FEATURE EXTRACTION DESIGN FOR EMBEDDED NEURAL NETWORK URBAN SOUND CLASSIFIER

LIM CHIN SHEN

A project report submitted in partial fulfilment of the
Requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic System)

School of Electrical Engineering
Faculty of Engineering
Universiti Teknologi Malaysia

JULY 2021

DEDICATION

This project report is dedicated to my parents, who taught me the importance of education. I also dedicate this research to my manager, who encouraged me to embark on this journey of knowledge.

ACKNOWLEDGEMENT

I would like to express my gratitude to my family and friends, who has always been there for me and giving me the much-needed moral support.

I would also like to thank Dr Shahidatul, my supervisor for always giving me advice and guidance throughout the project, which helped me to understand and get into the field of study of machine learning and audio processing.

ABSTRACT

Urban sound research has become a hot topic in recent years for city growth observation and surveillance application through noise source identification. However, the sound identification is challenging due to the multiple sound sources that are blended. There are also new sounds that are unclassified by recent studies as the region of the city becomes more developed. In recent work of audio classification, the features of sound are extracted by its image which is obtained from the pattern of time-frequency representation or otherwise known as spectrogram. This project aims to design a noise robust, neural network urban sounds classifier that is implemented on an embedded system. Two feature extractors that converts audio to image will be explored and compared to produce better features for urban sound. Mel Frequency Cepstral Coefficient (MFCC) is commonly used throughout all sound classifiers with good results while Gammatone Frequency Cepstral Coefficient (GFCC) is an emerging feature extractor said to be better at extracting noisy data. Urbansound8k, which contains 8732 labelled sound classified into eight classes, is used as the dataset. Different decibels of noise were added to the dataset to simulate the actual urban sound scenario and to explore the noise robustness of the two feature extractors. To classify urban sound, the audio is converted into an image. Therefore, Convolutional Neural Network (CNN) model is employed because it is one of the best machine learning models for image. Since the design are focusing on embedded system application, lightweight CNN model MobileNetV2 will be used in this project. The feature extractor and the neural network model will be developed using a python language and TensorFlow library. The experimental result shows that MFCC outperforms GFCC in terms of classification accuracy by an average of 14.34% across all SNR levels. MFCC is also more robust to noise in dataset, with 2.75% and 2.87% drop in accuracy at 30dB and 10dB noise signal respectively compared to baseline of noiseless signal, whereas GFCC has a drop of 6.18% and 3.87% at 30dB and 10dB noise signal respectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	-	Convolutional Neural Network
MFCC	-	Mel Frequency Cepstral Coefficient
GFCC	-	Gammatone Frequency Cepstral Coefficient
IoT	-	Internet of Things
SNR	-	Signal to Noise Ratio
RMS	-	Root Mean Square
WGN	-	White Gaussian Noise
STFT	-	Short Time Fourier Transform
DFT	-	Discrete Fourier Transform
DCT	-	Discrete Cosine Transform

LIST OF SYMBOLS

δ	-	Minimal error
D, d	-	Diameter
F	-	Force
v	-	Velocity
p	-	Pressure
I	-	Moment of Inersia
r	-	Radius
Re	-	Reynold Number

CHAPTER 1

INTRODUCTION

1.1 Problem Background

As a city becomes more bustling and developed, the increase in undesirable urban noise is often inevitable. Some examples of urban noise come from car engine, construction sites, people shouting and so on. Exposing to high level of urban noise leaves the city's netizen susceptible to several health issues such as impaired sleep, depression or heart attack [1]. To battle that, urban sound classification surges in importance to help identify the problematic noise source for reduction or elimination.

One of the most challenging aspect of urban sound classification is the background noise that exists in the signal due to the city's environment. Previous studies in [2] [3] have found that image is the best performing input feature when it comes to sound recognition in noisy environment. Mel Frequency Cepstral Coefficient (MFCC) is an image-like sound feature commonly used in sound recognition and classification. However, MFCC is very susceptible to noise, which quickly degrades the performance of the classification model [4]. Gammatone Frequency Cepstral Coefficient (GFCC) on the other hand is an underappreciated sound feature that shows promising classification result [5] [6].

Urban sound classification using supervised machine learning has been extensively studied in the past [7]. With majority of urban sound applications being in various parts of the city, IoT solutions and embedded system implementations such as audio-based surveillance systems are seeing higher usage [1] [8]. Therefore, hardware limitations in the form of device size and computational power require us to select machine learning models that are lightweight in nature. MobileNetV2 is one such model aiming to deliver high accuracy while keeping a relatively low number of parameters and mathematical operations.

In this project, different levels of noise are added to the dataset to mimic the noise environment in a city. The feature extraction methods based on Mel filter and Gammatone filter are introduced, where audio is converted to an image of time-frequency representation. Then, the sound features are extracted from the dataset and preprocessed to have uniform dimensions. Both features are fed to the MobileNetV2 machine learning model and trained with an optimized set of parameters.

1.2 Problem Statement

The identification of urban sound is challenging due to the multiple sound sources that are blended. Most researches on urban sound classifiers are using a clean dataset, which does not reflect the actual urban sound with background noises. Moreover, the noise robustness of different feature extractors to urban sound has not been extensively studied.

1.3 Research Questions

There are two research questions identified for this study:

Q1. Which feature extractor is better for urban sound?
Q2. Which feature extractor is more robust to noise?
Q3. How is the performance of MobileNetV2 as a low-complexity sound classifier compared to heavyweight machine learning models?

1.4 Research Objectives

The objectives of the research are:

1. To explore and compare different feature extractors that converts audio to image to produce better performance results

2. To add noise to the urban sound dataset and explore the noise robustness of different sound features
3. To simulate an embedded system implementation through lightweight machine learning model

1.5 Project Scope

The project focuses on building a sound identification system using MFCC and GFCC feature extraction for lightweight CNN learning architecture. The development is done in Python and Keras-TensorFlow, which are employed as deep learning framework. The dataset used to train our model is UrbanSound8K [9], a popular dataset used in many other urban sound researches. It consists of 8732 labelled sounds coming from 10 different classes. White gaussian noise (WGN) with different signal-to-noise ratio are added to the dataset to simulate noise environment. The features of the modified dataset extracted are MFCCs and GFCCs, which are used as the input to the CNN model. The neural network implementation is focusing on embedded system and the targeting CNN model is MobileNetV2, which is a lightweight CNN.

REFERENCES

[1] Y. Alsouda, S. Pllana and A. Kurti, "A machine learning driven IoT solution for noise classification in smart cities," *arXiv preprint arXiv:1809.00238,* 2018.

[2] I. McLoughlin, H. Zhang, Z. Xie, Y. Song and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, p. 540–552, 2015.

[3] H. Zhang, I. McLoughlin and Y. Song, "Robust sound event recognition using convolutional neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP),* 2015.

[4] Z. Wu and Z. Cao, "Improved MFCC-Based Feature for Robust Speaker Identification," *Tsinghua Science and Technology,* vol. 2, 2005.

[5] G. K. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *arXiv preprint arXiv:1806.09010,* 2018.

[6] W. Zhang, Y. Wu, D. Wang, Y. Wang, Y. Wang and L. Zhang, "Underwater Target Feature Extraction and Classification Based on Gammatone Filter and Machine Learning," in *2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR),* 2018.

[7] B. da Silva, A. W Happi, A. Braeken and A. Touhafi, "Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems," *Applied Sciences,* vol. 9, p. 3885, 2019.

[8] A. Zanella, N. Bui, A. Castellani, L. Vangelista and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things journal,* vol. 1, p. 22–32, 2014.

[9] J. Salamon, C. Jacoby and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[10] V. Valerio, *Mel-Frequency Cepstral Coefficients Explained Easily,* 2020.

[11] Z. Huang, C. Liu, H. Fei, W. Li, J. Yu and Y. Cao, "Urban sound classification based on 2-order dense convolutional network using dual features," *Applied Acoustics*, vol. 164, p. 107243, 2020.

[12] D. Bardou, K. Zhang and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial intelligence in medicine*, vol. 88, p. 58–69, 2018.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[14] M. Jeevan, A. Dhingra, M. Hanmandlu and B. K. Panigrahi, "Robust speaker verification using GFCC based i-vectors," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, 2017.