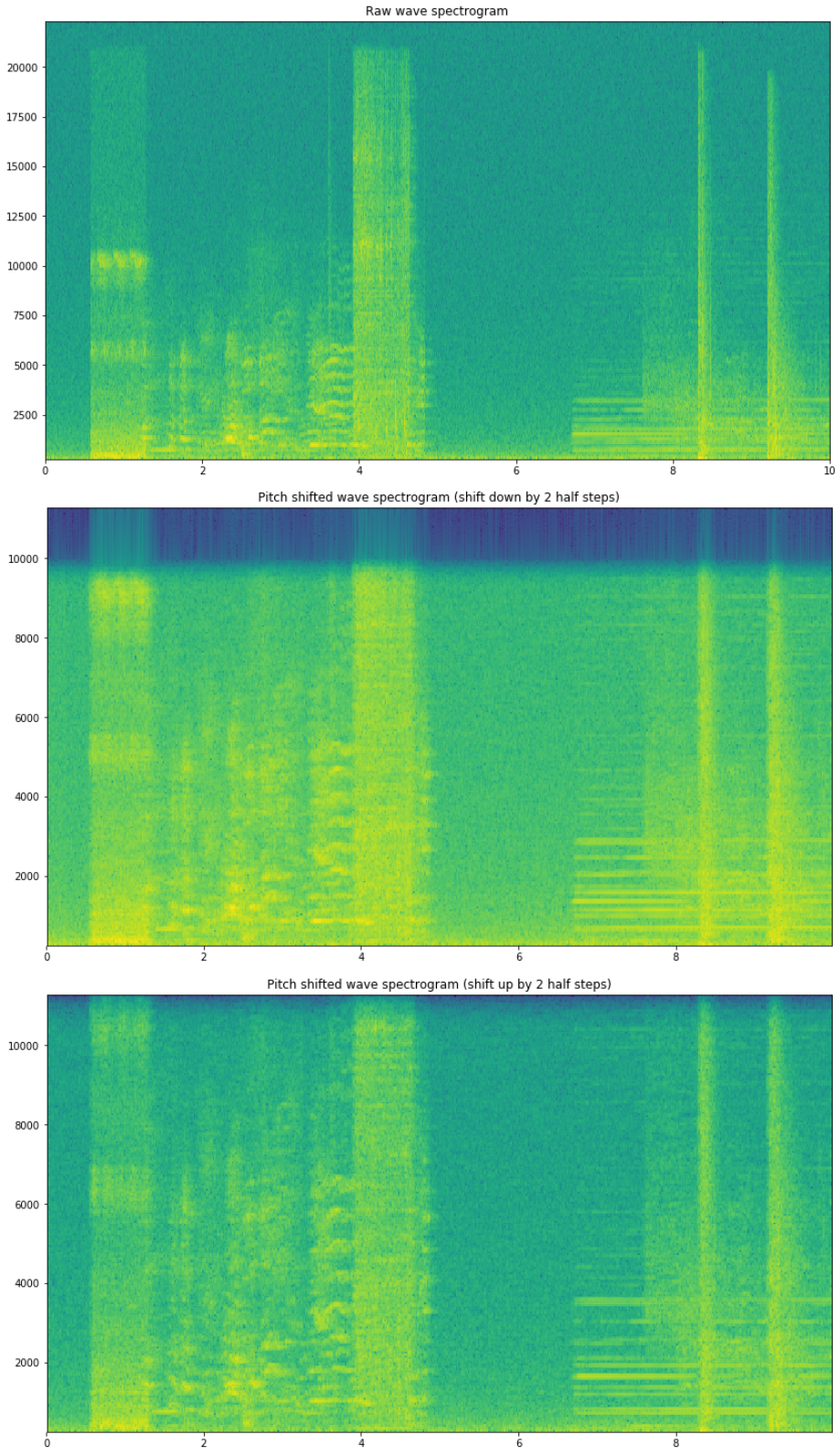

Figure 4.3: Spectrograms of the original soundscape (top), shift down by 2 semi-tones (middle) and shift up by 2 semi-tones (bottom)

50 ms long frames of the soundscapes were considered to extract features. Since the soundscapes are sampled at 44,100 Hz in the dataset, a 50 ms frame consists of nearly 2,250 samples per frame. Overlapping between frames was kept at 50%. This results in 431 frames per 10 s long soundscape. Therefore, the length of the feature vector was 431. Window function that was used for both feature sets was Hamming window. It is defined as the below equation.

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{4.1}$$

Where $0 \leq n \leq N-1$ and N is the length of the window.

Extracted feature vectors were saved as NumPy arrays. Desktop computer with intel core i7 processor, 16GB of RAM, and 2GB NVIDIA GPU was used for the feature extraction. That computer was placed in the university and TeamViewer software was used to connect to the PC.

Then, frame-based multi-label annotations were prepared using the text-based annotations in the dataset. Dataset gives start and end times of each sound event present in the soundscape. Based on those timestamps, respective frames that contain particular sound events are identified. A two-dimensional binary vector was prepared for each soundscape considering the frames that were used for feature extraction. Those annotation vectors were also saved as NumPy arrays. The same desktop computer which was used for the feature extraction was used to prepare annotations.

4.2.1 MFCCs Extraction

librosa.feature.mfcc function was used to extract the MFCC feature vector. The function parameter *n_mfcc* was set to 40 to get the first 40 coefficients. DCT type 2 was used with an ortho-normal DCT basis. Those parameters were selected based on the findings of the previous literature. Figure 4.4 shows the representation of extracted MFCCs on a training set soundscape.

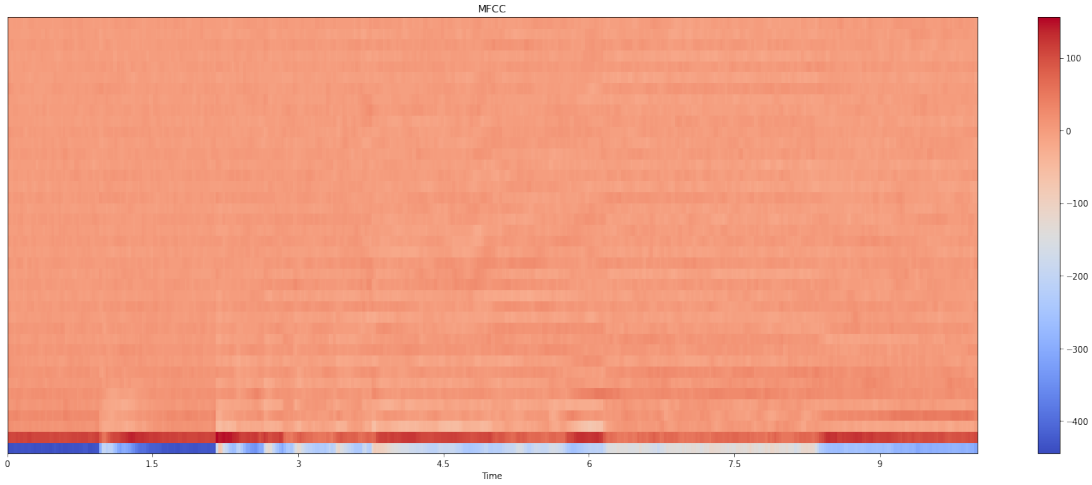

Figure 4.4: Representation of MFCCs

4.2.2 Log Mel-band Energies Extraction

Same as MFCCs, 40 Mel scaled filter banks (Mel bands) were used to calculate the log scaled energy. *librosa.feature.melspectrogram* function was used to compute the log Mel-band energies. Figure 4.5 shows the representation of extracted log Mel-band energies on a training set soundscape.

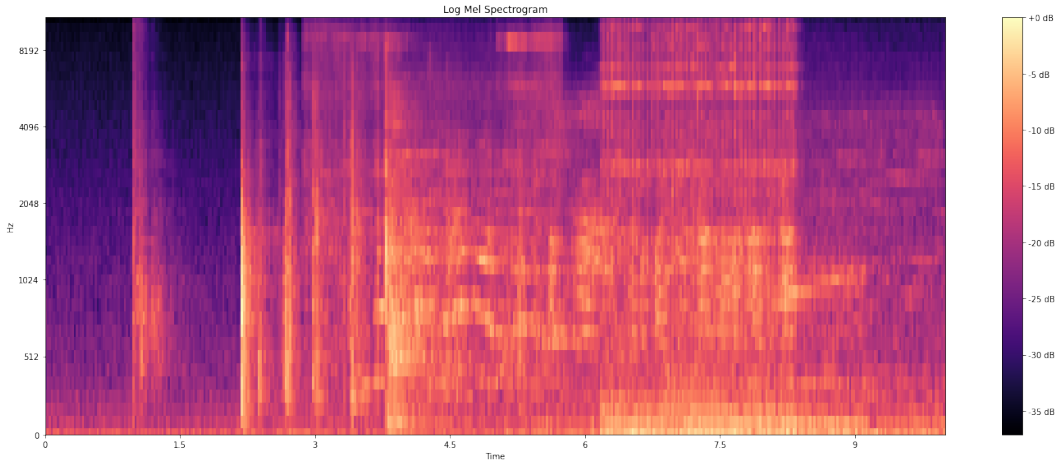

Figure 4.5: Representation of log Mel-band energies

4.2.3 Context Window

As mentioned in [40], the auditory sensitivity of the human ear decreases when the duration of the sound is less than 1 s. According to the authors, the average

duration of the sounds should be between 200 ms to 300 ms for low-intensity sounds for middle age person. Based on that finding I used a context windowing method to preserve the temporal properties of the audio. Since the feature extraction was done taking 50 ms frames with 50% overlapping, a context window with 5 frames was used resulting in 150 ms total duration. It was prepared by concatenating two preceding and two succeeding feature vectors with the feature vector at a given time frame t.

Context window x_t at the time frame t can be denoted as,

$$x_t = [u_{t-2} + u_{t-1} + u_t + u_{t+1} + u_{t+2}] \tag{4.2}$$

where ut is the feature vector extracted at time t.

Prepared context window x_t was used as a training instance for the classifier.

4.3 Classification

Classification is the main step in any machine learning problem. This is the process where we use the extracted features and the encoded annotations to train the classifier model. This section describes the selected classifier architectures and the implementation of those.

4.3.1 Classifiers

At the classification step, frame-wise classification was done using the 5-frames context window as the input. Classification result y_t of length N was obtained for each time frame, where N is the number of sound event classes. For the ML classifier, N is 10 and for the CSL classifier, N is 1. Variable y_t is a vector with binary elements.

In the proposed classifiers, the activation function of the output layer was sigmoid activation. The sigmoid function gives the event activity probability for a particular event class which is a value between 0 and 1. That probability was then binarized using a threshold to obtain the binary result for the classification

results. The binarization module was implemented and experimented in two ways as explained in the section 4.4.

Grid search was used to optimize the hyperparameters of the neural network models. The first 800 soundscapes of fold 1 - 4 of the traing set and 200 augmented soundscapes that were taken from fold 5 and 6 of the training set were used to run the grid search. 200 soundscapes of the fold 7 and 8 (100 soundscapes each) were taken to validate the grid search results. The results of the grid search which are the parameters used for the final models are listed in the section 5.2.

Neural networks were initially trained on the same PC that was used for feature extraction. But due to the power failures at the university, it could not be completed. Therefore, a remote computer on the Google Cloud Platform was used to train models. Google Cloud GPU machine was used with NVIDIA Tesla T4 GPU. GPU memory was 16GB GDDR6 with 4 vCPU and 32 GB memory. Keras with Tensorflow GPU backend was used to train models while utilizing the GPU of the computer. Training of both ML and CSL neural networks and post-processing of the results were done on this remote computer.

In this research, two neural network architectures have been implemented.

1. Multi-label deep feed-forward neural network
2. Combined single label deep feed-forward neural network

Multi-label Deep Feed-forward Neural Network

In the ML DNN, one model is trained to identify all the sound event classes. The number of nodes in the output layer is equal to the number of classes N ($N = 10$ in our case). Therefore, we get output vector y_t of length N for each frame we considered in feature extraction.

The designed neural network was a fully connected DNN with 3 hidden layers and having $ReLU$ as the activation function. A *Sigmoid* function was used as the activation function of the output layer because it bounds the output value between 0 and 1. That output value can be interpreted as the detection probability of a

particular sound event class. Probabilities at the output layer are then binarized using a threshold value as explained in section 4.4.

At the training phase of the DNN, Stochastic Gradient Descent (SGD) was used as the learning algorithm. Binary Cross-entropy was used as the loss function because it measures how far away is the predicted value from the actual value (which is either 0 or 1) for each class and then averages the class-wise errors to get the final loss. Binary cross-entropy can be mathematically explained as,

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=0}^{N}(y * log(\hat{y}) + (1 - y) * log(1 - \hat{y})) \tag{4.3}$$

Where, $\hat{y}$ is the predicted value obtained from the output layer and y is the target output value according to the annotations.

Hyper-parameters of the DNN such as number of hidden layers, number of hidden units, learning rate, initial weight, and bias values were selected by a grid search. The best result was given by the a rchitecture with 3 hidden layers and 800 hidden units each.

Combined Single Label Deep Feed-forward Neural Network

In the CSL DNN, ten different models were trained for each sound event label independent from others. This was a time-consuming task compared with the ML DNN training. Same as in the ML DNN context window of features was fed to the inputs layer. The activation function of hidden layers was ReLU and the output layer was a single node with sigmoid activation. Predictions of all models are concatenated to get the multi-label detection probabilities and the multi-label output vector was binarized by a threshold as explained in section 4.4.

Source presence prediction $\hat{y}$ is a value between 0 and 1 which is the detection probability of a particular class. The actual target value y for each training instance is either 0 or 1. Cross entropy loss function was used to calculate the distance between the predicted value and actual value with the SGD learning algorithm. For a single label classifier, equation 4.3 can be simplified as,

$$L(y, \hat{y}) = -(y * log(\hat{y}) - (1 - y) * log(1 - \hat{y})) \tag{4.4}$$

Same as the ML DNN, hyperparameters were selected by a grid search.

4.4 Binarization

The purpose of the binarization is to assign positive or negative class labels based on the probabilities at the output layer sigmoid activations. Two binarization methods were implemented and compared in this research.

4.4.1 Fixed Threshold Binarization

A constant threshold value was used to binarize the neural network outputs for all frames. The threshold value was set to 0.5 based on the findings of the [32]. This is a computationally faster method since we only need to compare the output probability against the threshold and assign it to the positive or negative result set.

4.4.2 Dynamic Threshold Binarization

Cakir et al in [32] claims that higher polyphony levels show better accuracy for small binarizing thresholds and vice versa. Based on that claim, we propose a dynamic threshold binarization which adapts the threshold value based on the soundscape characteristics. I came up with this method based on the assumption that the energy of the sound at a particular time instance depends on the polyphony level at that time. We evaluated the results of the proposed DTB against the FTB. The steps of the DTB calculation are as below:

Root Mean Square Energy (RMSE) is calculated through the soundscape with a 50ms frame window and 50% overlappings. Let $x(n)$ be the value of a sample, N be the total samples in a frame and R be the RMSE at a particular time frame, E can be defined as,

$$E = \sqrt{\frac{1}{N}\sum_{n}|x_{(n)}|^2} \tag{4.5}$$

Then the ratio of E at frame t to the maximum E is calculated and T value is derived as,

$$T'_{(t)} = 1 - \frac{E_{(t)}}{max(E)} \tag{4.6}$$

Then, to avoid extremely large or extremely small values for the threshold, T' is interpolated between $[0.3, 0.7]$ interval. RMSE and threshold variation throughout the soundscape is shown in figure 4.6.

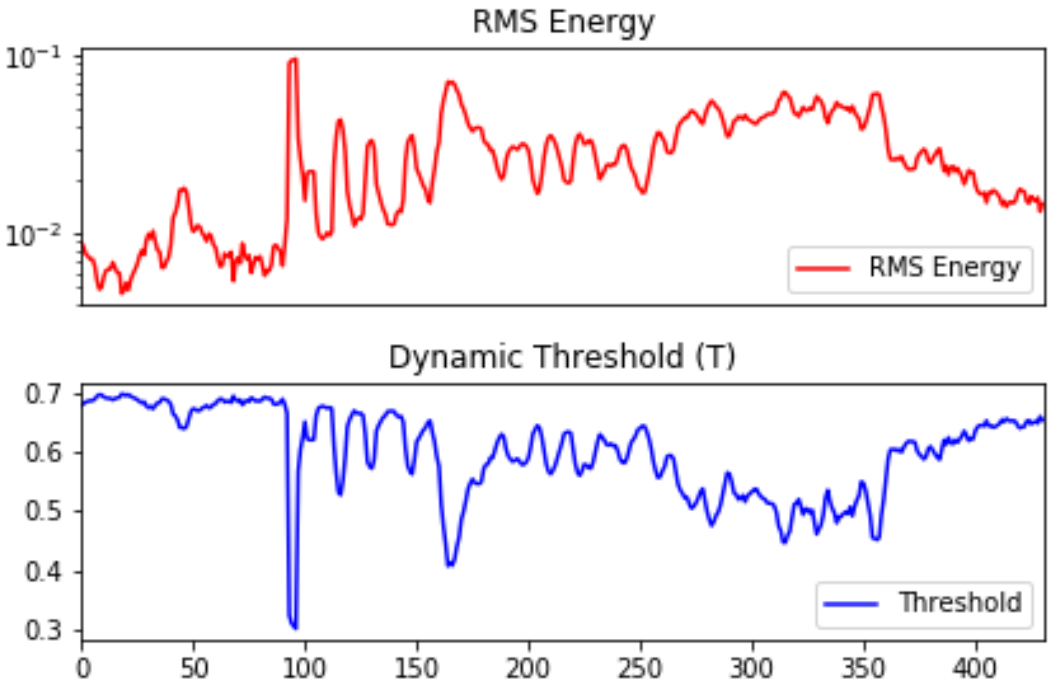

Figure 4.6: RMSE variation of a signal (top) and dynamic threshold value calculated by the algorithm (bottom)

Ten probabilities given by the sigmoid functions of the output layer were compared with the threshold value and binarized the result to get source presence predictions. Evaluation of the event identification was done comparing the output of the binarization step.

4.5 Post-processing

As mentioned in section 3.8, median filter based post-processing was used to smoothen the classification results. This section describes the implementation of the post-processing method with the parameters that were used.

4.5.1 Median Filter Based Post Processing

The median filter is widely used in previous literature to remove the noisy data. In most of the image processing applications, the median filter has been used to filter the salt and pepper noise of the image [41]. However, some researchers used it as a post-processing method to enhance the classification results [42].

In my research, I used the median filter to filter the noisy results and combine the frame-wise classification results over consecutive frames. This filtering was done for each class separately considering the binarized classification results of a particular class. The filtering window was applied through all the frames of the soundscape. Frame window with 9 consecutive frames was used so that it sums up the duration to 250 ms which is an acceptable value that a human listener could identify. Filtering was done by sliding the window by 1 frame through the result vectors.

Let $\hat{Y}_t$ be the filtered result and $\hat{y}_t$ be the original result after binarizing. For a particular class at frame t, the median filter can be explained as,

$$\hat{Y}_t = \begin{cases} 1; & median(\hat{y}_{(t-4):(t+4)}) = 1 \\ 0; & otherwise \end{cases} \tag{4.7}$$

As shown in figure 4.7, exceptionally short positive and negative classification results are filtered after applying the median filter. To evaluate the performance of the post-processing method, classification accuracy before and after applying the filter is compared.

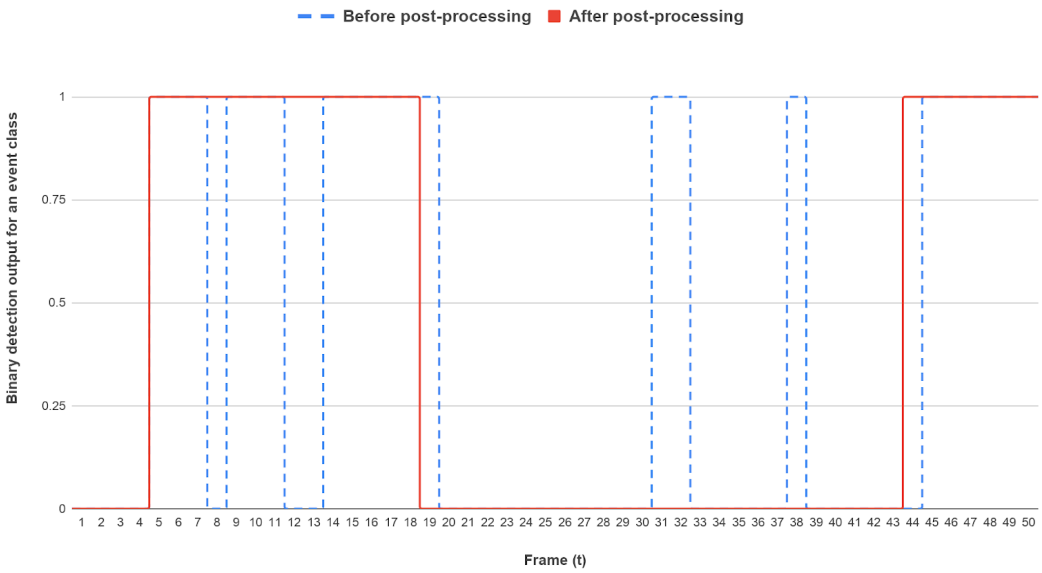

Figure 4.7: Sample representation of median filtering based post-processing

4.6 Summary

This chapter explained the implementation procedure of the methodology proposed in the chapter 3. Data augmentation, feature extraction, classification, binarization, and post-processing methods are explained with all the configuration parameters, equations, and graphs. At the first step, two data augmentation methods were implemented as pitch shifting and time stretching. Then MFCC and log Mel-band energies were extracted from the soundscapes as the features sets for the experiments. Two neural network architectures were implemented as classifiers of this study followed by a binarization module to make binary output vector from the source presence probabilities. Widely used FTB and a novel energy-based DTB method was implemented and compared the performance. As the final stage, moving window of median filter was used to smoothen the detection results by filtering the exceptionally short positive and negative detections.

Chapter 5

RESULTS AND DISCUSSION

This chapter presents the results and a discussion of the the experiments presented in the chapter 4. The first section provides the evaluation metrics that have been considered in the research. Then the experimental results are presented in detail. The final section provides a detailed discussion based on the obtained outcomes.

5.1 Evaluation Metrics

In this research, the evaluation metric to compare the performance or the results of the proposed methods was the F_1 *Score* (or the *F Score*). F_1 score can be defined as the below equations.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.2}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.3}$$

The term *accuracy* refers to the F_1 *score* throughout the thesis. True positives, true negatives, false positives, and false negatives can be explained as below.

- True Positive: If a classifier labels sound event as present at a particular time frame and if it is also present according to the annotations
- True Negative: If a classifier labels sound event as absent at a particular time frame and if it is also absent according to the annotations
- False Positive: If a classifier labels sound event as present at a particular time frame and if it is absent according to the annotations

- False Negatives: If a classifier labels sound event as absent at a particular time frame and if it is present according to the annotations

This work is considered as the baseline model for this research and the results are evaluated compared to this model.

5.2 Results

In this section, the results of the different experiments are explained and compared. Comparison of the selected feature sets (MFCCs and Log Mel-band Energies), classifier architectures (ML DNN and CSL DNN), and binarization methods (fixed threshold and dynamic threshold) are explained with the experimented post-processing method.

Table 5.1 shows the experimental results of the comparison of the DNN architectures and binarization methods for different feature sets where Figure 5.1 shows the comparison in Table 5.1 graphically.

Feature Set	DNN Architecture	F_1 Score % for FTB	F_1 Score % for DTB
Log Mel-band Energies	ML DNN	54.61	57.14
	CSL DNN	52.28	54.16
MFCCs	ML DNN	52.67	55.73
	CSL DNN	49.56	53.62

Table 5.1: Comparison of the binarization methods

Based on the results in Table 5.1, we can see DTB performed well compared to the FTB. The post-processing method proposed in section 4.5 was tested with the DTB outputs to further increase the detection accuracy. Table 5.2 compares the detection accuracy results with and without the post-processing. Figure 5.2 shows the comparison results graphically.

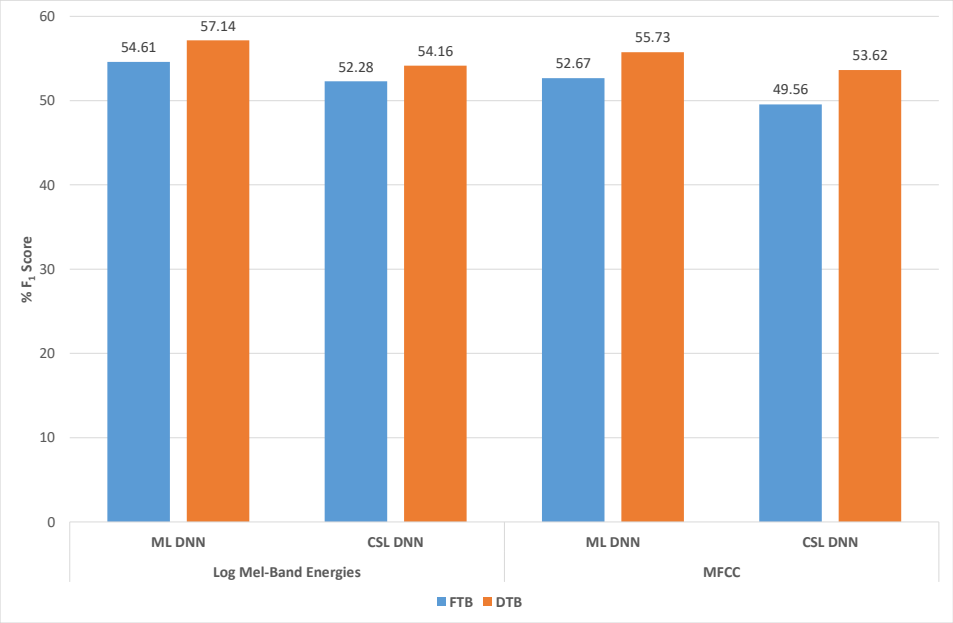

Figure 5.1: Comparison of the binarization methods

Feature Set	DNN Architecture	F_1 Score % Before Post-processing	F_1 Score % After Post-processing
Log Mel-band Energies	ML DNN	57.14	62.50
	CSL DNN	54.16	58.12
MFCCs	ML DNN	55.73	59.53
	CSL DNN	53.62	57.84

Table 5.2: Comparison of the results before and after the post-processing

For all the above experiments, the train set was used with the augmented data. Fold 5 and 6 of the original dataset were replaced by the augmented data which were prepared using the same folds. To evaluate the effect of the data augmentation on the classification results, the best-performed architecture was trained using the original test set of the dataset. ML DNN was used with the same hyperparameters for the experiment. The model trained with only the

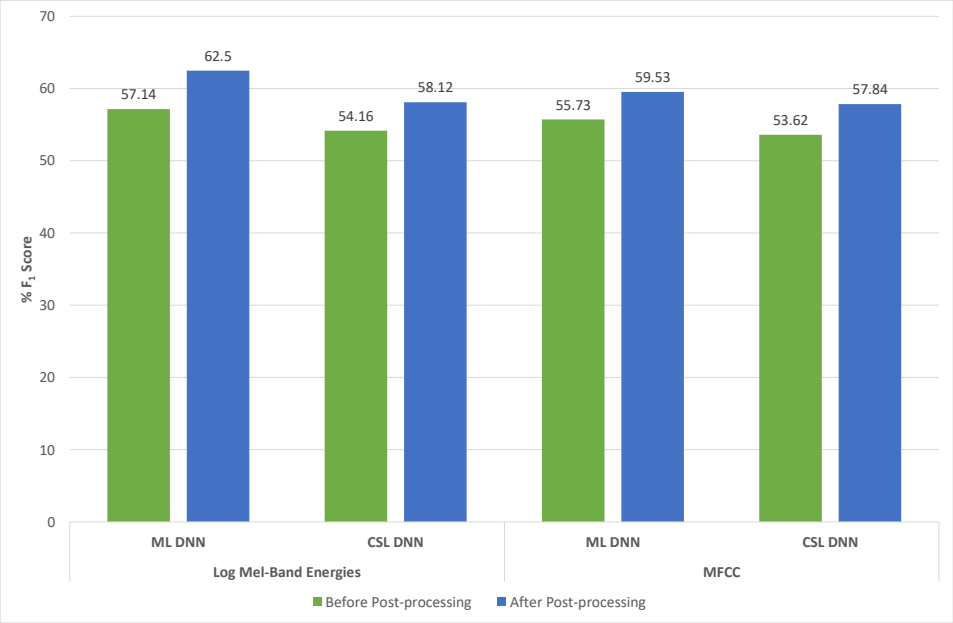

Figure 5.2: Comparison of F_1 score before and after the post-processing

original data recorded 60.29% F1 score which is a nearly 2% drop compared to the augmented data.

As mentioned in section 4.3.1, grid search was used to select the best hyper-parameter of the DNN. Table 5.3 summarizes the results of the grid search that have been used as the final values for the different models.

Class-wise detection accuracies were calculated for the best performed model which is the ML DNN with Log Mel-band energies. Figure 5.3 shows the comparison of the detection accuracy of each class.

5.3 Discussion

When we compare the overall results of the different experiments that are presented in the table 5.1, 5.2 and graphs in figure 5.1, 5.2 it emphasizes that Log Mel-band Energy features are more effective than the MFCCs for polyphonic AED. Selecting the prominent coefficients after the Discrete Cosine Transform that involves in the MFCC results in a loss of information. That might be the

DNN Model	Parameters			
	Learning Rate	# Hidden Layers	# Hidden Units (in each layer)	Initial Weight Range
ML DNN	0.02	3	800	± 0.002
CSL DNN - children playing	0.04	3	600	± 0.002
CSL DNN - air conditioner	0.02	3	600	± 0.004
CSL DNN - engine idling	0.02	3	600	± 0.002
CSL DNN - car horn	0.04	2	800	± 0.004
CSL DNN - gunshot	0.04	2	600	± 0.002
CSL DNN - dog bark	0.02	2	800	± 0.002
CSL DNN - siren	0.02	2	800	± 0.002
CSL DNN - drilling	0.04	2	800	± 0.002
CSL DNN - street music	0.02	3	600	± 0.004
CSL DNN - jackhammer	0.02	2	600	± 0.002

Table 5.3: Final hyperparameter values based on the grid search results

reason for the lower accuracy compared to the Log Mel-band Energies.

In all the experiments, ML DNN outperformed the CSL DNN in a significant amount. Decomposing a multi-label classification problem into a set of single label

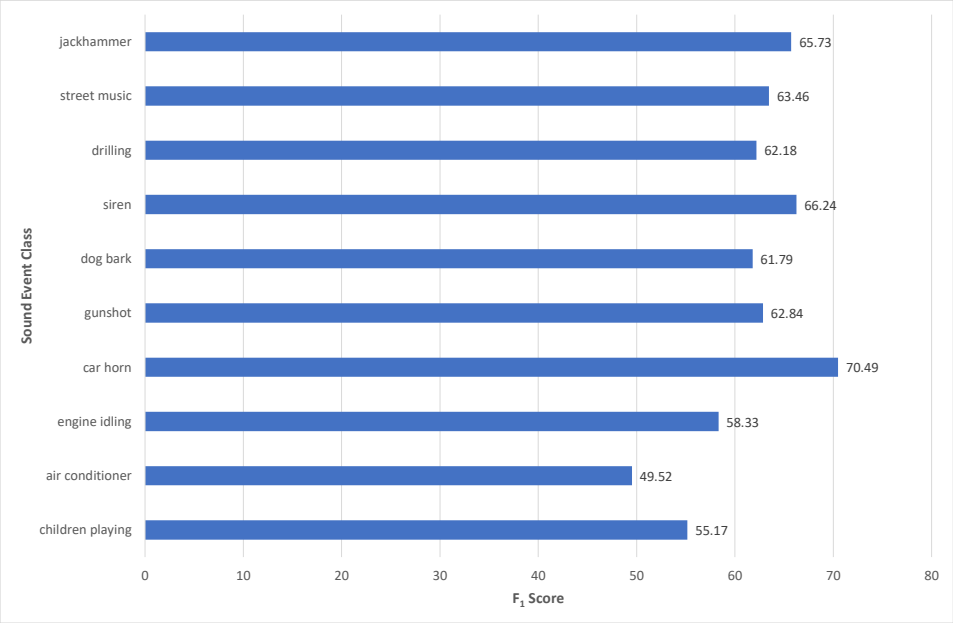

Figure 5.3: Comparison of class-wise F_1 score

classification problems may cause a loss of correlation patterns between labels. This might be the reason for the lower accuracy of the CSL DNN compared to the ML DNN. However, when we consider the advantages of the CSL DNN, like introducing a new class or taking a subset of the classes, this accuracy loss might be negligible based on the context.

When comparing the performance of the binarization methods, DTB outperformed FTB in all experiments. 4.1% which is the highest increase of the accuracy compared to the FTB was given for the CSL DNN which was trained using MFCCs. But the other three experimental setups have shown only a 1.9% - 3% accuracy increase for DTB. When the polyphony level increases at a particular frame, the output probabilities of the sigmoid activation functions are getting lower. Therefore, the threshold value should also go lower. As explained in the section 4.4.2, DTB was proposed based on the assumption where the energy of the signal depends on the polyphony level. Since DTB increased the detection accuracy for all NN architectures and feature sets, we can say that the assump-

tion is correct. However, tryout more complex threshold calculation methods will further improve the accuracy.

Comparison in the table 5.2 and the graph in figure 5.2 emphasize the effectiveness of the proposed post-processing method. It increased the accuracy of all the experimented architectures methods by 3.8% - 5.4% resulting in the highest values as 62.50%. When we observe the accuracy improvement of both ML DNN and CSL DNN, we can see a clear improvement in ML DNN results. Because of that, we can decide that the number of noisy classifications (ie. very short term false positives and false negatives) are higher in ML DNN compared to CSL DNN.

We can decide that the data augmentation helped to improve the accuracy since the model which was trained with the original 6 folds of the training set could only reach 60.29% in the F1 score. As explained in section 3.4, there could be slight variations and deformations even in the sound events that are emitted from the same source. Implemented data augmentation methods made a more robust training dataset resulting better results in the detection as well.

When comparing the obtained results with the baseline work, this research has reported a slight improvement in accuracy (3.7%). However, when we compare the complexity of the classifier architectures, this work has proposed simpler architecture with deep feedforward neural network compared to the CRNN architecture proposed in the baseline study. Implementation of data augmentation, dynamic threshold binarization, and post-processing increased the performance of the DNN classifier even it is not complex as the baseline work. Trying out more complex classifier architectures with the proposed pre-processing, binarization, and post-processing methods may increase the accuracy further.

5.4 Summary

This chapter presented the results of the experiments carried out. Different comparisons were done to compare the feature sets, NN architectures, binarization methods, and those results are presented in the chapter. Then the effect of the

post-processing and data augmentation was also presented with the experiment done using the best-performed binarization method. According to the experimental results, log Mel-band energies outperformed the MFCC feature set in all the setups. Slight accuracy improvement was given by the DTB for all experiments while showing the highest 4.1% improvement for CSL DNN with MFCC. Median filter based post-processing performed well in smoothing the detection results. Finally, the augmented data in the training dataset helped to train a robust while improving the accuracy nearly by 2%. The best-performed setup reached a 62.5% F_1 score which is a good value in the domain of polyphonic AED for the considered dataset.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This study addressed the acoustic event identification in polyphonic environments which is a complex area of study in the automatic sound recognition and identification domain. When we analyze the previous literature, we can see that this research area is still to be improved since none of them exceeded the 75% accuracy. Creating a robust AED requires a significant amount of data, processing, and also advance machine learning techniques.

This thesis provides a detailed explanation of two different implementations of the artificial neural networks for the above task with the use of two different feature sets. And also the methods that can be used to further improve the accuracy are evaluated.

Based on the findings of the research, ML DNN performed well compared to the CSL DNN. Also, it is easier to implement ML DNN since we need to train only one classifier where we have to train multiple classifiers in the CSL DNN approach. But if we can do more experiments to fine-tune the models in the CSL DNN separately, we might gain better results.

The Log Mel-band Energy feature set has given better results compared to the MFCC feature set for all NN architectures. Possible reason for this would be selection of only the prominent coefficients after the discrete cosine transformation in the MFCC calculation. However, Log Mel-band Energies might also not the best feature combination for the task.

When comparing the experimental results of the FTB and DTB methods, DTB increased the detection accuracy. However, it was not to the expected level for three experimental setups except CSL DNN with MFCCs. The assumption I made to come up with the DTB equation seems correct but not strong enough

to calculate a robust threshold value for all architectures.

The median filter based post-processing method has performed well to filter the noisy classification results. Although the accuracy of the CSL DNN is lower than the ML DNN, noisy classifications such as sudden positives or sudden negatives are lower in the CSL DNN. That is the reason where accuracy improvement after the median filter is lower for the CSL DNN.

In summary, this research did a comparison and a detailed evaluation of different approaches that can be applied to polyphonic acoustic event identification. A novel dynamic threshold calculation method for binarization was also proposed and evaluated. And finally, the median filter based post-processing method was also implemented to further increase the accuracy. Based on the experimental results, a 62.5% F1 score which is the highest value was achieved by the multi-label deep neural network with log Mel-band energy feature set, dynamic threshold binarization, and median filter-based post-processing. The recorded value is a 3.7% increase compared to the baseline study mentioned in section 3.1.

In conclusion the proposed system has performed well (62.5%) compared to state-of-the-art systems (58.8%) with additional implementation of data augmentation, dynamic threshold binarization, and median filtering based post-processing.

6.2 Future Work

This study can be extended in different directions in the future. One possible area is implementing more complex neural network architectures such as CNN, RNN with the proposed DTB method and evaluate the results. Although several research works have been performed with these architectures, they have used basic binarization methods. Therefore, trying out DTB with those architectures will increase the accuracy further.

The proposed DTB method increased the accuracy of all experiments. However, I suggest to try out better binarizing methods as future work. Since there is a clear dependency between the polyphony level and the threshold value, a more

accurate polyphony level estimation method can be tried to increase the DTB method. Polyphony level estimation would be an interesting research area that will increase the performance of polyphonic AED applications.

References

[1] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertilä, Toni Heittola, and Tuomas Virtanen. Sound event detection in multichannel audio using spatial and harmonic features. *arXiv preprint arXiv:1706.02293*, 2017.

[2] Emmanouil Benetos, Grégoire Lafay, Mathieu Lagrange, and Mark D Plumbley. Detection of overlapping acoustic events using a temporally-constrained probabilistic model. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6450–6454. IEEE, 2016.

[3] Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303, 2017.

[4] Woohyun Choi, Jinsang Rho, David K Han, and Hanseok Ko. Selective background adaptation based abnormal acoustic event recognition for audio surveillance. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 118–123. IEEE, 2012.

[5] Qi Li, Huadong Ma, and Dong Zhao. A neural network based framework for audio scene analysis in audio sensor networks. In *Pacific-Rim Conference on Multimedia*, pages 480–490. Springer, 2009.

[6] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In *Hellenic conference on artificial intelligence*, pages 91–100. Springer, 2010.

[7] Asma Rabaoui, Zied Lachiri, and Noureddine Ellouze. Using hmm-based classifier adapted to background noises with improved sounds features for audio surveillance application. *Int. J. Signal Process*, 3:535–545, 2009.

[8] Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis. On acoustic surveillance of hazardous situations. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 165–168. IEEE, 2009.

[9] Burak Uzkent, Buket D Barkana, and Hakan Cevikalp. Non-speech environmental sound classification using svms with a new set of features. *International Journal of Innovative Computing, Information and Control*, 8(5):3511–3524, 2012.

[10] Xiaodan Zhuang, Xi Zhou, Mark A Hasegawa-Johnson, and Thomas S Huang. Real-world acoustic event detection. *Pattern Recognition Letters*, 31(12):1543–1551, 2010.

[11] Sharath Adavanne, Pasi Pertilä, and Tuomas Virtanen. Sound event detection using spatial features and convolutional recurrent neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 771–775. IEEE, 2017.

[12] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. Acoustic scene classification with matrix factorization for unsupervised feature learning. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6445–6449. IEEE, 2016.

[13] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Audio surveillance using a bag of aural words classifier. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 81–86. IEEE, 2013.

[14] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1458–1465. IEEE, 2005.

[15] Anurag Kumar, Pranay Dighe, Rita Singh, Sourish Chaudhuri, and Bhiksha Raj. Audio event detection from acoustic unit occurrence patterns. In

2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 489–492. IEEE, 2012.

[16] Donatello Conte, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, and Mario Vento. An ensemble of rejecting classifiers for anomaly detection of audio events. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 76–81. IEEE, 2012.

[17] Michele Lai Chin and Juan José Burred. Audio event detection based on layered symbolic sequence representations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1953–1956. IEEE, 2012.

[18] Sam T Roweis. One microphone source separation. In *Advances in neural information processing systems*, pages 793–799, 2001.

[19] Laurent Benaroya, Frédéric Bimbot, and Rémi Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):191–199, 2005.

[20] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[21] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

[22] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z Fern, Raviv Raich, Sarah JK Hadley, Adam S Hadley, and Matthew G Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012.

[23] Luigi Gerosa, Giuseppe Valenzise, Marco Tagliasacchi, Fabio Antonacci, and Augusto Sarti. Scream and gunshot detection in noisy environments. In *2007 15th European Signal Processing Conference*, pages 1216–1220. IEEE, 2007.

[24] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE, 2016.

[25] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. Deep convolutional neural networks and data augmentation for acoustic event detection. *arXiv preprint arXiv:1604.07160*, 2016.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[28] Huy Phan, Marco Maaß, Radoslaw Mazur, and Alfred Mertins. Random regression forests for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):20–31, 2014.

[29] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[30] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348. IEEE, 2017.

[31] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. Chime-home: A dataset for sound source recognition in a domestic

environment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2015.

[32] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. Polyphonic sound event detection using multi label deep neural networks. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2015.

[33] M. Cartwright P. Li J. Salamon, D. MacConnell and J. P. Bello. The URBAN-SED Dataset, year = 2017, url = http://urbansed.weebly.com.

[34] Elmar Messner, Matthias Zöhrer, and Franz Pernkopf. Heart sound segmentation—an event detection approach using deep recurrent neural networks. *IEEE transactions on biomedical engineering*, 65(9):1964–1974, 2018.

[35] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[36] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126, 2015.

[37] Wenbo Wang, Sichun Li, Jianshe Yang, Zhao Liu, and Weicun Zhou. Feature extraction of underwater target in auditory sensation area based on mfcc. In *2016 IEEE/OES China Ocean Acoustics (COA)*, pages 1–6. IEEE, 2016.

[38] Stanley A Gelfand. *Hearing: An introduction to psychological and physiological acoustics*. CRC Press, 2016.

[39] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[40] Trevor R Agus, Simon J Thorpe, Clara Suied, and Daniel Pressnitzer. Characteristics of human voice processing. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 509–512. IEEE, 2010.

[41] Petros S Karvelis, Dimitrios I Fotiadis, Dimitrios G Tsalikakis, and Ioannis A Georgiou. Enhancement of multichannel chromosome classification using a region-based classifier and vector median filtering. *IEEE Transactions on Information Technology in Biomedicine*, 13(4):561–570, 2009.

[42] S Esakkirajan, T Veerakumar, Adabala N Subramanyam, and CH PremChand. Removal of high density salt and pepper noise through modified decision based unsymmetric trimmed median filter. *IEEE Signal processing letters*, 18(5):287–290, 2011.