SPEAKER IDENTIFICATION:

TIME-FREQUENCY ANALYSIS WITH DEEP LEARNING

A thesis

Presented to

Department of Computer Science

Tennessee State University

Nashville, TN

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Hui Chen

September 2018

Keywords: deep learning, subspace separation, deep convolutional networks,

To the Graduate School,

We are submitting a thesis by ”Hui Chen” entitled ”Speaker Identification: Time-Frequency Analysis with Deep Learning. We recommend that it be accepted in partial fulfillment of the requirements for the degree, Master of Science in Computer Science.

Dr. Ali Sekmen

Chairperson

Dr. Tamara Rogers

Committee Member

Dr. Erdem Erdemir

Committee Member

Accepted for the Graduate School:

Robbie K. Melton, Ph.D.

Dean of the Graduate School

ABSTRACT

Speaker identification with deep learning commonly use time-frequency representation of the voice signals. This research experiments with spectrogram based, Mel-Frequency Cepstral Coefficients (MFCCs) training on different Neural Networks (NNs) Topologies. The NNs ability to separating human voice biometrics features for identifying speakers. MFCCs are commonly used as feature extractor and combines with a Neural Networks (NNs) in speech recognition systems. This research shows that MFCCs with Convolutional Neural Networks (CNNs) shown a better accuracy for identifying speakers, comparing to other NNs topologies.

This research also proposes a network for speaker identification, combining Wigner Ville Distribution (WVD) with deep learning. WVD has been used for time-frequency (TF) transformation and successfully implemented for other sound identifying tasks, and its representations are known which have a better resolution of properties. In this research, instead of directly extracting features through MFCCs, WVD is implemented with CNNs together as feature extraction network, and trained on the dataset. Even though the result is inconclusive, it still provided many useful insights of the approach.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Ali Sekmen for guiding and supporting. Also, I would like to thank my thesis committee members Dr. Erdem Erdemir and Dr. Tamara Rogers, for providing your invaluable input; and thank you to fellow graduate students, Ayad Abdul-Malek for helping with instruments and thank Mustafa Parlaktuna and as my research collaborator. I am very grateful to all of you. Finally, I would like to thank my amazing family for love, support, and encouragement through out the years. Especially my parents, I undoubtedly could not have done this without you. Thank you!

This research is supported by DoD Grant W911NF-15-1-0495. I would like to acknowledge the support from the Army Research Lab (ARL) and the Air Force Research Lab (AFRL).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF EQUATIONS

CHAPTER I

INTRODUCTION

Research in speaker identification has been studied in experimenting with statistical models. In recent years, as High Performance Computing (HPC) and deep learning frameworks becoming more accessible, speaker identification problem have been researched with the availability of the GPUs and deep learning. The first phase of this research uses Mel-Frequency Cepstral Coefficients (MFCCs) as feature extractor, combines with existing sound identification deep learning architectures, and trained on speaker dataset to identify speakers. Comparing the accuracy of different deep learning architectures, and identify the best architecture for performing speaker identification task. The second phase of this research combines time-frequency (TF) transformation method Wigner Ville distribution (WVD), with the best performing deep learning architecture, which is identified in the first phase of the research. WVD and deep learning are proposed for direct feature extraction of the voice data.

I.1 Problem Statement

Speaker identification with deep learning has been a field thats has attracted less attention of researchers, and this research is focused on experimenting human voice identification through deep neural networks. The first phase to examine whether deep neural network is able to identify speaker based on their utterances, and determine the best topology and settings for speaker identification. The second phase closely assesses the process of conventional speaker identification tasks, and experiments with alternative approaches. Almost all speech tasks uses cepstrum, a sequence of numbers that characterize a frame of speech. Cepstrum is usually combined with Hidden Markov Model or Neural Network for automatic speech and speaker recognition. Although MFCCs extracted features have proven to be useful and are widely used to characterize the features of speech, the reasoning behind why the spectrogram feature extracting steps works well remains unclear. While CNNs are able to calculate and define a set of features throughout each layer, therefore utilizes WVD to replace MFCCs as a more direct speech feature extractor, and with CNNs is experimented at the second phase of this research.

I.2 Goals and Objectives

I.2.1 Goal-1: Dataset

Deep learning relies heavily on a large dataset with good labeling. For different deep learning tasks, there are many benchmark datasets available; for nature language modeling there are datasets including WordNet[7], Imdb Reviews[8], Yelp reviews[9] and the Wikipedia Corpus[10]; for image recognition, there are MNIST[11], ImageNet[12], CIFAR[13]; for voice recognition, there are WaveNet[14], Urban Sound Classification[15] and Bird Sounds[16]; that have been the standard database for these deep learning research areas. However there is no currently publicly available speaker identification dataset available, generating a good dataset thats suitable for such task is the critical first step for the research.

Objectives:

1. Creating a good dataset of clean data with good recording quality.

2. Easy to expand dataset with sufficient amount of utterances per speaker.