



Cairo University

# **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:** 38<sup>th</sup> 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  
Prof. Mohamed Fathy Abu-elyazeed (Internal Examiner)  
Prof. 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 Hierarchical Agglomerative Clustering. Such enhancement including enhancing clustering by classification methods as SVM, DNN and Random Forest. Finally we investigated enhancing the i-vector representation via extracting them from a DNN based background model

# Table of Contents

<b>List of Tables</b>	<b>iv</b>
<b>List of Figures</b>	<b>v</b>
<b>List of Symbols and Abbreviations</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>Abstract</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Important Notes . . . . .	1
1.2 Problem Definition . . . . .	1
<b>2 Literature Review</b>	<b>4</b>
2.1 Speech Activity Detection SAD Literature . . . . .	4
2.1.1 SAD Overview . . . . .	4
2.1.2 Feature Set for SAD system . . . . .	4
2.1.3 Signal Processing Based Speech Activity Detectors . . . . .	7
2.1.4 Model Based Speech Activity Detection . . . . .	8
2.2 Speaker Clustering Literature . . . . .	10
2.2.1 Classical Speaker Clustering techniques . . . . .	10
2.2.2 GMM Based Diarization Systems . . . . .	12
2.2.3 Advances in Speaker Clustering . . . . .	13
2.2.4 Speaker Clustering Based on I-vectors . . . . .	14
<b>3 Mathematical Background</b>	<b>18</b>
3.1 Metric Based Distance . . . . .	18
3.1.1 Bayesian Information Criterion BIC . . . . .	18
3.1.2 Kullback-Leibler distance (KL2) . . . . .	19
3.1.3 Hotelling $T^2$ -statistic distance (or $T^2$ distance) . . . . .	19
3.2 Speaker Modeling . . . . .	20
3.2.1 Training Universal Background Model . . . . .	20
3.2.2 Super vectors for Speaker . . . . .	20
3.2.3 Maximum Aposteriori Estimate MAP . . . . .	21
3.2.4 Joint Factor Analysis . . . . .	22
<b>4 Speech Activity Detection</b>	<b>27</b>
4.1 Why We Need Speech Activity Detection . . . . .	27
4.2 SVM and GMM Experiments . . . . .	27
4.2.1 Data Set . . . . .	27
4.2.2 SVM SAD System . . . . .	27
4.2.3 GMM Classifier SAD System . . . . .	28
4.2.4 Viterbi-Smoothing: . . . . .	28

4.3	HMM Phoneme Recognition SAD System . . . . .	28
4.3.1	Feature Set . . . . .	28
4.3.2	Transcribing training data . . . . .	28
4.3.3	Training mono-phone System . . . . .	28
4.3.4	Decoding Mono-phone System . . . . .	30
4.4	Results . . . . .	31
4.5	Experimentation Summary . . . . .	32
4.6	I-vectors For Speech Activity Detection . . . . .	32
4.6.1	Proof of Concept . . . . .	32
4.6.2	SVM Music/Speech Detector . . . . .	33
4.6.3	Linear SVM based Speech Activity Detector . . . . .	34
<b>5</b>	<b>Speaker Clustering</b>	<b>36</b>
5.1	Experiments . . . . .	36
5.1.1	Dataset Description . . . . .	36
5.1.2	Evaluation Metric . . . . .	37
5.1.3	NCLR Experiments . . . . .	38
5.1.4	Growing Neural Gas and i-vectors . . . . .	39
5.1.5	Ward Clustering . . . . .	42
5.1.6	Cosine Distance based HAC . . . . .	45
5.1.7	Enhancing Cosine based HAC . . . . .	45
5.1.8	Enhancing HAC Cosine with Random Forrest . . . . .	55
5.1.9	Compare Results of Three Enhancement Methods . . . . .	58
5.1.10	Adjusting stopping criterion . . . . .	58
5.1.11	Self Organizing Map with I-Vectors . . . . .	59
5.2	Enhanced I-Vectors . . . . .	64
5.2.1	Increasing Training data for UBM . . . . .	64
5.2.2	The DNN-UBM approach . . . . .	65
5.3	Wrapping Up Experiments . . . . .	67
<b>6</b>	<b>Error Analysis</b>	<b>69</b>
6.1	Error Analysis per Speakers . . . . .	69
6.2	Detailed Error Analysis . . . . .	69
6.2.1	Error Types in Dominant Speakers Verses less frequent speakers . . . . .	69
6.3	Comparison With English Results . . . . .	70
6.3.1	English Dataset VS Arabic Dataset . . . . .	70
6.3.2	English System Results . . . . .	70
<b>7</b>	<b>Future Work</b>	<b>71</b>
7.1	Enhanced Speech Activity Detector . . . . .	71
7.1.1	Multiclass Classification . . . . .	71
7.1.2	Different Modeling Techniques . . . . .	71
7.2	Overlap/Non-Overlap Detector . . . . .	72
7.3	Channel Compensation . . . . .	72
7.3.1	Linear Discriminant Analysis (LDA) . . . . .	72
7.3.2	Within Class Covariance Normalization (WCCN) . . . . .	72
7.4	Different Clustering Techniques . . . . .	72

7.5 Future System Description . . . . .	72
<b>8 Conclusion</b>	<b>74</b>
<b>References</b>	<b>75</b>
<b>Appendix A Toolkits Used in Thesis</b>	<b>79</b>

# List of Tables

4.1	Missed Speech Results . . . . .	31
4.2	False Alarm Results . . . . .	31
4.3	FA and Missed Speech rates After rescoring . . . . .	31
4.4	The results for SVM classification . . . . .	33
4.5	Missed Speech Final Results . . . . .	34
4.6	False Alarm Final Results . . . . .	35
5.1	Structured verses Unstructured DER . . . . .	44
5.2	Compared Results of several HAC systems . . . . .	58
5.3	DER of HAC+ SVM before and After adjustment . . . . .	58
5.4	DER before and after increasing training hours . . . . .	64
5.5	DER fot three developed systems . . . . .	66
5.6	Full Results of Each experiment . . . . .	68



# List of Figures

1.1	Diaization Block Diagram . . . . .	2
2.1	MFCCS extraction Block Diagram . . . . .	6
2.2	LPCs extraction Block Diagram . . . . .	7
2.3	GMM System Block Diagram . . . . .	9
2.4	SVM System Block Diagram . . . . .	9
2.5	DNN-HMM Block Diagram . . . . .	10
2.6	Bottom up verses Top Bottom approaches . . . . .	11
2.7	Flow Diagram for Clustering Algorithms . . . . .	11
2.8	Block Diagram for solaria system . . . . .	13
2.9	HAC cosine clustering approach . . . . .	15
2.10	Block Diagram for PLDA HAC system . . . . .	16
2.11	Block Diagram for dused cosine + PLDA . . . . .	17
3.1	Speaker Modeling using GMM methods . . . . .	21
4.1	Missed Speech VS No. Of phonemes . . . . .	29
4.2	False Alarm VS No. Of phonemes . . . . .	30
4.3	i-vectors of speech and music . . . . .	33
4.4	i-vectors based SAD . . . . .	34
5.1	HAC distribution of MGB challenge data . . . . .	37
5.2	distribution of Development data . . . . .	37
5.3	NCLR system Block Diagram . . . . .	39
5.4	Global Error of GNG . . . . .	40
5.5	Accumulated local error . . . . .	41
5.6	Final Clusters for single episode . . . . .	41
5.7	Cluster Evolution though iterations . . . . .	42
5.8	Ward clustering block diagram . . . . .	43
5.9	i-vector dimension verses DER . . . . .	43
5.10	DER verses i-vector dimension for structured ward . . . . .	44
5.11	DER verses i-vector dimension for HAC . . . . .	45
5.12	SER verses top percentage of data . . . . .	46
5.13	proposed enhanced HAC . . . . .	46
5.14	Percentage of training verse DER for 100 dim i-vector . . . . .	47
5.15	DER verses Percentage of training for 200 dim i-vectors . . . . .	48
5.16	DER verses Percentage of training for 300 dim i-vectors . . . . .	48
5.17	DER verses Percentage of training for 400 dim i-vectors . . . . .	49
5.18	DER verses Percentage of training for 500 dim i-vectors . . . . .	49
5.19	DER verses percentage of training for 100 dim i-vectors . . . . .	50
5.20	DER verses percentage of training for 200 dim i-vectors . . . . .	51
5.21	DER verses percentage of training for 300 dim i-vectors . . . . .	51
5.22	DER verses percentage of training for 400 dim i-vectors . . . . .	52
5.23	DER verses percentage of training for 100 dim i-vectors . . . . .	52

5.24	DER verses percentage of training for 200 dim i-vectors . . . . .	53
5.25	DER verses percentage of training for 300 dim i-vectors . . . . .	53
5.26	DER verses percentage of training for 400 dim i-vectors . . . . .	54
5.27	DER verses percentage of training for 500 dim i-vectors . . . . .	54
5.28	DER verses percentage of training for 100 dim i-vectors . . . . .	55
5.29	DER verses percentage of training for 200 dim i-vectors . . . . .	56
5.30	DER verses percentage of training for 300 dim i-vectors . . . . .	56
5.31	DER verses percentage of training for 400 dim i-vectors . . . . .	57
5.32	DER verses percentage of training for 500 dim i-vectors . . . . .	57
5.33	DER verses Stopping Criterion Threshold . . . . .	59
5.34	SOM Example . . . . .	60
5.35	SOM Clustering Flow Chart . . . . .	62
5.36	SOM Neighborhood Connections . . . . .	63
5.37	SOM Topology Example . . . . .	64
5.38	Supervised UBM Block Diagram . . . . .	65
5.39	DNN-HMM ASR system . . . . .	66
5.40	DNN-UBM based Clustering system . . . . .	66
6.1	Error Analysis per Speaker . . . . .	69
6.2	Error Analysis per Speaker for CALLHOME . . . . .	70
7.1	IBM SAD system . . . . .	71
7.2	Block Diagram for Future Speaker Diarizer . . . . .	73

# List of Symbols and Abbreviations

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

<i>JFA</i>	.....	Joint Factor Analysis
<i>RMS</i>	.....	Root Mean Square
<i>UBM</i>	.....	Universal Background Model
<i>MAP</i>	.....	Maximum A posteriori Probability Estimate
<i>LLR</i>	.....	Log Likelihood Ratio
<i>PLDA</i>	.....	Probabilistic Linear Discriminant Analysis
<i>HAC</i>	.....	Hierarchical Agglomerative Clustering
<i>GNG</i>	.....	Growing Neural Gas
<i>DER</i>	.....	Diarization Error Rate
<i>WCCN</i>	.....	Within Class Covariance Normalization
<i>KL2</i>	.....	Kullback-Leibler Distance
$T^2$	.....	Hotelling - Statistic distance
<i>MFCCs</i>	.....	Mel Frequency Cepstral Coefficients
$\Sigma$	.....	Covariance Matrix of a Gaussian Model
$\mu$	.....	Mean of a Gaussian Model
$\sigma_a$	.....	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 peace be upon his messenger Mohamed and his companions. 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 not least, I would like to show my gratitude to my family, and my fiancée 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 associated 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