ADVANCED TECHNIQUES IN SPEAKER DIARIZATION FOR ARABIC TV BRPADCAST

By

Mohamed Salem Mohamed Elhady

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JULY 2017

ADVANCED TECHNIQUES IN SPEAKER DIARIZATION FOR ARABIC TV BRPADCAST

By
Mohamed Salem Mohamed Elhady

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Under the Supervision of

Prof. Mohsen Abdelrazeq Rashwan

Professor of Communication Engineering
Electrical Communication and Electronics Engineering
Faculty of Engineering, Cairo University

Prof. Sehrif Mahdy Abdou

Professor
Information Technology department
Faculty of Computer and Information Science, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JULY 2017

ADVANCED TECHNIQUES IN SPEAKER DIARIZATION FOR ARABIC TV BRPADCAST

By
Mohamed Salem Mohamed Elhady

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communications Engineering

Approved by the
Examining Committee

_____
Prof. Dr. Mohsen Abdulrazeq Rashwah, Thesis Main Advisor

_____
Prof. Dr. Sherif Mahdy Abdou, Thesis Advisor
Faculty of computer and Information science, Cairo Universiy

_____
Prof. Dr. Mohamed Fathy Abu-Elyazeed, Internal Examiner

_____
Prof. Dr. Mohamed Waleed Talaat Fakhr, External Examiner
Faculty of Computer Information Technology , Arab Academy for Science and Technology

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JULY 2017

Engineer's Name: Mohamed Salem Mohamed Elhady
Date of Birth: 15/11/1992
Nationality: Egyptian
E-mail: mhmd.sl.elhady@gmail.com
Phone: 01121978364
Address: 38th Cablat street Wadi Hof Helwan

Registration Date: 1/10/2014
Awarding Date: / /2017
Degree: Master of Science
Department: Electronics and Communications Engineering

Supervisors:

Prof. Mohsen Abdelrazeq Rashwan
Prof. Sherif Mahdy Abdou
Faculty of Computer and Information Science, Cairo University

Examiners:

Prof. Mohsen Abdulrazeq Rashwan (Thesis main Advisor)
Prof. Sherif Mahdy Abdou (Thesis Advisor)
Faculty of computer and Information science
Porf. Mohamed Fathy Abu-elyazeed (Internal Examiner)
Porf. Mohamed Waleed Talaat Fakhr (External Examiner)
Faculty of computer and Information Technology, Arab Academy for science and technology

Title of Thesis:

ADVANCED TECHNIQUES IN SPEAKER DIARIZATION FOR ARABIC TV BROADCAST

Key Words:
Speaker Diarization; Speech Processing; Machine Learning; speech activity detector

Summary:

Speaker Diarization is known as the task that answers the question, who spoke, when in an audio file or a set of audio files that contain unknown number of speakers. The determination of speaker segments is done in an unsupervised manner. Our Speaker Diarization system composed of two main blocks; Speech Activity Detector and Speaker Clustering. In speech activity detection we propose several solutions including; Phoneme Recognition system, SVMHMM system and i-vector based system. In speaker clustering area we propose an enhancement over state of the art techniques as cosine based Hierarchal Agglomerative Clustering. Such enhancement including enhancing clustering by classification methods as SVM, DNN and Random Forrest. Finally we investigated enhancing the i-vector representation via extracting them from a DNN based background model

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

BIC ..... Bayesian Information Criterion
$NCLR$ ..... Normalized Cross Likelihood Ratio
DNN ..... Deep Neural Networks
ZCR ..... Zero Crossing Rate
GMM ..... Gaussian Mixture Models
RF ..... Random Forrest
SVM ..... Supporting Vector Machines
HMM ..... Hidden Markov Models
JFA ..... Joint Factor Analysis
RMS ..... Root Mean Square
UBM ..... Universal Background Model
MAP ..... Maximum Aposteriori Probability Estimate
LLR ..... Log Likelihood Ratio
$PLDA$ ..... Probabilistic Linear Discriminant Analysis
HAC ..... Hierarchal Agglomerative Clustering
GNG ..... Growing Neural Gas
DER ..... Diarization Error Rate
$WCCN$ ..... Within Class Covariance Normalization
$KL2$ ..... Kullback-Leibar Distance
T^2 ..... Hostelling - Statistic distance
$MFCCs$ ..... Mel Frequency Cepstral Coefficients
Σ ..... Covariance Matrix of a Gaussian Model
μ ..... Mean of a Gaussian Model
σ_a .....variance of a Gaussian Model
a

BIC ..... Bayesian Information Criterion
$NCLR$ ..... Normalized Cross Likelihood Ratio
DNN ..... Deep Neural Networks
ZCR ..... Zero Crossing Rate
GMM ..... Gaussian Mixture Models
RF ..... Random Forrest
SVM ..... Supporting Vector Machines
HMM ..... Hidden Markov Models

JFA ..... Joint Factor Analysis

RMS ..... Root Mean Square

UBM ..... Universal Background Model

MAP ..... Maximum Aposteriori Probability Estimate

LLR ..... Log Likelihood Ratio

$PLDA$ ..... Probabilistic Linear Discriminant Analysis

HAC ..... Hierarchal Agglomerative Clustering

GNG ..... Growing Neural Gas

DER ..... Diarization Error Rate

$WCCN$ ..... Within Class Covariance Normalization

$KL2$ ..... Kullback-Leibar Distance

T^2 ..... Hostelling - Statistic distance

$MFCCs$ ..... Mel Frequency Cepstral Coefficients

Σ ..... Covariance Matrix of a Gaussian Model

μ ..... Mean of a Gaussian Model

σ_a .....variance of a Gaussian Model a

Symbol Description

Acknowledgements

In the name of Allah the most merciful the most gracious; all thanks to Allah and peacebeuponhismessengerMohamedandhiscompanions. Thence, I would like to thank all those who helped and supported me during the course of my research.

I would like to thank Prof. Mohsen Rashwan and Prof. Sherif Abdou for their continuous support during and before this work. Their guidance and help have had the most remarkable effect on my life choices.

I would also like to thank RDI (Research and Development International) for, without their support with sufficient Tools, this work would have never been brought into the light.

Thanks, Shall also go to, Eng. Hani Ahmed from RDI whose technical guidance helped me to go through this work.

Last but least, I would like to show my gratitude to my family, and my fiance for their continuous support and for inspiring me to go forward whenever it is needed.

Abstract

Speaker Diarization is known as the task that answers the question; who spoke, when and where in an audio file or set of audio files that contain an unknown number of speakers. The determination of speaker segments is done in an unsupervised manner. Originally, Speaker Diarization was proposed as a research topic related to speech recognition. In recent years, it has been introduced as an independent research topic. Competitions and workshops have been dedicated to that area. In this thesis, we propose advanced techniques in Speaker Diarization for Arabic TV broadcast. We focus on Arabic as considered one of the most complex spread languages and the strongest representative of Semitic languages. Our Speaker Diarization system composed of two main blocks; Speech Activity Detector and Speaker Clustering.

In Speech Activity Detection we tackle the problem of speech/non-speech segmentation. We propose two main enhancements in that area; first, a phoneme based speech activity detector. In the phoneme recognition system, we utilize Speech Recognition techniques to solve the problem of speech and non-speech discrimination. Developing a phoneme recognition system could achieve an accuracy of 99% in speech detection and over 97.2% in non-speech detection. Second; i-vectors for speech activity detection. In that experiment, we developed a technique based on speaker recognition techniques. We start by a classification experiment of speech and non-speech using SVM. Classification results achieved 98% to classify speech and non-speech. Those results were motivating to install the i-vector technique in our Speech Activity Detection system. We compare the proposed systems with famous state of the art techniques as SVM-HMM and GMM-HMM.

The second problem we investigate is Speaker Clustering. We started by developing state of the art techniques in Speaker diarization which currently based on i-vectors and cosine based Hierarchal Agglomerative Clustering (HAC). In this area, we propose enhanced clustering technique based on i-vectors and Agglomerative clustering associated with the supervised classification. We experiment three main classification techniques SVM, DNN, and Random Forrest. We compare the enhanced techniques with state of the art techniques. Results show improvement over state of the art techniques using SVM enhancement by 1.7% reducing Diarization Error Rate from State of the art Baseline system of 24.4% to 22.67%

Chapter 1: Introduction

The Arabic language is considered one of the most complex languages morphologically. Due to its variety and lexical sparseness, its speech-related tasks considered difficult compared to other languages which have already wide speeded tools and scripts. For example, In Automatic Speech Recognition tasks, English language has already begun using Kaldi scripts, CMU Sphinx and HTK. In other tasks such as Speaker Verification, Speaker Recognition, and Speaker diarization, there are LIUM scripts for French broadcast speaker diarization, Kaldi scripts for English speaker recognition and Sidekit for English Speaker verification and recognition. In this thesis, we focus our research on tackling the problem of speaker diarization in Arabic Broadcast TV. Such a problem hasn't been studied in depth for Arabic TV broadcast tasks. We investigate in this thesis most of the researches that have been developed recently in Speaker diarization for other languages (especially English as the widest spread language in Speech Research). We implement state of the art algorithms in order to achieve comparable performance with other speaker diarization systems. Finally, we propose our own enhancements and collaboration to the area for the sake of developing a better Speaker Diarization system for the Arabic Language. This research is done with the hope it would be a good reference for researchers who want to work on the same problem for the Arabic Language.

1.1 Important Notes

- **Mel Frequency Cepstral Coefficients MFCCs:** The number of cepstral used in MFCCs during this thesis is 25 unless mentioned other wise. All used MFCCs are asscoiated with delta and delta delta features giving total feature vector length of 75.
- **Frame length:** Acoustic Frame length is composed of 10 ms. unless mentioned otherwise.
- **Results:** All results obtained in each experiment are on the development dataset of the MGB-2 Challenge broadcast.

1.2 Problem Definition

The problem this thesis focuses on is Speaker Diarization for Arabic Broadcast News. Speaker Diarization answers the question of who spoke when and where during an episode or an audio file containing an unknown number of speakers. Speaker Diarization was originally a research topic related to speech recognition in order to boost the performance of Speaker Adaptive Training. However, in the few recent years, it has been an independent standing research point for wide range research topics such information retrieval, navigation and higher level inference. Thus, a wide interest has motivated researchers to contribute their research in this particular area. Moreover, there have been dedicated sessions and workshops for Speaker Diarization track. The increase of interest can date back to 2006 where the first challenge released that focuses on the English Broadcast