

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,

ProQuest Number: 13419654

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13419654

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

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	vii
List of Tables	viii
I Introduction	1
I.1 Problem Statement	1
I.2 Goals and Objectives	2
I.2.1 Goal-1: Dataset	2
I.2.2 Goal-2: Deep Learning	3
I.2.3 Goal-3: Wigner Ville Distribution	3
I.3 Literature Review	4
I.4 Content Organization	5
II Speech Properties	7
II.1 Human Voice	7
II.2 Audio Recording and Data Storage Format	7
II.2.1 WAV Format	8
II.2.2 MP3 Format	9
II.3 Speaker Identification Methods	9
II.3.1 Nearest Neighbor	10
II.3.2 Support Vector Machines (SVMs) Classifier	10
II.3.3 Hidden Markov Models (HMM)	11
II.3.4 Neural Nets (NNs)	12
II.3.5 Decision Tree	12
II.4 Speaker Identification Application	12
III Time-Frequency and Feature Extraction Methods	14
III.1 Linear Prediction Cepstral Coefficients (LPCCs)	14
III.2 Mel-Frequency Cepstral Coefficients (MFCCs)	14
III.2.1 Pre-emphasis	15
III.2.2 Framing and Windowing	16
III.2.3 Short Term Fourier Transform (STFT) and Power Spectrum	17
III.2.4 Filter Banks	17
III.2.5 Normalization	19
III.3 Wigner Ville Distribution (WVD)	20
III.3.1 Auxiliary Function	20
III.3.2 Fourier Transform	20
III.3.3 Wigner Ville Distribution (WVD) Function	20
III.3.4 Modified Wigner distribution function	21
III.3.5 Fractional Fourier Transform (FrFT) and Rotation of WVD	21

III.4	Methods Analysis	22
III.4.1	Runtime	22
III.4.2	Leakage and Cross Term	22
IV	Deep Learning	24
IV.1	Development of Deep Learning	24
IV.2	Back-propagation	26
IV.3	Convolutional Neural Network	28
IV.4	Recurrent Neural Network	29
V	Speaker Identification Research	31
V.1	Dataset	31
V.2	Preliminary Study	32
V.2.1	Single Fully Connected Neural Network	32
V.2.2	Convolutional Neural Network	32
V.2.3	Low Latency Convolutional Neural Network	33
V.2.4	Low Latency SVDF	34
V.2.5	WVD on Small Audio Segments	35
V.3	Phase1: Deep Learning Improvement	35
V.4	Phase2: WVD with Deep Learning	36
V.4.1	WVD running time analysis	36
V.4.2	WVD with Down-sampling	38
V.4.3	Convolutional Neural Network Topology	39
V.5	High Performance Computing (HPC)	39
V.5.1	Data Pre-process with Multi-core CPUs Nodes	39
V.5.2	Tensorflow on GPU	39
VI	Research Results	41
VI.1	Preliminary: Comparing Deep Learning Architectures	41
VI.2	Results of Phase 1: Deep Learning Improvements	41
VI.3	Results of Phase 2: WVD with Deep Learning	42
VII	Conclusion	43
VII.1	Possible Improvements	44
VII.2	Future Direction	44
VIII	Bibliography	46

LIST OF FIGURES

II.1	WAV Audio File Format [1]	8
II.2	Hidden Markov Model [2]	11
III.1	Comparing Signals, before and after signal emphasis [3]	16
III.2	Filter bank on a Mel-Scale [3]	17
III.3	Spectrogram and MFCCs [3]	19
III.4	Normalized Spectrogram and MFCCs [3]	19
III.5	Utterances converted to TF domain through WVD	21
IV.1	Time Line: Artificial Intelligence, Machine Learning and Deep Learning [4] . . .	25
IV.2	How computer perceive a cat [5]	28
IV.3	Architecture of a CNN [5]	29
IV.4	FNNs vs. RNNs [6]	30
V.1	Architecture of Single FC: Basic Neural Network	32
V.2	Architecture of Convolutional Neural Network	33
V.3	Architecture of Low Latency Convolutional Neural Network	34
V.4	Architecture of Low Latency SVDF	35
V.5	Wigner Ville (left) vs. Wigner Ville with window (right)	37
V.6	Down-Sampled, Filtered and Resized WVD 2D Image	38
VII.1	Tesla K20m GPU, Memory utilization during training	43

LIST OF TABLES

III.1	TF Runtime Comparison with 100 repeats	22
V.1	TF Functions Computational Time Comparison, (minutes)	36
V.2	Nvidia Tesla K20m GPU Information	40
VI.1	Speaker Identification with Deep Learning	41

PREVIEW

LIST OF ALGORITHMS

1	Pre-Emphasis Algorithm	15
2	Window Function with WVD	37

PREVIEW

LIST OF EQUATIONS

III.1	Signal Pre-Emphasis Filter	15
III.2	Hamming Windows Function	17
III.3	Short Term Fourier Transform (STFT)	17
III.4	Power Spectrum	17
III.5	Mel-Scale on Power Spectrum	18
III.6	Frequency Bands	18
III.7	Filter Bank	18
III.8	Auxiliary Function (AF)	20
III.9	Fourier Transform (FT)	20
III.10	Wigner Ville Distribution Function (WVD)	21
III.11	Rotation of Wigner Ville Distribution Function	22
IV.1	Back-Propagation Activation Function	27
IV.2	Quadratic Cost Function	27

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 that 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 use cepstrum, a sequence of numbers that characterize a frame of speech. Cepstrum is usually combined with Hid-

den 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 that's 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.