

**ACOUSTIC EVENT DETECTION IN POLYPHONIC  
ENVIRONMENTS USING ARTIFICIAL NEURAL  
NETWORKS**

Jayawardhana Pathirana 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

|            |  |    |
|------------|--|----|
| Figure 1.1 | Representation of acoustic events in a monophonic environment  | 3  |
| Figure 1.2 | Representation of acoustic events in a polyphonic environment  | 4  |
| Figure 2.1 | Mel scale  | 12 |
| Figure 2.2 | Triangular Mel-filterbank $w_{k,h}$  | 13 |
| Figure 3.1 | Overview of the methodology  | 31 |
| Figure 3.2 | Difference between ML-DNN and CSL-DNN  | 36 |
| Figure 4.1 | Experimental design  | 41 |
| Figure 4.2 | Waveforms of the original soundscape (top), stretch factor = 0.8 (middle) and stretch factor = 1.2 (bottom)              | 42 |
| Figure 4.3 | Spectrograms of the original soundscape (top), shift down by 2 semi-tones (middle) and shift up by 2 semi-tones (bottom) | 44 |
| Figure 4.4 | Representation of MFCCs  | 46 |
| Figure 4.5 | Representation of log Mel-band energies  | 46 |
| Figure 4.6 | RMSE variation of a signal (top) and dynamic threshold value calculated by the algorithm (bottom)                        | 51 |
| Figure 4.7 | Sample representation of median filtering based post-processing  | 53 |
| Figure 5.1 | Comparison of the binarization methods   | 56 |
| Figure 5.2 | Comparison of $F_1$ score before and after the post-processing   | 57 |
| Figure 5.3 | Comparison of class-wise $F_1$ score   | 59 |

## LIST OF TABLES

|           |  |    |
|-----------|--|----|
| Table 5.1 | Comparison of the binarization methods                         | 55 |
| Table 5.2 | Comparison of the results before and after the post-processing | 56 |
| Table 5.3 | Final hyperparameter values based on the grid search results   | 58 |

## 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

|  |      |
|--|------|
| Declaration  | i    |
| Acknowledgement  | ii   |
| Abstract   | iii  |
| List of Figures  | iv   |
| List of Tables   | v    |
| List of Abbreviations                                  | vi   |
| Table of Contents                                      | viii |
| 1 Introduction   | 1    |
| 1.1 Acoustic Event Detection and Analysis              | 1    |
| 1.1.1 Monophonic acoustic event detection and analysis | 2    |
| 1.1.2 Polyphonic acoustic event detection and analysis | 2    |
| 1.2 Motivation   | 4    |
| 1.3 Problem  | 5    |
| 1.4 Research Objectives                                | 5    |
| 1.5 Thesis Outline                                     | 6    |
| 2 Literature Review                                    | 7    |
| 2.1 Feature Extraction                                 | 7    |
| 2.1.1 Temporal Features                                | 8    |
| 2.1.2 Spectral Features                                | 9    |
| 2.1.3 Time-frequency-based Features                    | 10   |
| 2.1.4 Cepstrum-based Features                          | 11   |
| 2.1.5 Energy-based Features                            | 14   |
| 2.1.6 Biologically/ Perceptually Driven Features       | 15   |
| 2.1.7 Usage of the Features                            | 17   |
| 2.2 Detection and Classification                       | 20   |
| 2.2.1 Generative Classifiers                           | 20   |
| 2.2.2 Discriminative Classifiers                       | 21   |
| 2.2.3 Hybrid Classifiers                               | 21   |

|       |                                    |    |
|-------|------------------------------------|----|
| 2.2.4 | Usage of the Different Classifiers | 22 |
| 2.3   | Datasets                           | 26 |
| 2.3.1 | TUT-SED Synthetic 2016             | 26 |
| 2.3.2 | UrbanSound8k                       | 27 |
| 2.3.3 | The URBAN-SED                      | 27 |
| 2.3.4 | CHiME-Home                         | 28 |
| 2.4   | Summary                            | 29 |
| 3     | Methodology                        | 30 |
| 3.1   | Baseline Study                     | 30 |
| 3.2   | Formation of the Methodology       | 32 |
| 3.3   | Selecting the Dataset              | 32 |
| 3.4   | Data Augmentation                  | 33 |
| 3.5   | Feature Extraction                 | 34 |
| 3.6   | Classification                     | 36 |
| 3.7   | Binarization                       | 37 |
| 3.8   | Post-processing                    | 38 |
| 3.9   | Summary                            | 38 |
| 4     | System and Experimental Design     | 40 |
| 4.1   | Data Augmentation                  | 40 |
| 4.1.1 | Time Stretching                    | 40 |
| 4.1.2 | Pitch Shifting                     | 43 |
| 4.2   | Feature Extraction                 | 43 |
| 4.2.1 | MFCCs Extraction                   | 45 |
| 4.2.2 | Log Mel-band Energies Extraction   | 46 |
| 4.2.3 | Context Window                     | 46 |
| 4.3   | Classification                     | 47 |
| 4.3.1 | Classifiers                        | 47 |
| 4.4   | Binarization                       | 50 |
| 4.4.1 | Fixed Threshold Binarization       | 50 |
| 4.4.2 | Dynamic Threshold Binarization     | 50 |
| 4.5   | Post-processing                    | 52 |

|       |                                     |    |
|-------|-------------------------------------|----|
| 4.5.1 | Median Filter Based Post Processing | 52 |
| 4.6   | Summary                             | 53 |
| 5     | Results and Discussion              | 54 |
| 5.1   | Evaluation Metrics                  | 54 |
| 5.2   | Results                             | 55 |
| 5.3   | Discussion                          | 57 |
| 5.4   | Summary                             | 60 |
| 6     | Conclusion and Future Work          | 62 |
| 6.1   | Conclusion                          | 62 |
| 6.2   | Future Work                         | 63 |
|       | References                          | 65 |