SPEAKER IDENTIFICATION USING MINIMUM VOLUME ELLIPSOIDS AND LARGE-MARGIN CRITERION

By

Eng. Omar Abdallah Abdelfatah Elgendy

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Engineering Mathematics

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JUNE 2015

SPEAKER IDENTIFICATION USING MINIMUM VOLUME ELLIPSOIDS AND LARGE-MARGIN CRITERION

By

Eng. Omar Abdallah Abdelfatah Elgendy

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Engineering Mathematics

Under the Supervision of

Prof. Dr. Abdel-Karim S. O. Hassan
Professor of Engineering Mathematics
Engineering Mathematics Department
Faculty of Engineering, Cairo University

Dr. Moataz M. H. El Ayadi
Assistant Professor
Engineering Mathematics and Physics Department
Faculty of Engineering, Cairo University

Dr. Ahmed Abdel-Naby Ahmed Mahmoud
Assistant Professor
Engineering Mathematics and Physics Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JUNE 2015

SPEAKER IDENTIFICATION USING MINIMUM VOLUME ELLIPSOIDS AND LARGE-MARGIN CRITERION

By

Eng. Omar Abdallah Abdelfatah Elgendy

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Engineering Mathematics

Approved by the Examining Committee:

Prof. Dr. Abdel-Karim S. O. Hassan, Thesis Main Advisor

Prof. Dr. Mohamed A. El-Gamal, Internal Examiner
(Faculty of Engineering, Cairo University)

Prof. Dr. Reda A. El-Khoribi, External Examiner
(Faculty of Computers and Information, Cairo University)

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
JUNE 2015

Engineer's Name:	Eng. Omar Abdallah Abdelfatah Elgendy
Date of Birth:	27/8/1988
Nationality:	Egyptian
E-mail:	omar.abdallah@eng.cu.edu.eg
Phone:	01005118321
Address:	Mokattam 15711, Cairo, Egypt
Registration Date:	11/10/2011
Awarding Date:	–/–/—-
Degree:	Master of Science
Department:	Engineering Mathematics

Supervisors:

Prof. Dr. Abdel-Karim S. O. Hassan
Dr. Moataz M. H. El Ayadi
Dr. Ahmed Abdel-Naby Ahmed Mahmoud

Examiners:

Prof. Dr. Abdel-Karim S. O. Hassan	(Thesis main advisor)
Prof. Dr. Mohamed A. El-Gamal	(Internal examiner)
Prof. Dr. Reda A. El-Khoribi	(External examiner)

Title of Thesis:

SPEAKER IDENTIFICATION USING MINIMUM VOLUME ELLIPSOIDS AND LARGE-MARGIN CRITERION

Key Words:

Speaker Identification; Statistical Pattern Recognition; Discriminative Training; Large-Margin Estimation; Robust Estimation; Minimum Volume Ellipsoid; Minimum Covariance Determinant

Summary:
This thesis proposes two approaches building robust text-independent speaker identification systems in noisy environments. The first approach is based using robsut statistical estimators such as the Minimum Volume Ellipsoid and Minimum Covariance Determinant to improve the maximum likelhood based classifiers by making them robust to outliers. The second approach is based on using an improved version for the large-margin discriminative critrerion, whcih is charectrerized by its simplicty compared to the standard large-margin approach and other discriminative approaches. Experimental results show that our proposed techniques outperform the baseline techniques

Table of Contents

List of Tables

List of Figures

Nomenclature

AULS	Accelerated Unconstrained Line Search
CDHMM	Continuous-Density Hidden Markov Model
CDS	Cosine Distance Scoring
DT	Discriminative Training
EM	Expectation Maximization
FBLPC	Formant Based Linear Prediction Coefficients
GMER	Generalized Minimum Error Rate
GMM	Gaussian Mixture Model
GSV	Gaussian SuperVector
GT	Generative Training
HMM	Hidden Markov Models
IDTTDR	IDentification Time to Test Duration Ratio
JFA	Joint Factor Analysis
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LME	Large-Margin Estimation
LMI	Linear Matrix Inequality
LPA	Linear Prediction Analysis
LPC	Linear Prediction Coefficients
MAP	Maximum APosteriori
MFCC	Mel-Frequency Cepstrum Coefficients
MCD	Minimum Covariance Determinant
MCE	Minimum Classification Error
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MMI	Maximum Mutual Information
MSE	Minimum Squared Error
MVE	Minimum Volume Ellipsoid
NAP	Nuisance Attribute Projection
PCA	Principal Component Analysis
PDF	Probability Density Function
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction

RPF	Radial Basis function
SID	Speaker IDentification
SDP	Semi-Definite Programming
SNR	Signal-to-Noise Ratio
SOCP	Second-Order Cone Programming
SVM	Support Vector Machine
TLE	Trimmed Likelihood Estimators
UBM	Universal Background Model
WCCN	Within-Class Covariance Normalization

Acknowledgements

"Praise be to God, who guided us to this: had God not guided us, We would never have found the way.", 7:43.

I would like to express my deep appreciation to my supervisors; Prof. Dr. Abdel-Karim S. O. Hassan, Dr. Moataz Elayadi, and Dr. Ahmed Abdel-Naby. I consider myself as a lucky person to work under their supervision because I have learnt a lot from them, both in scientific and personal aspects. Actually, this work would not have been developped in this form without their guidance, scientific support, and fruitful discussions.

I sincerely acknowledge my friends Mahmoud Taha, Mohamed Fouda, Ahmed Elsheikh, Mostafa Abdalla, Mohamed Elshafey, Ahmed Essam, and Ahmed Etman for their continuous encouragement, support and sharing experiences and knowledge.

Also, I would like to express my gratitude to all my professors and colleagues in Engineering Mathematics and Physics Department for their enthusiastic encouragements especially, Prof. Hany Abdel-Malak, Prof. Nabila Philib, Prof. Mohamed El-Gammal, Prof. Adel Mohsen, Prof. Nadia Hussein, Prof. Said Grace, Prof. Labib Eskandar, Prof. Mohamed Hesham Assoc. Prof. Maha Amin, Assoc. Prof. Ahmed Gomaa, Dr. Amany El-Gamal, Dr. Eman El-Maghrabi, Dr. Ahmed Abdel-Samea and Dr. Mohamed Ibrahim.

In addition, for their faithful friendship and encouragement, it is pleasure to thank my friends Ahmed Atef, Islam Refaat, Islam Esmat, Amr Mahmoud, Mohamed Yehia, Usama Toson, Ahmed Imam, Shimaa Ebid, Wafaa Saber, Mahmoud Ayad and Yassin Essam.

Finally, I would like to express my profound gratitude to my parents, sisters, and wife whose patience, continuous encouragement, support, and constant presence enabled me to complete this work.

Omar Abdallah

Dedication

This dissertation is dedicated to my parents

and my little sisters.

I also dedicate this work to my beloved wife and daughter.

Abstract

Building a robust Text-Independent Speaker Identification (SID) system that can work effectively in different environments is a very challenging task. It is well known that the performance of Text-Independent SID systems deteriorates significantly with the presence of noise and spectral distortion in the training and testing utterances.

In this thesis, two approaches for improving the performance of standard Gaussian Mixture Model (GMM)-based speaker identification systems are introduced. For these approaches, the initial model is a GMM trained by the standard Maximum Likelihood Estimation (MLE) method. In the first approach, the robustness of the GMM classifier to outliers is increased by deploying techniques proposed in the literature of robust statistics. We consider prominent estimators that belong to the family of Minimum Volume Ellipsoids (MVEs), where these ellipsoids are called the Löwner-John MVE, the Rousseeuw MVE, and the Minimum Covariance Determinant (MCD). Compared to the traditional method, the introduced methods are less sensitive to outliers (in the feature-vector space), caused by additive noise and spectral distortion. At the same time, they can be efficiently implemented using modern day computers. Moreover, in the testing phase, we use a very simple distance metric for comparing the unknown testing utterance against the speakers' models. The proposed methods have been applied to the NIST 2000 speaker recognition evaluation and compared against state-of-the-art techniques such as the supervectors method and the i-vectors methods. Experimental results show that the proposed method provides up to 16% relative improvement in the identification performance over the i-vector methods and up to 40% reduction in testing time when compared to the MLE method.

In the second proposal, the main aim is to increase the discriminative ability of the GMM classifier by using a modified version of the Large-Margin discriminative criterion to estimate its parameters. Generally, discriminative techniques such as Large Margin Estimation (LME) outperform standard generative techniques at the expense of additional training complexity [98, 47, 59]. In this thesis, we introduce some modifications to the standard LME criterion to decrease its complexity without much sacrifice in the classification performance. Simulation results reveal that our proposed LME outperforms the MLE and the Minimum Classification Error (MCE) criterion by 11.07% and 8.76%, respectively. Moreover, the LME criterion is faster than MCE criterion where the amount of relative reduction in the calculation time for the gradient and cost functions are estimated to be 82.92% and 37.75%, respectively.

Chapter 1: Introduction

Speech signal is one of the most rich signals in information; it carries the following types of information: linguistic information (spoken words and the language), speaker information (e.g., identity, emotional state, accent) and environmental information (e.g., the signal to noise ratio and the transmission bandwidth). Unlike some other biometrics, it does not require special sensors for acquisition; all what it needs for measurements is a microphone. Therefore, human voice has long been considered as an important characteristic to be used for authentication. The process of recognizing people by their voices is generally referred to as *speaker recognition*. Nowadays, speaker recognition has various potential applications such as user authentication in call centres and E-commerce systems, recognizing persons in a conversation for forensics, and security check in military environments.

Recently, research in speaker recognition has gained much momentum thanks to the widespread of low-cost and powerful computing devices. There are two main tasks of speaker recognition: speaker identification (SID) and speaker verification. In speaker identification, it is required to search among a set of individuals for the speaker of a given utterance. The set of individuals are commonly described as *registered* or *enrolled* speakers in the system. On the other hand, speaker verification is concerned with authenticating the claimed identity of a person using his voice. An SID system is said to be *open-set* if it can decide whether speaker of the unknown utterance is already enrolled in the system or not; otherwise, it is called a *closed-set* SID system. On another front, an SID system is said to be text-independent if it does not depend on the spoken text. On the other hand, text-dependent SID systems requires modelling the linguistic content of the given utterances. In this thesis, we will focus on the text-independent closed-set SID problem. A brief review on SID Systems will be given in chapter 3.

In fact, it is still infeasible to implement practical SID systems that can be deployed in real-life applications [62]. Unlike speaker verification, the performance of SID systems deteriorates significantly when the number of enrolled speakers increases. Several factors attribute to this degradation in performance including the noisy recording conditions (additive and convolutional noise), the spectral distortion caused by the communication channel, and the mismatch between training and testing recording conditions.

Our main objective, in this thesis, is to compensate these effects so as to improve the performance of existing SID systems. According to the literature, the compensation can be performed in three domains:

Feature domain

In this approach, signal processing techniques are applied to estimate and remove the noise spectral components from the raw speech signal and/or the extracted speech features [117, 94, 137, 13, 64].

Classifier domain

In this approach, the estimation procedure of the classifier parameters is modified to account for the presence of noise and spectral distortion in the speech signal. This is achieved by either explicitly modelling the effects of noise and spectral distortion or by employing robust estimation criteria [125, 113, 32, 126, 48].

Score domain

In this approach, the scores of the candidate models are normalized to account for the different recording environments of the training and testing utterances.

Our proposed techniques belong to the category of classifier-domain compensation.

1.1 Thesis Contributions

In this thesis, we propose two techniques for building a robust SID system in noisy environment and under mismatch between training and testing conditions. In both methods, we use the Gaussian Mixture Model (GMM) as the main statistical classifier [90].

In the first technique, we estimate the GMM parameters based on two robust criteria [42]: the Minimum Volume Ellipsoids (MVE) [118, 109, 121] and the Minimum Covariance Determinants (MCD) [41, 33]. It is sought that those criteria are more insensitive to outliers than the standard Maximum Likelihood (ML) criterion leading to a significant improvement in the classification performance of the overall SID system. Experimental results show an increase in the classification accuracy that reach 6.16% compared to the regular ML approach for the full covariance case. Moreover, results show that our testing criterion is simpler and faster than the classical log-likelihood criterion used in the regular ML approach for both cases of full and diagonal covariance matrices. Furthermore, our approach outperforms the state-of-the-art i-vector method for short testing utterance and the amount of relative improvement reaches 16% in some cases.

In the second approach, we propose a novel discriminative criterion for GMM training. Our proposed criterion is a modification to the popular Large Margin Estimation (LME) criterion [59, 103], successfully employed in other speech recognition applications [47, 131, 134]. In particular, we approximate the LME criterion to speed up the training and testing processes without much sacrifice in the classification performance. Simulation results reveal that our proposed technique outperforms other discriminative criteria such as the Minimum Classification Error (MCE) [51] and the Generalized Minimum Error Rate (GMER) [65]. The MCE criterion tends to minimize an empirical loss criterion which is