SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

AUTOMATIC SPEAKER RECOGNITION

M.Sc. THESIS

Yussouf NAHAYO

Department	**:**	**COMPUTER AND INFORMATION ENGINEERING**
Field of Science	**:**	**COMPUTER ENGINEERING**
Supervisor	**:**	**Assist. Prof. Dr. Seçkin ARI**

July 2015

SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

AUTOMATIC SPEAKER RECOGNITION

M.Sc. THESIS

Yussouf NAHAYO

Department	**:**	**COMPUTER AND INFORMATION ENGINEERING**
Field of Science	**:**	**COMPUTER ENGINEERING**
Supervisor	**:**	**Assist. Prof. Dr. Seçkin ARI**

This thesis has been accepted unanimously / with majority of votes by the examination committee on .....

…….	…….	…….
Head of Jury	**Jury Member**	**Jury Member**

DECLERATION

I declare that all the data in this thesis was obtained by myself in academic rules, all visual and written information and results were presented in accordance with academic and ethical rules, there is no distortion in the presented data, in case of utilizing other people's works they were refereed properly to scientific norms, the data presented in this thesis has not been used in any other thesis in this university or in any other university.

Yussouf NAHAYO

27.07.2015

PREFACE

This master's thesis is based on the study of data classification using different discriminative approaches for automatic speaker recognition. The research took place at the Department of Computer Engineering, University of SAKARYA.

My gratitude is expressed to YTB (Yurtdışı Türkler ve Akraba Topluluklar Başkanlığı) for giving me an opportunity and support to study in a beautiful country which has a high level in education such as Turkey.

I express my sincere gratitude to my supervisor Dr. Seçkin ARI for his guidance and the great number of counseling he provided to me that contributed to the successful completion of this work.

My thanks are also expressed to the department of Computer and Information Engineering, the field of Computer Engineering, the Faculty of Science and Technology, and the SAKARYA University.

I'm also very grateful to my parents, relatives, different families and friends deserve praise for the pain and sacrifices endured during my education.

I would not forget to appreciate the company and friendship from my classmates especially those who contributed to the completion of this work.

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

ASI	: Automatic Speaker Identification
ASR	: Automatic Speaker Recognition
ASV	: Automatic Speaker Verification
EM	: Expectation Maximization
FFT	: First Fourier Transform
GMM	: Gaussian Mixture Model
HMM	: Hidden Markov Models
IR	: Identification Rate
K	: Nearest neighbors number
K-NN	: K-Nearest Neighbors
LLR	: Log likelihood ratio
MAP	: Maximum A Posteriori
MFCC	: Mel Frequency Cepstral Coefficients
NB	: Naive Bayes
SNR	: Signal to Noise Ratio
SVM	: Support Vector Machines
TIMIT	: Acoustic-Phonetic Continuous Speech Corpus
UBM	: Universal Background Model
VQ	: Vector Quantization

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Keywords: SVM, K-NN, NB, GMM, TIMIT, Combination

This master project focuses on the study of data classification using different discriminative approaches for speaker recognition with and without GMM modeling: Automatic speaker identification on text-independent case.

First; a study of different classifiers (SVM, K-NN, NB) was applied by adopting certain parameters of each approach. In step two, a multi Gaussian model based on the Expectation Maximization (EM) algorithm for generating a dictionary of reference models has been implemented. The generated models are the input vectors for these different hybrid systems implemented: GMM-SVM, GMM-KNN and GMM-NB. In step three, a combination of the hybrid systems was developed. The study results showed the effectiveness of the implemented methods. In the end, in order to test the robustness of the implemented systems, random noises have been added to the database (TIMIT) used during this study.

OTOMATİK KONUŞMACI TANIMA

ÖZET

Anahtar Kelimeler: SVM, K-NN, NB, GMM, TIMIT, Kombinasyon

Metin-Bağımsız Durumda Otomatik Konuşmacı Tanımlama başlıklı ana projede GMM modelleme olmadan konuşmacı tanınması için farklı ayrımcı yaklaşımlar kullanarak veri sınıflandırma çalışması üzerine odaklanmaktadır.

İlk olarak; farklı sınıflandırıcıların çalışması (SVM, K-NN, NB) için her yaklaşımda bazı parametreler adapte edilerek uygulanmıştır. İkinci adımda, referans modellerinin bir sözlüğünü oluşturmak için Beklenti Maksimizasyonu (EM) algoritmasına dayanan çoklu Gauss modeli uygulanmıştır. GMM-SVM, GMM-KNN ve GMM-NB modelleri uygulanan bu farklı hibrid sistemler için giriş vektörleridir.Üçüncü adımda, hibrit sistemlerin bir kombinasyonu geliştirilmiştir. Çalışmanın sonuçları uygulanan yöntemlerin etkinliğini göstermiştir. Son olarak, Bu çalışma esnasında uygulama sistemlerin sağlamlığını test etmek amacıyla, rastgele gürültüler veri tabanına (TIMIT) eklenmiştir.

Konuşma işareti, kelime veya konuşulan anlam hakkında bilgi taşımakla birlikte konuşanın fizyolojisi, ruh hali, yaşı, cinsiyeti, lehçesi gibi birçok bilgiyi aynı anda barındırabilen karmaşık bir işarettir. Bu bilgilerin birine veya birkaçına odaklanarak, farklı sistemler gerçeklestirebilir. Örneğin konuşma tanıma, dil tanıma, cinsiyet tanıma, konuşmacı tanıma… Konuşma tanıma, söylenen sözcüğün anlamı ile ilgilenilirken konuşmacı tanıma ise sözcüğü söyleyen kişinin kimliği ile ilgilenilir.

Insanlar konuşanın kimliğini belirlemek için sözle ilgisi olmayan pek çok ipucu kullanmaktadır. Bu ipuçları pek iyi anlaşılmamakla birlikte kabaca anlam ile ilişkili olanlar “yüksek seviye”, konuşmanın akustik yanı ile ilişkili olanları “düşük seviye” ipuçları olarak gruplandırılmaktadır. Yüksek seviye ipuçları, kelime kullanımı, söyleyişteki kişisel özellik ve konuşma karakteristiği ile ilişkili olmayan konuşmacıya özel karakteristik özellikler içerir. Bu ipuçları kişinin konuşma söyleyiş biçimi dolayısıyla değişik yaşam biçimlerine bağlı olarak farklılıklar gösterir. Bu tip ipuçları öğrenilmiş davranış olarak ortaya çıkar. Düşük seviye ipuçları kişinin sesiyle direkt ilişkili olup yumuşak, sert, kaba, açık, yavaş veya hızlı gibi nitelikler içerir. Düşük seviye ipuçları konuşmacının anatomik yapısı ile doğrudan bağlantılıdır. Konuşmacılar arasındaki anatomik farklılıklar, konuşmacıların ses sistemlerinde bulunan bileşenlerinin boyutları ve şekillerinin farklı olmasından kaynaklanır.

Bu nedenle konuşma sinyalleri güvenilir ve ayırt edici bir özellik olarak kullanılmaya başlanmıştır. Sesin bu öneminden dolayı konuşmacı tanıma sistemleri de önem kazanmaktadır. Konuşmacı tanıma sistemi, genellikle güvenliğin ön planda olduğu yerlerde, kriminal laboratuarlarında, telefon ve internet üzerinden çalışan uygulamalarda kullanılmaktadır.

Konuşmacı tanıma iki ana bölüme ayrılabilir; konuşmacı doğrulama (speaker verification) ve konuşmacı saptama (speaker identification). Konuşmacı doğrulama, bilinmeyen bir ses örneğinin, iddia edilen kişiye ait olup olmadığının belirlenmesidir. Konuşmacı saptama ise bilinmeyen bir ses örneğinin, belli konuşmacıların ses kayıtlarından oluşan bir veritabanı içerisinde hangi kisiye ait olduğunun bulunmasıdır.

Konuşmacı tanıma metne bağımlılık yönünden iki alt gruba ayrılır. Bunlar metne bağımlı ve metinden bağımsız konuşmacı tanımadır. Metne bağımlı sistemlerde konuşulan metin sistem tarafından önceden bilinmektedir. Metinden bağımsız sistemlerde ise, metin, herhangi bir sözdizimi olabilir. Diğer taraftan; konuşmacı tanıma, açık küme ya da kapalı küme olabilir. Kapalı kümede bilinmeyen ses örneği, veritabanındaki konuşmacılardan birisine aittir. Açık kümede ise ses örneği veritabanındaki konuşmacılardan hiç birisine ait olmayabilir. Dolayısı ile açık küme konuşmacı tanıma sistemlerinde, ret sonucunu da içeren fazladan bir olasılık daha vardır.

Bu tez çalışmasında, konuşmacı saptama kapalı kümede metinden bağımsız konuşmacı tanıma sistemi kullanılmıştır.

Konuşmacı tanıma sistemleri iki aşamadan oluşmaktadır. Birincisi eğitim, ikincisi ise test aşaması. Egitim aşamasında tüm kullanıcılar, bir referans modeli oluşturmak için ses örnekleri verir, ikinci aşamada ise giriş sinyali referans modelleri ile karşılaştırılarak saptama yapılır.

Konuşmacı tanıma sistemi Öznitelik Vektörleri çıkarma ve Modelleme olarak iki ana kısımdan oluşur. Konuşmacı tanımada öznitelik vektörü çıkarma önemli bir yer oluşturmaktadır. Bu şekilde kişileri temsil eden sayısal vektörler oluşur. Özellik vektörleri daha sonra önceden belirlenen modeli eğitmek için kullanılır. Sistemin en sonunda karar mekanizması vardır. Karar mekanizmasının girişindeki test vektörü ve eğitilmis model kullanılarak test örneğindeki sesin hangi konusmacıya ait olduğu tespit edilir.

Konuşmacı tanımanın ilk aşamasında kullanılan tekniklerin amacı sınıflandırma için öznitelik vektörleri çıkarmaktır. Amaç çok fazla olan konuşma verilerinin, konuşmacıyı tanımlayabilecek vektörlere indirgenmesi ve bir sonraki aşama olan sınıflandırma için kullanıslı veriler üretmektir. Konuşmacı tanımada kullanılacak özniteliklerin, zamanla değişmemesi, gürültüden etkilenmemesi ve diğer konuşmacılardan kolay ayrılabilir olması istenir. Öznitelik vektörü çıkarma için kullanılan yöntemler genel olarak iki gruba ayrılır. Bunlar parametrik ve parametrik olmayan yaklaşımlardır.

Parametrik yaklaşım: Sesli ifadenin üretiliş mekanizmasının tahmin edilmesine yönelik bir modeldir. Bir sesli ifade üretim sistemi öngörülür. Bu yöntemde giriş (kesin olarak bilinmez fakat tahmin edilir), ve çıkış (sesli ifadenin kendisi) arasında bir sesli ifade üretim fonksiyonu oluşturulur. Bu fonksiyonun parametreleri sesli ifade tanıma sisteminde öznitelik vektörü olarak kullanılır. Ses işleme alanında en çok kullanılan ve daha önce yapılan çalışmalarda en iyi sonuç vermiş olan öznitelikler Mel frekans kepstrum katsayıları (Mel-Frequency Cepstrum Coefficients,MFCC) ve Doğrusal Öngörü Katsayılarıdır (Lineer Prediction Coefficients, LPC). Bu ndenle, bu tez çalısmasında öznitelik olarak MFCC kullanılmıştır.

Konuşmacı tanıma alanında, veritabana şecmeye çok onmelidir. Bu tez çalışmasında, Konuşmacı tanıma deneylerinde TIMIT veritabanı kullanılmıştır. TIMIT veritabanı Amerikan İngilizcesinin 8 ana lehçesini konuşan, 438'i erkek 192'si kadın olmak üzere toplam 630 konuşmacının her birinin fonetik yönden zengin 10'ar adet cümlesini içerir. Öznitelik vektörü olarak mel ölçekli kepstrum katsayıları kullanılmıştır. Ses işareti, 10 ms'lik kısmı örtüşen 20 ms uzunluğundaki çerçevelere ayrılıp Hamming pencere uygulanarak işlenmektedir. Pencerelenen ses işaretinin 512 örnek uzunluklu Hızlı Fourier Dönüşümü (HFD) alınıp, elde edilen vektör mel ölçekte 0-8000 Hz arasına yerleştirilmiş üçgen süzgeç takımına uygulanmıştır. Her bir çerçeveye karşılık olarak TIMIT veritabanı için 8,12 ve 16 boyutlu öznitelik vektörleri elde edilmiştir.

Konuşmacı Modelleme üç grup halinde sınıflandırılabilir: Sablon modeller (Dynamic Time Warping, DTW...), İstatiksel modelleme (Gaussian Mixture Model, GMM ...) ve Diğer Yöntemler (Yapay Sinir Agları (Artificial Neural Network, ANN...))

Bu tez çalışında ,konuşmacı tanıma için: ilk olark farklı sınıflandırıcılar (SVM,K-NN, NB) GMM İstatiksel modellemesi olmadan uygulanmıştır. İkincisi olarak, bu sınıflandırıcılar İstatiksel modellemesi ila uygulanmıştır. İstatiksel metot, konuşmacının ortalama ifade özelliklerini kullanmak yerine olasılık dağılımını kullanarak modellemektir ve sınıflandırmayı ortalama özelliklere göre yapmak yerine olasılığa göre yapmaktır. Gauss Karısım Modeli, konuşmacı tanıma uygulamalarında en çok kullanılan istatiksel yaklaşımdır.

Bu tez calışmasında, SVM,K-NN ve NB sınıflandırma teknikleri kullanılmıştır.

Destek vektör makineleri (SVM) çok çeşitli görevler için uygulanan son zamanların en yaygın sınıflandırıcılarından birisidir. Bu sınıflandırma yöntemi, hastalık teşhisi, konuşmacı tanıma ve yazılan sayıyı tanıma gibi değişik alanlarda uygulanmıştır.

tarafından önerilmiş olup yapısal risk minimizasyonu prensibini kullanmaktadır. Bu yöntemde, iki sınıf arasındaki birbirine en yakın örneklerin uzaklıkların maksimumlaştırıldığı yüksek bir düzlem araştırılır. Doğrusal olarak ayrılamayan veriler için, SVM yardımıyla giriş vektörü yüksek boyutlu bir uzaya doğrusal

olmayan bir fonksiyon yardımıyla eşleştirilir. SVM eğitiminde ikinci dereceden bir optimizasyon problemi kullanılabilir.

K-En Yakın Komşuluk (KNN) algoritması sorgu vektörünün en yakın k komşuluktaki vektör ile sınıflandırılmasının bir sonucu olan denetlemeli, oldukça basit bir öğrenme algoritmasıdır. Buna göre; tanıma yapılacak öznitelik vektörüne en yakın k komşu bulunur. Daha sonra bu k komşu en fazla hangi sınıfa ait ise, o sınıf tanıma sonucu olarak atanır. K sayısını belirlemenin en pratik yolu k‟yı toplam eğitim örnekleri sayısının karekökünden daha az olarak seçmektir. Yöntemin performansını k en yakın komşu sayısı, eşik değer, benzerlik ölçümü ve öğrenme kümesindeki normal davranışların yeterli sayıda olması kriterleri etkilemektedir. KNN algoritmaları büyük boyutlu öznitelik vektörlerinde etkin olmamakla birlikte düşük boyutlu öznitelik vektörleri ile etkin olabilmektedirler.

Naive Bayes sınıflandırma yönteminde, öznitelik vektörünü oluşturan özniteliklerin tamamının istatistiksel olarak bağımsız olduğu kabul edilir. Naive Bayes sınıflandırıcı belirli bir sınıfa ait her bir örneğin olasılığını bulmak için Bayes istatistik ve Bayes teoremi kullanır. Varsayımların bağımsızlığı üzerine vurgu yapılması nedeniyle tecrübesiz,saf anlamına gelen Naive denilir. Naive Bayes sınıflandırıcı belirli bir sınıfa ait her örneğin o sınıfa ait olasılığını bulur.

Çalışan Matlab ortamında yapılmış olup SVM sınıflandırma GMM İstatiksel modellemesi olmadan kullanılarak verinin %3, KNN sınıflandırma için %27, ve NB sınflandırma için %11.

Bu tez çalışmasında; ikinci orarak, sistemin performansını geliştirmek için, hibrit sistemi geliştirmiştir. Bu hibrit sistemi GMM algoritması ile farklı sınıflandırıcılar (SVM,K-NN, NB) birleştirerek oluşturmaktadır.

Hibrit sistemi temel amacı tanımlama oranını artırma ve tanıma sisteminin hesaplama süresini azaltmaktır. Bu ndenle GMM sınıflandırıcılarının giriş matrisi azaltarak super vektörlerin girişine içine birçok kare girişi dönüştürerek ve bu super vektörleri farklı hibrid sistemleri için girişidir.

Çalışan Matlab ortamında yapılmış olup GMM-SVM hibrid sistemi kullanılarak verinin %96, GMM-KNN hibrid sistemi için %87, ve hibrid sistemi GMM-NB için %92. Bu alandaki başka çalışmalarda karşılaştırarak, çalışma sonuçları uygulanan metodların etkinliğini göstermiştir.

Bu tez çalışmasında; üçüncü orarak, en yüksek bir perrformans almak için, hibrid sistemlerin farklı kombinasyonu geliştirilmiştir. Dort tane kombinasyon (GMM-SVM + GMM-K-NN, GMM-SVM + GMM-NB, GMM-K-NN + GMM-NB, GMM-SVM + GMM+K-NN + GMM-NB) geliştirirlmiştir. İlk, her hibrid sistemi bağımsız bir şekilde konuşmacı tanırır sonra tüm sonuçları otomatik konuşmacı doğrulama sistemini kullanarak birleşecektir. GMM-SVM hibrid sistemi iyi sonuçları verdiği gibi, bir kombinasyon içeren GMM-SVM hibrit sistemi de iyi sonuçlar verdi.

Çalışan Matlab ortamında yapılmış olup GMM-SVM + GMM-KNN kombinasyon sistemi kullanılarak verinin %96, GMM-SVM + GMM-NB kombinasyon sistemi için %98, GMM-NB + GMM-KNN kombinasyon sistemi için %97 ve GMM-SVM + GMM-KNN + GMM-NB kombinasyon sistemi için %100. Bu alandaki başka çalışmalarda karşılaştırarak, çalışma sonuçları uygulanan metodların etkinliğini göstermiştir. Hatta en yüksek genel başarım GMM-SVM + GMM-KNN + GMM-NB kombinasyon sistemi ile % 100 olarak gerçekleşmiştir.

Bu tez çalışma sonunda, Bizim hibrid sistemi ve hibrid kombinasyon sisteminin sağlamlığını test etmek için çalışma sırasında kullanılan rastgele sesler veritabanına (TIMIT) eklenmiştir. Çalışmanın sonuçları gürültülü verilerin önünde bizim sistemlerinin etkinliğini göstermiştir.

Çalışan Matlab ortamında yapılmış olup GMM-SVM hibrid sistemi gürültülü verilerin önünde (10 dB) kullanılarak verinin %93, GMM-KNN hibrid sistemi için %72, ve hibrid sistemi GMM-NB için %36. GMM-SVM + GMM-KNN kombinasyon sistemi gürültülü verilerin önünde (10 dB) kullanılarak verinin %97, GMM-SVM + GMM-NB kombinasyon sistemi %96, GMM-NB + GMM-KNN kombinasyon sistemi %85 ve GMM-SVM + GMM-KNN + GMM-NB kombinasyon sistemi %98. Çalışma sonuçları uygulanan metodların etkinliğini göstermiştir.

Bu tez pespektivleri:

- Konuşma birkaç modaliteleri entegrasyonu (dudaklar hareketi, yüz resim) ve karakteristik parametreleri buları birleştirmektir
- Diğer akustik parametrelerin türlerini kullanmaktır
- Bizim sistemleri diğer veritabanları ile değerlendirilmiştir

CHAPTER 1. GENERAL INTRODUCTION

In many applications (access control, criminology, banking, ...) it is necessary to characterize a person by an imprint to distinguish him (or her) from others without ambiguity. Among biometric indices, the voiceprint remains an interesting way to exploit because the voice is the most natural means of communication and the most significant for people.

Automatic Speaker Recognition (ASR) is a study field in perpetual evolution and has a very varied scope which requires mostly further researches.

The ASR mainly contains tasks related to Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV). In [1], J.KHAROUBI finds their applications in various fields, including the security of access cards (credit cards, phone cards, etc.), the access control in databases, e-commerce security, information and booking services.

The ASI is to define the identity of the speaker who has delivered a message (word, sentence, text) from a known group of speakers.

Despite significant work on ASR systems, the ASI systems suffer from a lack of robustness due to the variability of the speech signal. Sources of variability of the speech signal are numerous, such as emotional state of the speaker, linguistic content of the message, recording conditions, stress, etc…

In the present work, we focus on ASI systems in text independent mode. We implement different discriminative classification methods as well as a hybrid of these approaches with the generative modeling GMM.

Among these discriminative approaches, we used the classification by Support Vector Machines (SVM) [2], K-Nearest Neighbors (K-NN) [3], and Naive Bayes (NB) [4] [5].

In [6] [7] [8]; J. Zeljkovic, I.TRABELSI, and L. LAZLI showed that even the use of those classification approaches is promising, their effectiveness has remained limited given to the sequential speech nature, particularly in the presence of a large amount of data.

The robustness improvement of the proposed systems is accomplished by a hybridization based on Gaussian Mixture Model (GMM) and the combination of different decisions of implemented systems.

This thesis contains five chapters, the remaining of the chapters are organized as follows: In the second chapter, we set up a Literature review with a detailed overview of the different modules of an ASR system. In the third chapter, we present a review of discriminative approaches; we also discuss the generative approach GMM as well as the different systems of combination proposed. The fourth chapter is dedicated to the presentation of experimental results of the systems studied. The whole of this document is ended by a general conclusion summarizing our contributions and our main results as well as the perspectives left open by this work.

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

This chapter presents an overview of Automatic Speaker Recognition. It presents various related tasks. It subsequently describes the speaker recognition system structure. To understand the challenges of this research, this chapter also outlines the main problems limiting the robustness of ASR systems. Finally, it introduces some examples of corpus used in ASR and its different domains of application.

2.2. Automatic Speaker Recognition Presentation

The automatic speaker recognition (ASR) processes the information of a speaker from his voice signal in order to identify or to verify him. Figure 2.1. shows the speaker recognition location in the speech processing system.

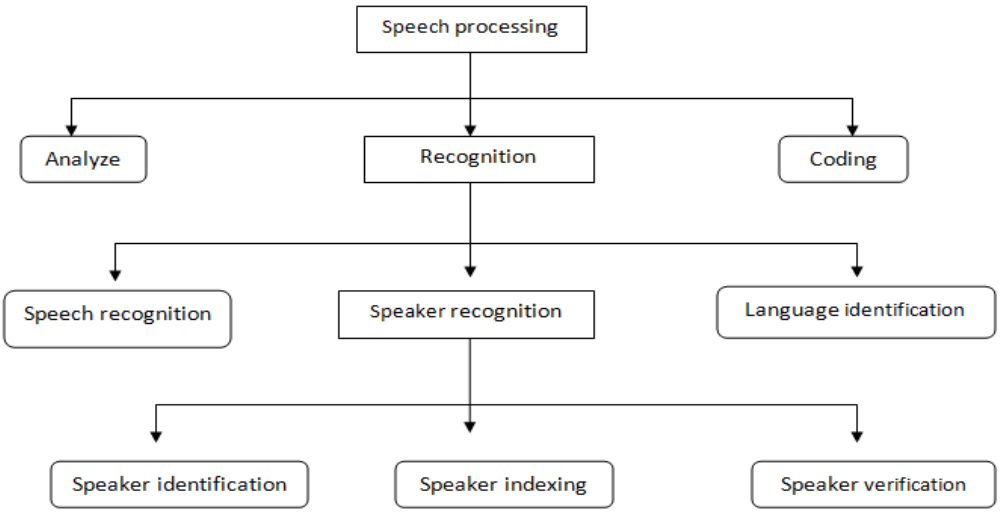

Figure 2.1. Speech processing system (Figure inspired on [9]).

Automatic speaker recognition is a part of the general speech processing field. ASR applications are grouped into three main parts: Automatic Speaker Identification, Automatic Speaker Verification and Automatic Speaker Indexation [9]. Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV) are two most common tasks in the ASR system.

2.2.1. Automatic speaker verification

An Automatic Speaker Verification (ASV) is a process used to verify a speaker identity, if the speaker claims to be of a certain identity and the voice is used to verify this claim. Figure 2.2. represents a modular schema of a speaker verification system. The user who is presented to the system must announce its identity and provide biometric data to the system. The system then compares the reference corresponding to the identity proclaimed in data provided by the user. Their similarity is compared with a threshold Ω. If the similarity measure is greater than that threshold, the user is accepted, otherwise, the user is rejected [9]. In [10] Reynolds presented high performance speaker verification systems based on Gaussian mixture speaker models applied in TIMIT, NTIMIT, Switchboard and YOHO databases. The identification rate varied between 82% and 99 %.

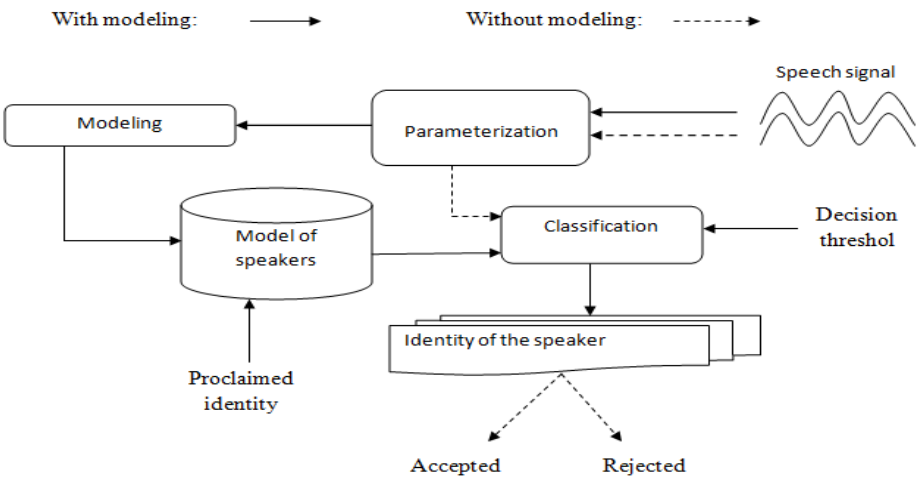

Figure 2.2. Modular schema of speaker verification system.

2.2.2. Automatic speaker identification

From a set of speakers referenced in the system, the task of Automatic Speaker Identification (ASI) is to determine the identity of the speaker by his voice signal (test signal) [11]. Speaker identification systems fall into two sets[9]:

- Open-set identification: it is possible that the unknown speaker is not in the set of speaker models. If no satisfactory match is found, a no-match decision is provided.
- Closed-set identification : the unknown speaker is one of the known speakers.

The Figure 2.3 represents a modular schema of the speaker identification system.

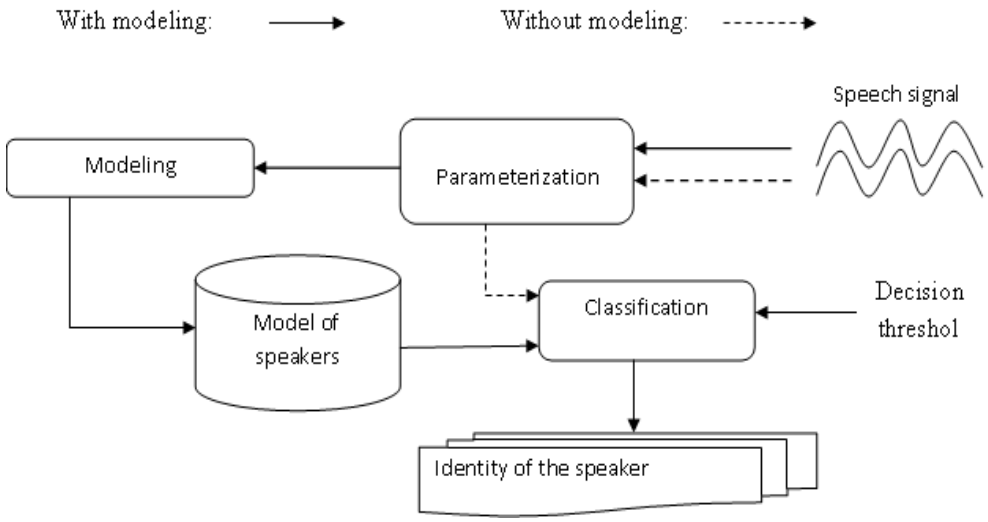

Figure 2.3.Modular schema of speaker identification system.

2.2.3. The dependence and independence of text

Speaker recognition systems fall into two categories: text-dependent and text-independent.

In text-dependent mode, during the test phase, the speaker pronounces the same speech (word, sentence, text) as the one that he pronounced during the training phase of his voice. In this case, systems are mainly distinguished by the context of the text.

In fact, the speech pronounced by the speaker must be known by the system and can be selected by the speaker (password, sentence) or imposed by the system (PIN code) [12] [13].

In text independent mode, the speaker can pronounce any speech to be recognized. In this case, there is no constraint on the speech pronounced or on the language used.

In [14], Besacier proved that, the performance of systems in text dependent mode is more important than the performance of systems in text independent mode due to the linguistic variability. Obviously, the priori knowledge of the voice message makes the task of identification systems easy and better performances. However, in the case of systems with databases in large vocabulary, the performance of systems in text dependent or text independent mode are practically the same.

2.3. The Limitation On The Robustness Of ASR Systems

In computer science, robustness is defined as the ability of a system to work correctly in the presence of invalid inputs or abnormal conditions. We briefly present some variability problems which limit the robustness of ASR systems: The variability due to the speaker and the conditions of registration.

2.3.1. Variability due to the speaker

Individual variations between speakers called inter-speaker variation have two main origins: First, the phonation characteristics are different for each speaker independently of the pronounced sentences. Then the same sentence is not pronounced in the same way by two speakers; differences are observed in elocution rates, in pitch variation range or even differences related to their backgrounds. Individual variations of the speaker himself called intra-speaker variation due to several factors such as pathological factors like tiredness, colds, stress or emotional factors [15].

2.3.2. Variability due to registration conditions and transmission

Registration support such as telephone causes the speech quality degradation due to the limitation of useful band and the distortion of transmission channel [16]. In [17], Reynolds proved the identification performance degradation from 99.7% in the TIMIT corpus (Texas Instruments Massa chusetts Institute of Technology) to 76.2% of NTIMIT corpus (Network TIMIT) for 168 speakers. In [18], VanVuuren proved the problems caused by differences between telephone environments. Thus, when the training data and the test data don't come from the same telephone environment, the degradation of the speaker identification performance is very important.

2.4. Modules Of ASR System

The ASR system is composed of three main modules: Parameterization, modeling and decision.

2.4.1. Parameterization

This is the first step of the ASR process, it is to extract characteristic parameters of speaker. These parameters are used to discriminate a speaker from others which reduces information redundancy and quantity.

The choice of parameterization technique is very important for speaker recognition system, because it determines the effectiveness of generated systems. The signal representation of the Cepstral Coefficients is a common task in ASR field. In this theme, the MFCC (Mel Frequency Cepstral Coefficients) parameters are referenced parameters [19].

The calculation of MFCC parameters uses a non-linear frequency scale that takes into consideration the characteristics of the human ear [20].

MFCC parameters are obtained by the signal frequency analysis and the use of filter banks that allow bringing closer the extracted information from that perceived by a human ear. The main steps of MFCC calculation are described in Figure 2.4.

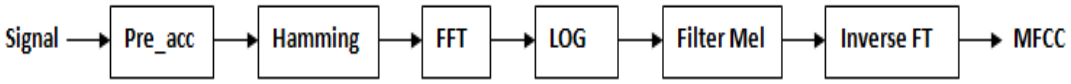

Figure 2.4. MFCC calculation steps.

The calculation process begins with windowing the signal into frames, and the steps to get MFCC are successively applied to those frames.

The steps are:

- The phase of pre_accentuation aims to enhance the spectrum high frequencies. This operation is given by the following equation, where S is the input signal.

$$S(i) = S(i+1) - 0.96 * S \tag{2.1}$$

- To reduce the spectral distortion, Hamming window is applied to the signal. The hamming function gives a good signal representation in windowing field and strongly reduces convolution effects.
- To calculate the Cepstral coefficients, we need to move from the temporal domain (signal) to the frequency domain (spectrum). For this, we use Fourier transforms.
- To simulate a human ear, filtration frequently following nonlinear Mel scale of the spectrum logarithm is applied.
- Applying an inverse Fourier transform of the portions, we obtain the Cepstral coefficients (MFCC).

In addition to extracting acoustic parameters, other additional operations may be added during the parameterization model such as:

- Voice activity detection. It's a technique used in speech processing in which the presence or absence of human speech is detected. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session. [21].
- Acoustic vectors normalization. This task aims to increase the system robustness by reducing the gap between conditions of observation during learning phase and test phase.

2.4.2. Modeling

It's to build a reference model for each speaker using the characteristic parameters extracted during parameterization phase. The modeling techniques are divided into two approaches: Generative approach and Discriminative approach. In this study we used generative approach which also called "modeling approach. The basic idea of this approach is to generate a reference model from the observed data which allow constructing a decision rule. The most generative approaches used in ASR are: Hidden Markov Models (HMM) in text dependent mode and Gaussian Mixture Models (GMM) in text independent mode.

2.4.3. Classification

It is to use one of discriminative approaches for identifying to which of a set of categories a new observation belongs. In ASI application, the decision specifies that the speaker is finally identified, whereas for the ASV application, the decision is a rejection or acceptance of tested speaker. The computing cost of this phase, increases linearly with the number of speakers.

2.5. Example Of Corpus

It's important to underline that ASR systems evaluation depends on the corpus used. Different corpuses have been conceived to measure ASR system performance. Table 2.1. gives an ASR corpus overview [22] [23] [24] [25] [26] [27].

Table 2.1. Example of corpus.

Corpus 1 : «TI-DIGITS »

– Year: since 1982
– Language: English
– Number of speakers: 326
  – 111 Men
  – 114 Women
  – 50 Boys
  – 51 Girls
– Type: noiseless, paying
– Each speaker pronounced 77 digit sequences

Corpus 2: «TIMIT »

– Year: since 1989
– Language: English
– Number of speakers: 630
  – 438 Men
  – 192 Women
– Type: noiseless, paying
– Each speaker pronounces 10 records

Corpus 3 : «NTIMIT »

– Year: since 1993
– Language: English
– Number of speakers: 630
  – 438 Men
  – 192 Women
– Type: noisy, paying
– Each speaker pronounces 10 records
– NTIMIT was collected from the transmission of all TIMIT records by a telephone line

Corpus 4 : «YOHO »

– Year: since 1994
– Language: French
– Number of speakers: 120
  – 55 Men
  – 65 Women

Table 2.2. Example of corpus (continued).

– Type: noiseless, paying – Each speaker pronounces 24 records
Corpus 5: «POLYVAR »

– Year: Since 1997
– Language: French
– Number of speakers: 143
  – 85 Men
  – 58 Women
– Type: noisy, paying
– Each speaker pronounces 10 records

Corpus 6: «SAAVB»
– Year: since 2002 – Language: Arabic – Number of speakers: 1033 – Type: noisy, paying – Each speaker pronounces 59 records

2.6. Applications Domains

In this section, we give some examples of ASR applications; they are grouped into three main categories: Applications on geographic sites, juridical applications and telephone applications [1].

2.6.1. Applications on geographical sites

This category concerns the applications on a particular geographic site; they are mainly used to limit an access of private places.

For examples:

- Automatic locking: It is an electronic lock application used to protect the access of a house, garage, building, etc.
- Transaction validation on website (such as additional control of banking distributors).

- Access to the factories private places: which in general are reserved for employees, workers and inspectors in order to protect production and materials confidentiality.

The advantages of these types of applications are:

- The environment is easily controllable.
- The speaker verification has a deterrent role.
- Speech recognition can be combined with other identity recognition techniques (e.g.: face analysis, fingerprints, etc….).

2.6.2. Telephone applications

This type of applications uses the telephone as communication medium equipment between human and machine. It's the most important category because it allows to verify or to identify the speaker within long distance. There are several applications of this category:

- Validation of banking transactions by phone (to improve the banking service, as well as to validate legally the completed transaction).
- Access to databases for more security and protection (ex: email consultation, consultation of answering machine, etc....).

Disadvantages of these types of applications are:

- It's very difficult to control the environment because the quality of the telephone lines can vary considerably from a call to another, as well as the background noise produces by the calling place (restaurant, office, etc....).
- Applications require to store the data in a centralized way.
- It's impossible to use other recognition techniques (except a digital code typed on touchstones').

2.6.3. Juridical applications

These application domains are currently the ones which pose the most problems. Speaker recognition is used for example, in:

- The investigations orientation.
- The evidence constitution during a trial.

In juridical applications there are more disadvantages than advantages:

- The amount of speech provided is generally very limited.
- The environmental conditions are very bad.
- Involved speakers are rarely cooperative.

2.7. Conclusion

In this chapter, we reviewed the state of the art of ASR system, regrouping main terminologies and concepts. We also presented the general structure of ASR system and its components. A set of corpus and application areas of this system are listed in the last section of this chapter.

CHAPTER 3. DISCRIMINATIVE APPROACHES

3.1. Introduction

Since the introduction of discriminative methods in pattern recognition field, they have given rise to new researches. It's in this context that fits our study, by adapting some discriminative methods in Automatic Speaker Identification(ASI) field. In the chapter, we presented first three discriminative methods of classification: SVM, K-NN and NB.

The effectiveness of these methods is limited given the sequential speech nature, particularly in the presence of a large amount of data. The robustness improvement of the applied discriminative methods is carried out by a hybridization based on multi-Gaussian modeling (GMM) which description is presented in the part two of this chapter, and by the combination of these various methods described in the part three of this chapter.

3.2. Classification By Support Vector Machines: SVM

Support Vector Machines (SVM) or Separators with Vast Margins (SVM) are new statistical learning techniques result directly from Vapnik's work in statistical learning theory [28] [29]. SVM is a classification method by supervised learning, well adapted to process data with very high dimension such as images, speech, etc… Since the introduction of SVM in pattern recognition field, several studies have been able to demonstrate the effectiveness of these techniques mainly in signal processing[6][7].

3.2.1. SVM principle

The principle of SVM, presented by the figure 3.1. consists in projecting data of input space (data belonging to two different classes) non-linearly separable in a space of greater dimension called space of characteristics in the way that data become linearly separable. In this space, an optimal hyperplane separating the classes is constructed such that:

- The vectors belonging to different classes are located on other sides of the hyperplane.
- The smallest distance between vectors and the hyperplane (the margin) is maximal.

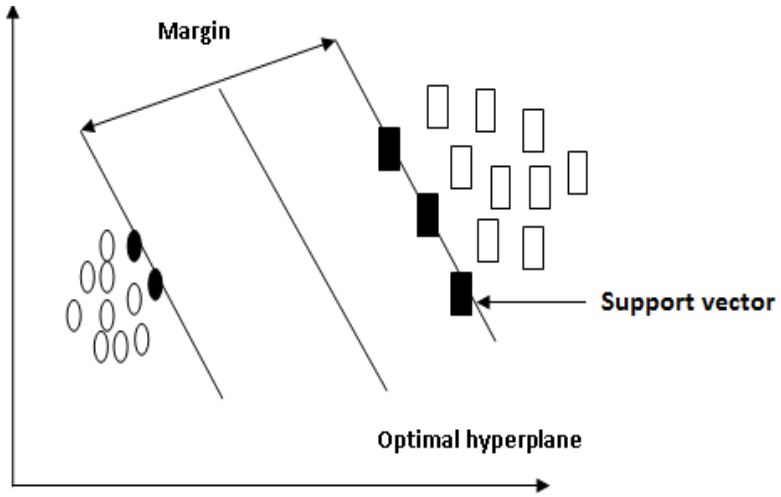

Figure 3.1. Example of optimal hyperplane for a binary classification.

By giving the basic example $D = \{ (x_i, y_i) \in R^d \text{ for } i = 1,...., m \}$ which is a data set where x_i represents an observation of R^d and y_i associated decision which is assumed binary. The SVM purpose is to search an optimal hyperplane of equation:

$$H(x) = w^T + b = 0 \text{ where } x, w \in R^d \text{ and } b \in R . \quad (3.1)$$

Two cases are possible depending on whether data is linearly separable or not. A classifier is called linear when it is possible to express its function decision by a linear function in x .

In case of linearly separable data, the optimal hyperplane is the solution of the following optimization problem:

$$\begin{cases} Min \frac{1}{2}\|w\|^2 & (3.2) \\ y_i\left(w^T x_i + b\right) \geq 1 \quad \forall i = 1,...,m & (3.3) \end{cases}$$

Figure 3.2 provides a visual representation of optimal hyperplane in case of linearly separable data.

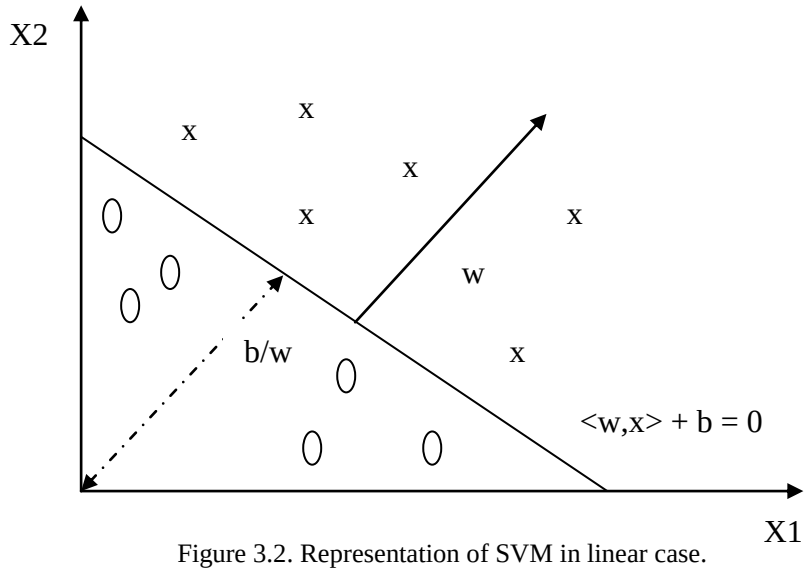

Figure 3.2. Representation of SVM in linear case.

In case of non-linearly separable data, the optimal hyperplane is the one that satisfies the following conditions:

- The distance between correctly classified vectors and optimal hyperplane must be maximum.
- The distance between misclassified vectors and optimal hyperplane should be minimal.

To formalize those conditions, we introduced the distance variables called gap variables ξ_i, where $i = 1,...,m$. These variables represent the distance which separates an example incorrectly classified to the hyperplane of its corresponding class.

Those variables transform the inequality as follows:

$$y_i\left(w^t x_i + b\right) \geq 1 - \xi_i \quad \forall i = 1,...,m \tag{3.4}$$

The objective is to minimize the following function:

$$Min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \tag{3.5}$$

Where C is a tolerance parameter for SVM to control the trade-off between maximizing the margin and minimizing the classification errors committed in the training set.

A second technique used to overcome the problems of non-linearly separable data is the use of kernel function allowing passage to a large space in which linear separation is possible.

The SVM operates by transforming data ϕ of the original space R^d into the space E of more higher-dimensional space. Thus the linear SVM algorithm applied to data $\phi(x)$ in space E produces uneven surfaces decision in the departure space.

Originally, the SVM have been designed primarily for the binary classification. Different methods have been proposed based on the idea of constructing a multi-class classifier combining several binary classifiers. Among these methods, we mention the approach "one against all" and "one against one".

The first approach is Q SVM learning which separates each class of all the other classes; with Q the number of classes.

In the second approach, we make the learning of Q (Q-1) / 2 SVM which each one separate a pair of classes[29].

3.2.2. SVM in automatic speaker recognition

Since the SVM emerged in 1995 [28], several researchers in pattern recognition field began to be interested on it. The first attempt to use SVM in speaker recognition was made by Schmidt in 1996 [2] [30].

In this application, Schmidt has used the frames obtained in parameterization phase as input vectors for SVM. The results obtained are encouraging but not sufficiently reliable.

After this first attempt, other laboratories were interested in these techniques such as IBM [31]. The system they proposed uses the SVM as additional system of decision support which comes into action only when the score obtained by the basic system using GMM modeling and Log Likelihood Ratio (LLR) is not reliable.

More recently, SVM hybridized with GMM modeling have made a breakthrough among the most effective methods in ASR; These works [6], [7], [32], [33], [34] have marked a step in SVM systems progression.

3.3. Classification By K-Nearest Neighbors: K-NN

The K-Nearest Neighbors (K-NN) algorithm is one of the simplest algorithms of automatic supervised study. Fix and Hodges are at the origin of K-NN approach [3]. It's a method based on the memory, which contrary to other statistical methods, doesn't require to adjust the model. Its principle is quite simple, but its implementation requires high computing resources.

3.3.1. K-NN principle

The principle of this classification algorithm is very simple: it's to provide for this algorithm a set of training data $D(x_1, x_2, ..x_n)$, a function of distance d and an integer k . For any new point of test $x \in R^n$ for which it must take a decision, the algorithm searches in D the k nearest points of x in function of the distance d , and assigns to x the class which is commonest among its neighbors.

The fact to consider in general case neighbors k , rather than the single nearest neighbor allows a certain robustness to labeling errors.
The basic K-NN algorithm:

```
Start
    For each (example x ) do
        Calculate the distance D(x , x_n)
    End
    For each (x_n ∈ K-NN (x))do
        Count the number of occurrences of each class
    End
    Attribute to x the most common class;
End
```

Figure 3.3. presents the principle of K-NN with k =3, left side before classification, right side after classification.

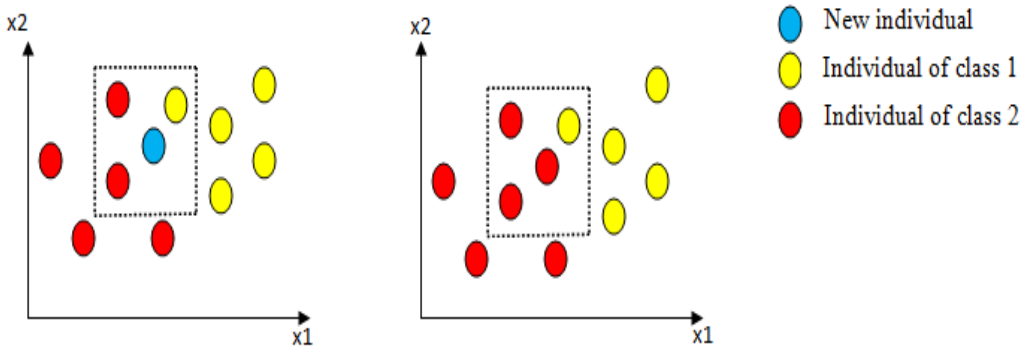

Figure 3.3. Principle of K-NN.

The parameter k must be determined by the user: $k \in \mathbb{N}$. In binary classification, it is helpful to choose odd k to avoid equal votes. The best choice of k depends on the dataset.

If k = 1:

- Borders of classes are very complex.
- Very sensitive to fluctuations in data (high variance).
- Risk of over-adjustment.
- Poor resistance to noisy data.

If k = n:

- Hard Border, constant prediction.
- Risk of over-learning.

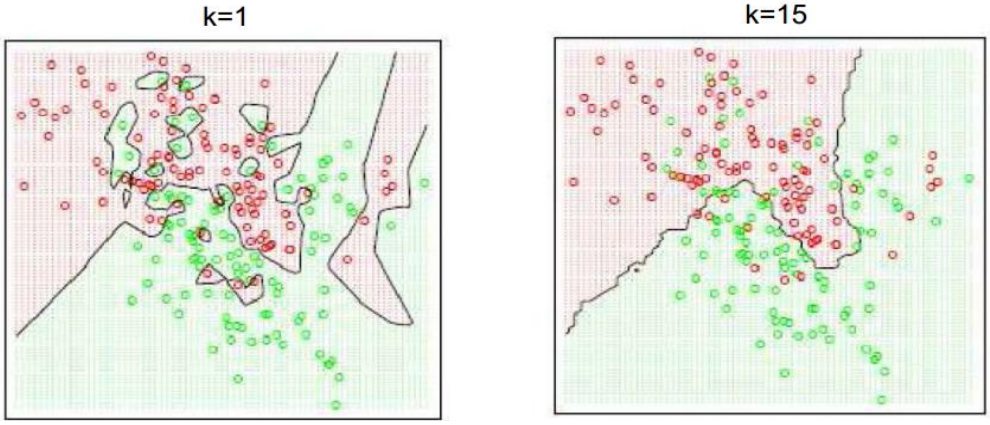

Figure 3.4. Effect of k on class boundaries.

3.3.2. The K-NN in automatic speaker recognition

The k-nearest neighbors have been successfully applied to the ASR, in protocols involving small corpus [10]. View the advantage of being very simple and effective, several researchers in ASR domain have been interested with this classification method [8][35][36].

3.4. Classification By Naive Bayes: NB

Bayesian networks have been the subject of several studies. They constitute an original proposal for automatic extraction of semantic concepts. These networks have already made their proof in several domains related in reasoning and learning. They are the result of the combination between the theory of graphs and probabilities which makes them natural and intuitive tools to treat complex and uncertain data. Indeed, their great capacity of modeling conditional dependencies between objects, allow representing the recognition in a simplified manner, visual and quantitative.

Bayesian Networks are directed and acyclic graphs where recognitions are represented in the form of variables. Each variable is a graph node that takes its values in a discrete or continuous set. The directed arcs represent links of direct dependency expressing in mostly the causality relations between network variables. Their powerful and flexible formalism favored their introduction in several domains of research.

Some Bayesian Networks have been designed for classification problems, the best-known, are those based on the model known as "Naive Bayes". This last constitutes a very simplistic modeling of supervised classification problem.

This model is easy to implement and has proven its effectiveness in many applications. For example, in [37], Spiegelhalter used this discriminative model in medical environment and it has been incorporated to electronic customers mails of Renom.

In [38], Sebe, Lew, Cohen, Garg have used this model to detect the emotion from the image of a person's face. In [39] Or, Zhou, Feng and Sears have used this method to automate the error detection of a speech recognition system.

3.4.1. Naive Bayes principle

NB is based on Bayes' theorem expressed by:

$$P(H \mid D) = \frac{p(H)p(D \mid H)}{p(D)} \tag{3.6}$$

In this equation, we want to calculate $P(H \mid D)$, the posterior probability of the hypothesis H, knowing the data D where:

- $p(H)$: the prior probability of the hypothesis H,
- $p(D)$: the probability of data D
- $P(H \mid D)$: the likelihood of the data D under the hypothesis H.

For a classification task, D represents the data to classify and H corresponds to a hypothesis of class. In other words, for a given x_i, the posterior probability that x_i belongs to the class C_j is estimated by:

$$P(H = C_j \mid D = x_i) = \frac{p(H = C_j)p(D = x_i \mid H = C_j)}{p(D = x_i)} \tag{3.7}$$

In that case, we try to identify the class to which belong x_i. We shall keep then the one which maximizes $P(H = C_j \mid D = x_i)$. This can be formulated as follow:

$$\hat{C}_j = \arg\max_{C_j} \frac{p(H = C_j)p(D = x_i \mid H = C_j)}{p(D = x_i)} \tag{3.8}$$

Since $P(D = x_i)$ doesn't depend on C_j, we can simplify the above equation:

$$\hat{C}_j = \arg\max_{C_j} \; p(H = C_j)p(D = x_i \mid H = C_j) \tag{3.9}$$

The data x_i is generally presented in form of elements vector. Each attribute of this vector corresponds to a characteristic value of x_i. The assignment of x_i to one of the classes depends only on its values. Thus, x_i will be given in the following form: $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$ and therefore we have:

$$\hat{C}_j = \arg\max_{C_j} \; p(H = C_j)p(D = x_{i1}, x_{i2}, ..., x_{ik} \mid H = C_j) \tag{3.10}$$

In Naive Bayes, it is assumed that the attributes of vector x_i are mutually independent. This assumption is not always correct and that is why this method is called naive. However and despite this constraint, Naive Bayes constitutes an effective and efficient method of classification. By adopting this assumption we can write:

$$p(x_{i1}, x_{i2}, ..., x_{ik} \mid C_j) = \prod_{k=1}^{k} p(x_{ik} \mid C_j) \tag{3.11}$$

Thus, the quantity $P(C_j)$ that we seek to maximize corresponds to the probability attached on Bayesian network which the structure is given by the following figure:

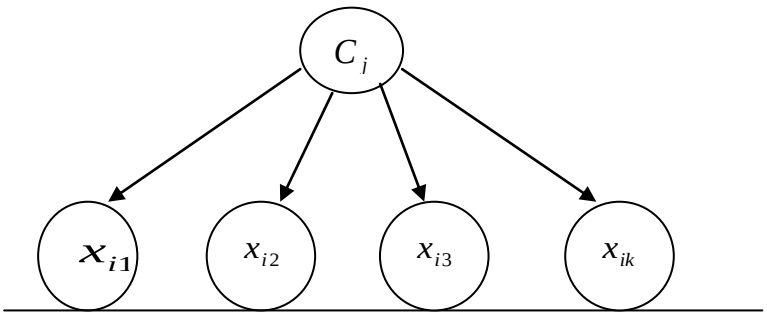

Figure 3.5. The general structure of NB.

In classification task, the learning step is to learn from a corpus labeled the different probabilities $P(C_j)$ and $p(x_{ik} \mid C_j)$. The test step is to look at the class $\hat{C}_j$ which maximizes the product[38].

3.4.2. Naive Bayes in automatic speaker recognition

During the last ten years, the Bayesian networks have become very popular in artificial intelligence due to many advances in various aspects of learning and inference.

For a classification problem, the Naive Bayes structure proved experimentally that it's able to give good results, especially in speaker's recognition [4] [40].

3.5. Modeling Strategy By Gaussian Mixture Models

Modeling speakers by Gaussian Mixture Models (GMM) is the most powerful and most common method for ASR systems in text independent mode [41]. GMM models are used for their ability to model the probabilities distribution of the cepstral coefficients.

3.5.1. GMM structure

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system [16]. The figure 3.6 bellow represents the weighted sum of M Gaussians multidimensional when M = 3.

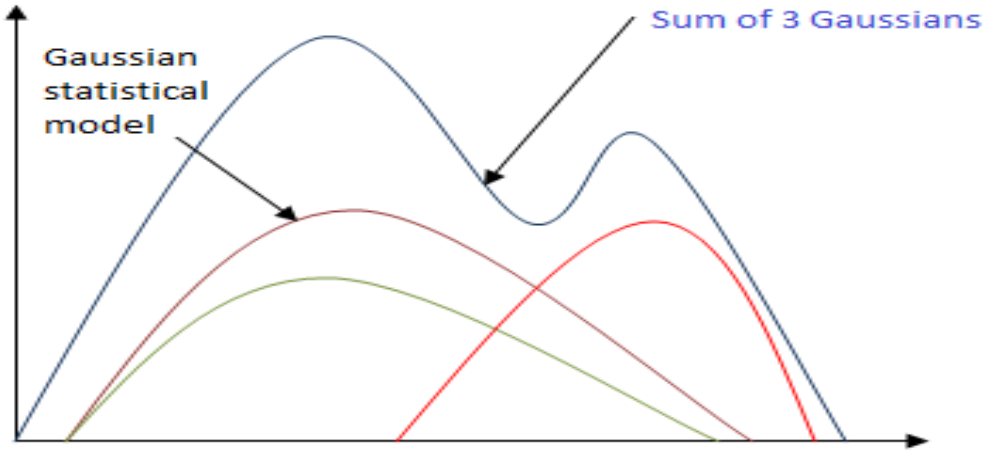

Figure 3.6. Mixture Model with 3 Gaussians.

In literature, each Gaussian g_i is presented by its weight p_i by a medium vector μ_i of dimension d and by a covariance matrix $\sum_i$ of dimension $d*d$. To define the model of a speaker, it is necessary to determine the set of these parameters ($p_i, \mu_i, \sum_i$). Determining the number of Gaussians M is a crucial issue since it constitutes a compromise between complexity and precision [16].

3.5.2. Universal background model construction

The UBM (Universal Background Model) introduced by Carey, Parris [42] and Reynolds [17]; is a generic model with independent speech of the speaker that collects all the training data by representing also the a priori distribution of the whole input acoustic space. Its parametric form is a mixture Gaussian models (GMM).

The initialization of Gaussians is done by a Vector Quantization (VQ), using classification algorithms like K-means or Fuzzy C-means (fcm) [43]. The initialization phase is very important, it allows to avoid the random initialization that can bring learning algorithms trapped toward optimal erroneous premises.

This paradigm gave superior performance to classical methods (for example, vector quantization). This model is learned by maximum likelihood via the EM algorithm.

The Expectation Maximization (EM) algorithm is proposed by Dempster [44]. It allows facing the learning problem with the maximum likelihood criterion. It also provides maximum likelihood of the training data (estimated by the equation below) compared to GMM model of UBM references.

$$\hat{\lambda}_x = \arg\max(x \mid \lambda) \tag{3.12}$$

This algorithm is based on two steps [45]:

- The expectation step which consists of determining the posterior probabilities that Gaussians have generated from the learning frames.
- The maximization step which consists of updating the model parameters in order to maximize the selected criterion.

This iterative algorithm converges to a local maximum. It introduces advantage to augment the likelihood of data with the estimated model at every iteration until the pseudo-stability [44].

3.5.3. Maximum a posteriori adaptation (MAP)

The adaptation MAP (Maximum A Posteriori) has been introduced in ASR field and more specifically in ASI by Reynolds [41]. It's to define a priori distributions $p(\lambda)$ for the reference model parameters (Universal Background Model that represents all the acoustic parameters of all speakers) and maximize the a posteriori probabilities $p(\lambda \mid x)$ on the training signal x (for obtaining the speaker model which the training signal is x).

The MAP adaptation is based on two steps [41]:

- The first step is to determine the statistical parameters for each Gaussian reference model based on the training data. In practical only the averages of the GMM are adapted. In [41], REYNOLDS showed that, by adapting just GMM

averages, the performance loss is negligible compared to the adjustment of all parameters that are costly in terms of time.

- The second step aims to combine those new parameters with the Universal Background Model (UBM) parameters by using weighted coefficients that depend on the training data. In fact, the Gaussian components are heavily weighted within estimation of final model parameters in case of a big amount of training data, and they are weakly weighted in the opposite case.

The output of this adaptation is a set of vectors called supervector. It's a vector concatenating the parameters of a statistical model which contains all the average parameters of GMM model [46].

Those supervectors will be provided as input to the different classifiers (SVM, K-NN, NB) in speaker identification case with GMM modeling.

3.6. Classifiers Combination Approaches

The combination of classifiers is a new strategy to integrate multiple information, derived from the same original source or from different sources in classification process by exploiting the best characteristics of each source . This strategy proved its aptitude to conceive and to set up powerful systems [47].

There is two approaches of combining classifiers: Sequential and parallel approaches.

3.6.1. Sequential approach

The sequential combination approach, also known as series combination is organized into successive levels of decisions. The methods are co-operated in series, ones after the others. The results found by one or more classifiers are used for the execution of other classifiers. In that case, the methods execution order is important.

A change of execution order causes also a change of the final result [48]. Figure 3.7 provides a representation of classifiers sequential combination.

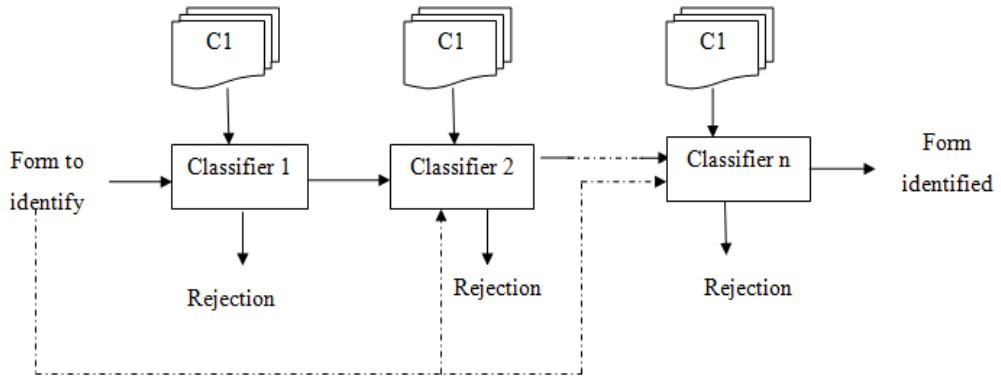

Figure 3.7. Sequential Combination of classifiers.

3.6.2. Parallel approach

In the parallel approach, the classifiers operate independently and then merge their respective results. This merger is made in democratic or directed way; the democratic merger way does not promotes any classifier over another, but for the directed merger way, a weight is attributed to the result of each classifier according to its performance. The execution order of the classifier doesn't intervene in this approach [48]. Figure 3.8 provides a representation of classifiers parallel combination.

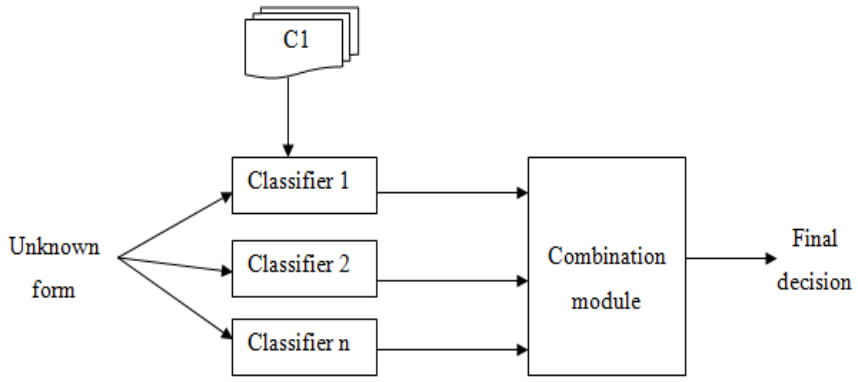

Figure 3.8. Parallel Combination of classifiers.

3.7. Some Results Of Classification In Literature

Table 3.1 shows examples of classification results in literature by using various discriminative approaches of classification.

Table 3.1. Some results for automatic speaker recognition in literature.

Authors	Corpus	approaches	parameters	TI (%)
[Trabelsi, 2011] [7]	TIMIT	SVM	12 MFCC coefficients + 3 energy coefficients + 26 dynamic coefficients	15
[Lefevre, 2000] [49]	TIMIT	K-NN	12 MFCC coefficients	49
[Trabelsi, 2011] [7]	TIMIT	GMM-SVM	12 MFCC coefficients + 3 energy coefficients + 26 dynamic coefficients	98
[Didé, 2007][50]	360 audio samples	ACP, K-NN	13 MFCC coefficients	84
[Amami and al., 2012] [51]	TIMIT (20 phoneme)	NB	12 MFCC coefficients + 24 dynamic coefficients	43

3.8. Conclusion

This chapter was devoted to the presentation of different discriminative classification methods used in this study. We presented three methods of classification : SVM, K-NN and NB. We also presented the contribution of modeling by Gaussian mixture models. In the last part, we presented different combination methods of classification and some results of automatic speaker recognition in literature.

CHAPTER 4. EXPERIMENTAL STUDY AND RESULTS

4.1. Introduction

The main objective of this chapter is to evaluate the effectiveness and robustness of different speaker identification approaches already described in the previous chapter. First, we describe the experimental protocol of our study, then we evaluate all classifiers (SVM, NB and K-NN) while studying the impact of some parameters on their performances. We evaluate also the hybridization of GMM with these classifiers. In order to design and implement powerful system, a combination of the hybrid classifiers has been adopted. Finally; to test the robustness of the hybrid systems, we generate different random noise in the TIMIT database used during this study.

4.2. Background And Experimentation

4.2.1. Development environment

The implementation of our study is performed under MATLAB version R2011b. For the parameterization, we appealed to the toolbox voicebox, for the SVM classification, we have appealed to the LIBSVM library, for GMM modeling, we appealed to the bookstore NETLAB. Execution was accomplished on Laptop Intel (R) Core(TM) i5-3210M CPU @ 2.50 GHz with 8GB of RAM, the operating system: Windows 7 and system type: 64-bit. The experiments were performed on the TIMIT corpus of speech whose description is given in the following section.

4.2.2. Description and organization of TIMIT corpus

General organization of TIMIT corpus is presented in the figure 4.1. It's a non-noisy, acoustic and phonetic speech database, dedicated to automatic speech recognition systems. It's composed by eight regional dialects (DR1 to DR8).

The different regions DR (Dialect Region) are:

- DR l : New England
- DR 2: Northem
- DR 3: North Midland
- DR 4: South Midland
- DR 5: Southem
- DR 6: New York City
- DR 7: Westem
- DR 8: Army Bra

These dialects are spoken by 630 speakers from the United States and divided between training set (462 speakers: 326 men and 136 women) and test set (168 speakers: 112 men and 56 women). This database has four types of file extensions which are respectively .wrd, .wav, .phn and .txt. The .wrd and .phn files contain words and phoneme segments of the sentence x_i which textual content is located in .txt file.

In our work, we explored all the basics set of training and test for DR1 dialect (New England). The DR1 is represented by 49 speakers (18 female and 31 male). It regroups 490 sentences in total where every speaker pronounces 10 sentences: the first 8 sentences are used in the training phase and the last 2 sentences are used during the test phase.

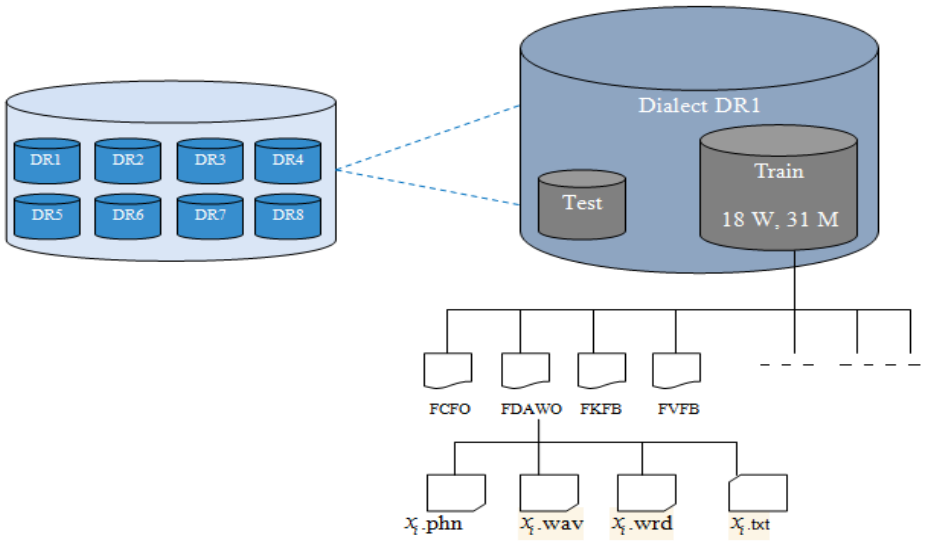

Figure 4.1. General Structure of TIMIT corpus.

4.2.3. Experimental condition

The experimental conditions are defined in the table below.

Table 4.1. Experimental Conditions.

Corpus : TIMIT	
Dialect: DR1	Number of Speaker female: 18 Number of Speaker male: 31 Number of training sentences per speaker: 8 Number of test sentences per speaker: 2 Average training sequence length: 2s Average test sequence length: 1s
Parameterization	
Coefficients: MFCC Sampling frequency: 16 kHz Window length: 16ms Sampling interval: 8ms Windowing: hamming Filters number : 24	

4.3. Evaluation Metrics

To evaluate the performance of the automatic speaker identification system, numerous metrics have been proposed by M. Siu and H. Gish research in 1999[52]. In our study, we used as performance measure the Identification rate (IR). This rate is obtained during the test phase and it's expressed by the following equation:

$$IR(\%) = \frac{\text{Number of speakers correctly Identified}}{\text{Total number of speakers}} \times 100 \quad (4.1)$$

4.4. Application Of Classifiers

We present the results of the different experiments in order to evaluate the behavior of the following classifiers: SVM, K-NN and NB.

The choice of these classifiers is justified due to the fact that discriminative approaches have been able to dominate the state of art of speaker recognition systems. Thus, the selected classifiers are the most used in automatic speaker recognition and gave promising results [50].

4.4.1. SVM identification system

We introduce throughout this section the SVM system architecture used as well as the results of its various performed tests. We studied the impact of some technical parameters on speaker identification rates.

4.4.1.1. SVM System Architecture

The SVM block diagram system is given in Figure 4.2. First of all is to extract the acoustic vectors directly from the sequences pronounced. This has to be done for speakers of training and testing. Dictionaries obtained for each speaker will be

concatenated and presented as input vector to the SVM classification module. We chose to study the linear kernel case of SVM.

During the training phase, SVM system determines the separator functions between the characteristics of different speakers by producing the SVM training model that includes all parameters related to different hyperplanes.

During the test phase, a study of similarity between the characteristics of the speaker to be identified and the characteristics of all speakers is established.

The identified speaker is the one for which the similarity is the greatest.

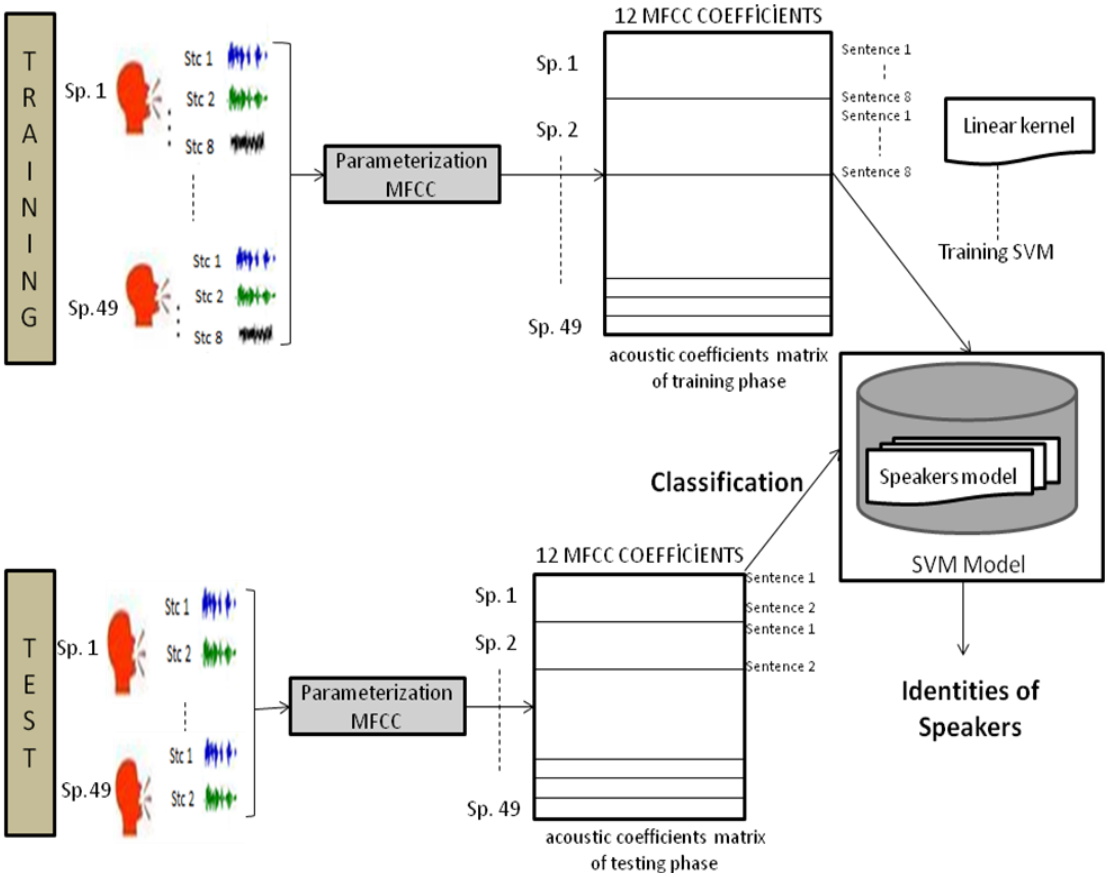

Figure 4.2. SVM system architecture.

4.4.1.2. Impact of MFCC coefficients number for SVM

The parameterization phase is very important for speaker identification because it directly contributes to the performance of the identification system. Tests on MFCC coefficients number of parameters vectors have been processed. Table 4.2. details the variation rates of speakers identification according to MFCC coefficients number, which are respectively 8, 12, 16.

Table 4.2. Impact of MFCC coefficients number on SVM identification rate system.

Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	2%	3%	5%
Execution time	30min10sec	36min32sec	43min55sec

We notice that IR increases proportionally with the increase of MFCC coefficient numbers. The execution time also increases with the increase of MFCC coefficient numbers to reach a maximum value of 43min55sec.

4.4.1.3. Impact of dynamic parameters for SVM

The main goal of extracting parameters, is to model the speech which is a highly variable signal. It is also necessary to resort to the local information on speech signal parameters evolution in time. The first derivative D' (speed) and the second derivative D" (acceleration) are added to the acoustic parameter vectors to model their trajectories in time.

We present in the table below the effect of adding those parameters to the SVM system while varying also MFCC coefficients number.

Table 4.3. Impact of dynamic parameters on SVM identification rate system.

Number of MFCC Coefficients + D' + D"	8	12	16
Identification Rate (IR)	5%	7%	9%
Execution time	39min15sec	44min10sec	56min5sec

We notice that, the addition of dynamic parameters slightly improves rates around 3% to 4% at most. For example, with 16MCCF the rate increased from 5% to 9% by adding the dynamic parameters.

We can explain those rates by the fact that the behavior of speakers has been amended with dynamic parameters. We observed that the system becomes accustomed to the speakers models.

4.4.2. K-NN Identification system

The use of nearest neighbors classification finds its foundation in discriminative analysis. Although that the K-NN algorithm has been successfully applied in the speech recognition field, there is still an essential question of choosing the number (K) of nearest neighbors. The choice of the number (K) may be achieved by using an experimental study [53]. In this section, we present the architecture of K-NN system proposed and the results of tests.

4.4.2.1. K-NN System Architecture

Figure 4.3. bellow shows the structure of K-NN speaker identification system.

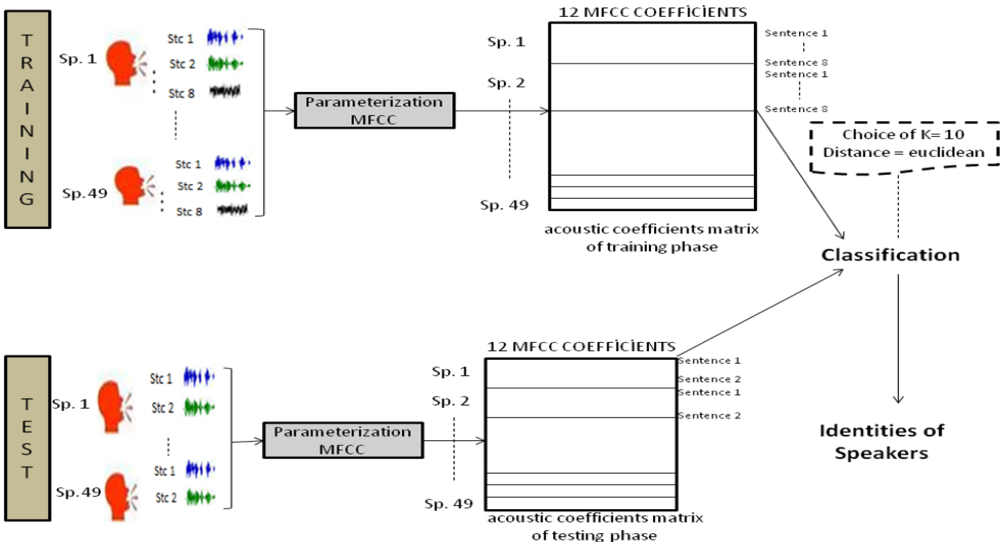

Figure 4.3. K-NN System Architecture.

The K-NN method is very simple, a direct classification is performed for all new speakers based on its Euclidean distance from K-nearest neighbors.

4.4.2.2. Impact of Nearest Neighbors number K

We have studied the impact of nearest neighbors' number on identification rate and on system performance when MFCC coefficients equal to 12. The numbers of nearest neighbors are, respectively: 1, 5, 10, 15, 20. The experimental results are described in table 4.4. below.

Table 4.4. Impact of nearest neighbors number on K-NN identification rate system.

Number of MFCC Coefficients = 12					
Number of K-NN	1	5	10	15	20
Identification Rate (IR)	23%	25%	27%	26%	25%
Execution time	1min38sec	1min39sec	1min45sec	1min39sec	1min39sec

We notice that for a K-NN system with 12 MFCC coefficients, by increasing the number (K) of nearest neighbors improves the identification rate up to 27% for K=10, Beyond this value, the performance of the system decreases.

In the literature is noted that usually the best value of K is the value that approaching the square root of the classes number [53]. For our case, we have 49 classes, the reason why the best rate obtained was for K = 10.

4.4.2.3. Impact of MFCC coefficients number for K-NN

We have accomplished some tests in order to evaluate the system performance. Table 4.5. below gives an overview of different tests made by changing MFCC coefficients number respectively in 8, 12, 16.

Table 4.5. Impact of MFCC coefficients number on K-NN identification rate system.

Number of K-NN = 10			
Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	16%	27%	28%
Execution time	11sec	1min26sec	1min45sec

This table shows that the identification rate (IR) increases proportionally with the increase of MFCC coefficients number at a negligible time; with 8 MFCC, the IR is 16%; with 12 MFCC the IR is 27% and with 16 MFCC the IR is 28%.

4.4.2.4. Impact of dynamic parameters for K-NN

Table 4.6. below presents the results of the various tests carried out by adding dynamic parameters to K-NN identification rate system based on different numbers of MCCF coefficients which are respectively 8, 12, 16.

Table 4.6. Impact of dynamic parameters on K-NN identification rate system.

Number of K-NN = 10			
Number of MFCC Coefficients + D' + D"	8	12	16
Identification Rate (IR)	18%	30%	33%
Execution time	1min10sec	3min25sec	4min8sec

Table 4.6. shows that adding dynamic parameters improves the rate of identification (IR), example with 16 MFCC coefficients, the identification rate without the addition of dynamic parameters was 28%, while appending dynamic parameters, the rate increased up to 33%.

4.4.3. NB identification system

Like others classifiers, we have accomplished various tests for NB classifier, by varying some parameters which are: MFCC coefficients number and dynamic parameters. First, we present NB system architecture.

4.4.3.1. NB System Architecture

Figure 4.4 shows the structure of NB speaker identification system.

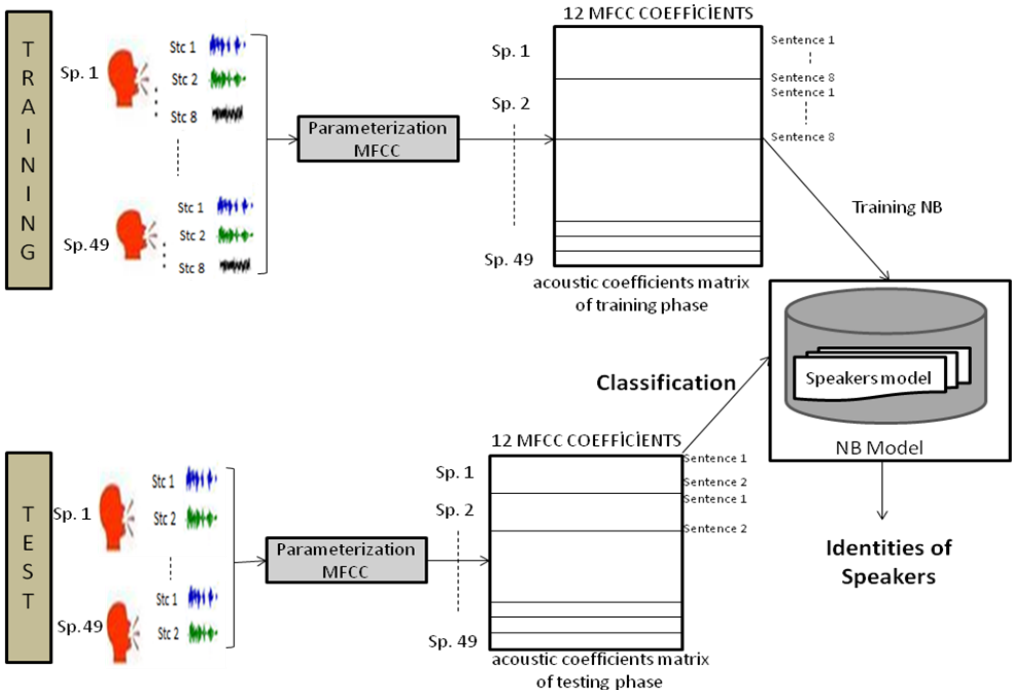

Figure 4.4. NB System Architecture.

A speakers model is constructed for the NB system during the training phase. In test phase, a score of the tested speaker is calculated in order to identify his identity.

4.4.3.2. Impact of MFCC coefficients number for NB

Table 4.7. below presents the variation of the NB identification rate system depending on MFCC coefficients numbers which are respectively 8, 12 and 16.

table 4.7. Impact of MFCC coefficients number on NB identification rate system.

Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	7%	11%	14%
Execution time	1sec	1sec	1.5sec

As is shown in table 4.7, by increasing MFCC coefficients number the identification rates increase around 3% to 4% at most, but these rates are very low and the best rate is just equal to 14%.

4.4.3.3. Impact of dynamic parameters for NB

After studying the impact of varying MFCC parameters number, we test the impact of adding dynamic parameters on NB classifier performance. The results are detailed in Table 4.8. below.

Table 4.8. Impact of dynamic parameters on NB identification rate system.

Number of MFCC Coefficients + D' + D"	8	12	16
Identification Rate (IR)	10%	16%	19%
Execution time	3 sec	4sec	4sec

According to the result table, we notice a slight improvement of identification rate, but the rates are low and the best result is 19 %.
According to the tests carried out, we observed a slight improvement of identification rate each time we tried to increase MFCC coefficients numbers or by adding dynamic parameters for the three systems tested (SVM, K-NN and NB).

4.4.4. Comparative study of different identification systems

In this section, we compare the results obtained from the different tested classifiers (SVM, K-NN, NB). The results of the comparison are presented in figure 4.5. below.

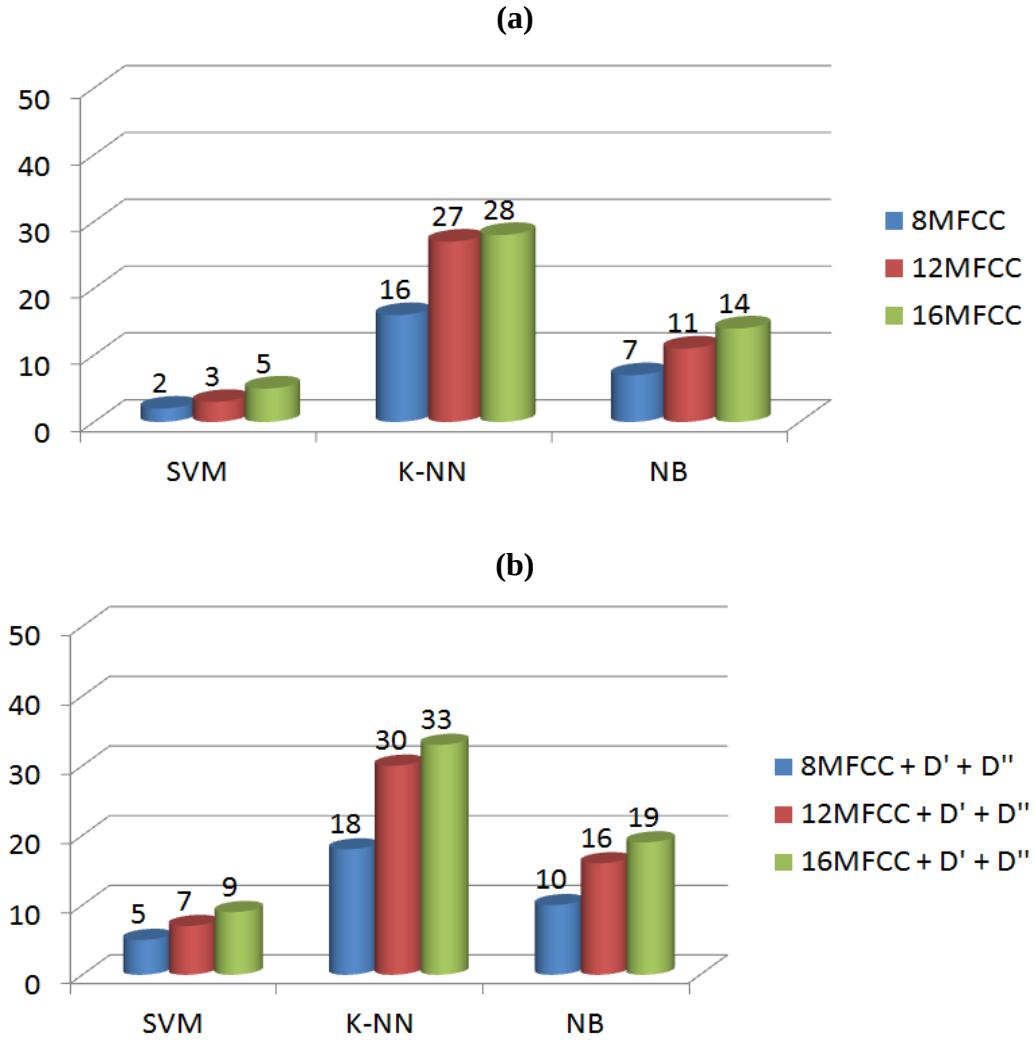

Figure 4.5. Comparative study between different identification systems without dynamic parameters (a) and with dynamic parameters (b).

We clearly notice that the identification rate had a slight increase by adding the dynamic parameters, this increase is around 1% to 7% for all classifiers.

For example, for SVM with 8 MFCC coefficients, the identification rate increased from 2% to 5% by adding dynamic parameters. For NB with 12 MFCC coefficients, the identification rate increased from 11% to 16% by adding dynamic parameters.

Generally, the identification rates are low, between 5% and 33%; SVM and NB present lower identification rates while K-NN provides the average identification

rates. The results of our experiments show that the performance of different systems is generally low especially for SVM and NB.

The performance weakness of the classifiers used in this study is caused by the size of entry matrix which is composed by thousands of frames, what augments the probabilities errors during the classification.

In order to improve the performance of the classifiers, we applied the Gaussian mixture modeling in our study.

4.5. Classifiers Application With GMM Modeling

Hybridization of generative and discriminative approaches for the classification was recently the subject of research in machine learning [54]. It's consist of projecting a wide variety of N-dimensional input signals into fixed dimension vector, using the parameters of generative models. Generative - discriminative hybrid models have been applied successfully in bioinformatics [55], computer vision and audio processing.

The idea of coupling these two approaches aims to combine the advantages of each of them. In particular, some of the intrinsic properties of GMM generative approach are imposed:

- Its ability to model a system without being limited by its complexity.
- Its ability to naturally treat the sequences of varying size.

The expectation maximization (EM), is an another GMM property that explains their success in classification. It's a powerful tool in parameter estimation.

The implementation of GMM approach, and especially the optimization of GMM estimation parameters by EM algorithm constitute the main object of this section.

The block diagram of hybrid approach based on GMM model proposed for speaker identification in text independent mode is given in figure 4.6 below.

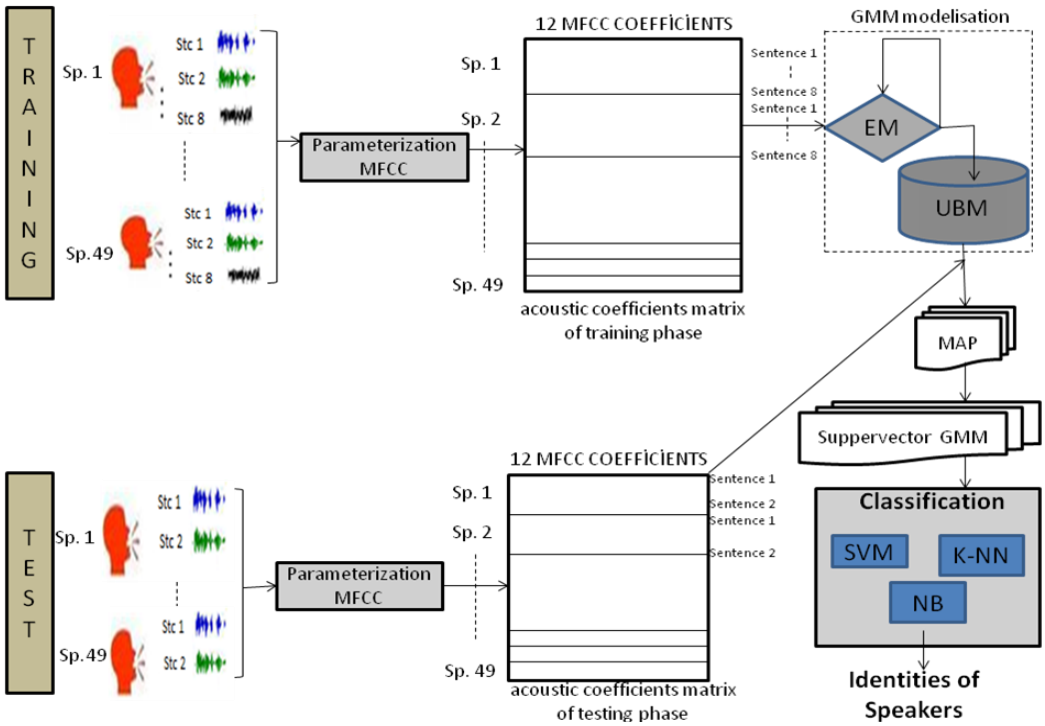

Figure 4.6. ASI System architecture based on GMM generative approach.

The system consists of two steps:

The first step is the training phase and the second step is the testing phase which uses one of the classifiers already described above, in order to calculate the pronounced sequence score of the unknown speaker. In the first step, the training phase is performed in three stages:

The first stage named GMM-UBM consists of creating the universal background model using the EM algorithm, it's a decisive and fundamental step since it represents the basic foundation of the training phase. The second stage is to create a GMM model of each speaker in the database via "instantiation" of the UBM model by using MAP adaptation. From this model, the third stage allows to extract the average supervectors for each speaker and these average supervectors contain the average parameters of the GMM model.

The concept of supervector reveals a transformation of a speech signal to a large dimension vector. This concept fits well with the idea of SVM kernel sequence [16] whose the basic idea is to compare directly a core with two speech signals (not with frames).

In the second step, the supervectors set is built just for the speaker who will be identified. The identity of the unknown speaker is determined following a similarity study between the test supervectors characteristics and the constructed classifier model.

To study the importance of GMM modeling on recognition rate, we have tested and compared multiple systems while varying certain parameters and this depend on the selected classifiers.

In order to ensure a multi-Gaussian modeling of speakers, we adopted diagonal covariance matrices of 128 Gaussians. Firstly, this choice is justified due to the fact that the covariance matrices are generally estimated for ASR under diagonal form. This has the advantage of reducing both the calculation time required for the operations and the modeling complexity [40]. Secondly, a high number of Gaussians may cause over-learning system problem and this may penalize its generalization capabilities.

4.5.1. GMM-SVM hybrid identification system

We present in this section the results of different tests performed in order to evaluate the robustness of the hybrid system GMM-SVM. For that purpose, we studied the impact of some technical parameters on its performance.

4.5.1.1. Impact of MFCC coefficients number

The test results obtained by varying the MFCC coefficients numbers are given in the table 4.9. bellow. The MFCC coefficients numbers are respectively : 8, 12 and 16

Table 4.9. Impact of MFCC coefficients number on identification rate of GMM-SVM system.

Number of Gaussian GMM = 128			
Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	94%	96%	97%
Execution time	4sec	6sec	7sec

We notice that the GMM-SVM hybrid identification system gives better identification rate (IR) of 97% for 16 MFCC, and especially in a very short time. It's also noted that by increasing the number of MFCC coefficients improves slightly GMM-SVM system performance. With 8 MFCC the rate increase from 94% to 96% with 12 MFCC and to 97% with 16 MFCC.

4.5.1.2. Impact of dynamic parameters

To improve the robustness of the GMM-SVM system, we studied its performance by adding dynamic parameters on speakers identification rate. The table 4.10. presents the results of GMM-SVM hybrid system evaluation based on 128 Gaussians and by varying MFCC coefficients numbers which are respectively 8, 12, 16.

Table 4.10. Impact of dynamic parameters on identification rate of GMM-SVM system.

Number of Gaussian GMM = 128			
Number of MFCC Coefficients + D' + D"	24	36	48
Identification Rate (IR)	96%	97%	98%
Execution time	1min	1min25sec	1min46sec

The test results by adding dynamic parameters show a slight improvement on system performance compared to the GMM-SVM system without the addition of dynamic parameters. For example, with 16 MFCC there is an increase of identification rate from 97% to 98% but this increase cause an augmentation of the execution time from 7sec to 1min46sec.

4.5.2. GMM-K-NN Hybrid identification System

We present in this section the results of different tests that we performed for GMM-K-NN hybrid system. We studied the impact of some technical parameters on identification rate, starting by the number of nearest neighbors K.

4.5.2.1. Impact of Nearest Neighbors numbers

The impact of nearest neighbors number (k) on identification rate is presented by The figure 4.7. bellow. The Numbers of nearest neighbors (k) , are from 1 to 20. The test results are described by the following curve.

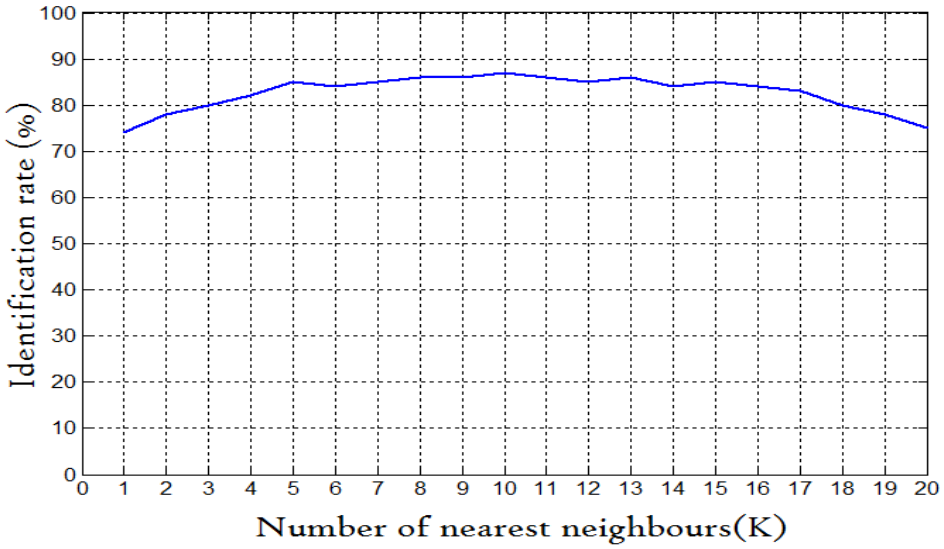

Figure 4.7. Impact of nearest neighbors number (k) on GMM- K-NN Identification rate system.

According to the curve, we notice that the identification rate augments proportionally with the increase of K to reach 87% for K = 10, from K > 10 the system performance decreases. In our study, we used k = 10 for all tests we performed.

4.5.2.2. Impact of MFCC coefficients numbers for GMM-K-NN

The test results by varying MFCC coefficients numbers are provided by table 4.11. below.

Table 4.11. Impact of MFCC coefficients number on identification rate of GMM-K-NN system.

Number of Gaussian GMM = 128 Number of nearest neighbors (k)= 10			
Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	85%	87%	90%
Execution time	1sec	1sec	1sec

The GMM-K-NN hybrid system identification rate increases proportionally with the increase of MFCC coefficients number and this in a very short time. Example with 8 MFCC the rate increases from 85% to 90% with 16 MFCC.

4.5.2.3. Impact of dynamic parameters for GMM-K-NN

After studying the impact of varying the number of MFCC parameters , we test the impact of adding dynamic parameters on the performance of GMM-K-NN hybrid system. The results are given on the table 4.12. below.

Table 4.12. Impact of dynamic parameters on identification rate of GMM-K-NN system.

Number of Gaussian GMM = 128 Number of nearest neighbors (k)= 10			
Number of MFCC Coefficients + D' + D"	26	36	48
Identification Rate (IR)	64	71	72
Time	2sec	4sec	12 sec

In contrast with other classifiers, by adding dynamic parameters on GMM-K-NN hybrid system; we notice that there is a decrease of identification rate. For example with 16 MFCC the identification rate decreased from 90 % to 72 % by adding dynamic parameters.

4.5.3. GMM-NB Hybrid identification System

We present the results of different tests we performed for GMM-NB hybrid system in order to evaluate its performance while diversifying some parameters.

4.5.3.1. Impact of MFCC coefficients numbers for GMM-NB

The table 4.13 below represents the test results of GMM-NN system by varying the number of MFCC coefficients which are respectively: 8, 12 and 16.

Table 4.13. Impact of MFCC coefficients number on identification rate of GMM-NB system.

Number of Gaussian GMM = 128			
Number of MFCC Coefficients	8	12	16
Identification Rate (IR)	95%	92%	88%
Time	1sec	1sec	2sec

According to the table 4.13. above, GMM-NB hybrid system presents a significant identification rate that reaches 95% for 8 MFCC in a negligible time of 1s. In contrary of others classifiers, by increasing MFCC coefficients number, the identification rate (IR) decrease .

4.5.3.2. Impact of dynamic parameters for NB

The table 4.14. bellow presents the impact of adding dynamic parameters on GMM-NB hybrid system performance.

Table 4.14. Impact of dynamic parameters on identification rate of GMM-NB system.

Number of Gaussian GMM = 128			
Number of MFCC Coefficients + D' + D"	8	12	16
Identification Rate (IR)	65%	75%	76%
Execution time	2sec	8sec	14sec

The addition of the dynamic parameters generates a significant decline of rates; with 8 MFCC the rate decreases from 95% to 65%, with 12 MFCC the rate decreases from 92% to 75% and with 16 MFCC the rate decreases from 88% to 76%.

4.6. Comparative Study Of Different Hybrid Identification Systems

We compare the best results obtained from different tested classifiers without (or with) addition of dynamic parameters. The comparison results are presented in figure 4.8. below.

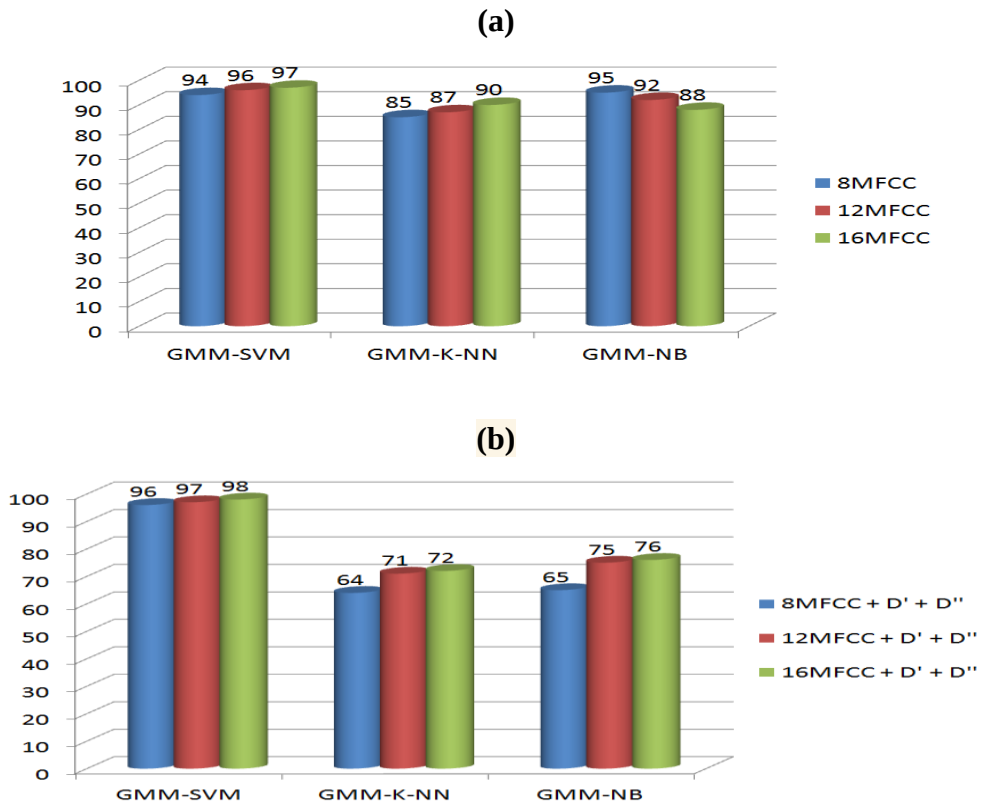

Figure 4.8. Comparative study between different hybrids identification systems without dynamic parameters (a) and with dynamic parameters (b).

For SVM classifier , we notice that the identification rate had a slight increase around 1% to 2% by adding the dynamic parameters while for K-NN and NB by adding dynamic parameters the identification rate decreases. Example for NB with 12 MFCC, the identification rate decreases from 92% to 75%.

Generally, the results of our different performed tests show that the performance of the hybrid systems (GMM-SVM, GMM-K-NN, GMM-NB) are good. The identification rate vary between 87% and 97%, with GMM-SVM which provided the highest identification rate.

4.7. Robustness Of Hybrid Systems

As we used a noiseless database TIMIT in our work, and after getting good results with hybrid systems, we tested their robustness by introducing randomly different noise and running the system several cases. The noises have been introduced respectively with SNR of -20dB to 20dB. The figure below 4.9. represents the plots of the performance of hybrid systems with noisy data.

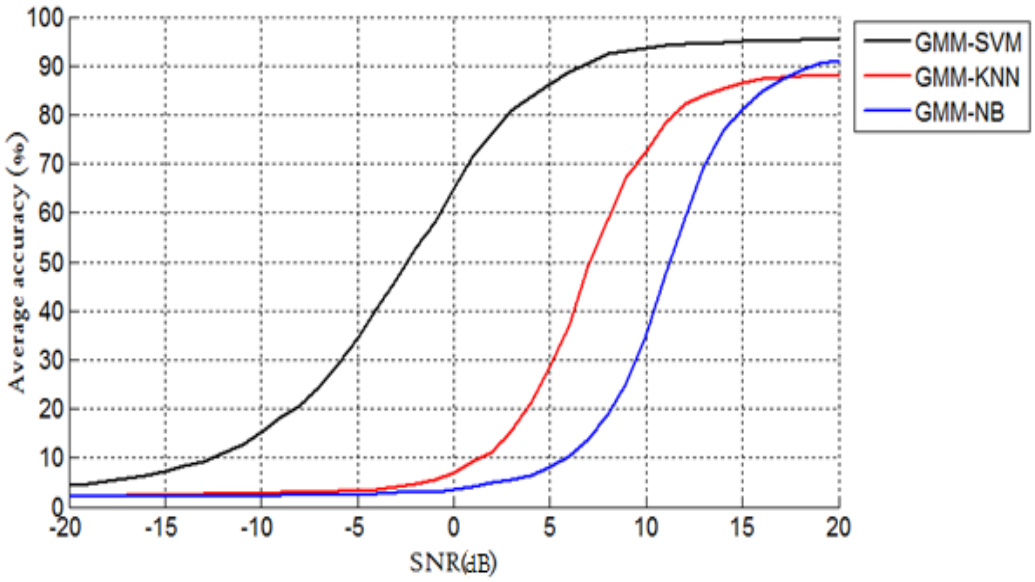

Figure 4.9. Performance of hybrid systems with noisy data.

As is shown by the figure 4.9 by introducing different generated noise in the database, our hybrid systems are robust and can resist to noise problem. Even if in the interval of -20dB to 5dB the classification rates are too much lower for other hybrid classifiers, the GMM-SVM hybrid system showed its robustness by giving a good result that can even reach an average accuracy of 88% at 5dB of SNR.

4.8. Combination Of Hybrid Systems

In order to exploit the best of each hybrid system characteristics, we propose to combine their decisions. For parameterization we used 12 MFCC coefficients for all classifiers which, according to our accomplished experiments gave the best identification rate in a short time. And for GMM Gaussians we used 128 for all test.

4.8.1. Combination architecture

The combination principle is shown in the following figure 4.10. This system is composed with two main steps:

The first step is the classification where each hybrid classifier (GMM-SVM, GMM-K-NN, GMM-NB) operates independently. The decision of all these classifiers are combined via majority vote mode. For this mode of combination, the output of each method is considered as a vote for a class. The number of votes for each classes is counted. The class having maximum votes will be selected. [56].

In the second step, the decision of each hybrid system is considered independent to others.

All possible combinations of the three hybrid systems have been tested and they gave four different systems as follows:

System 1: GMM-NB + GMM- K-NN
System 2: GMM-KNN + GMM-SVM
System 3: GMM-SVM + GMM-NB
System 4: GMM-KNN+ GMM-NB + GMM-SVM

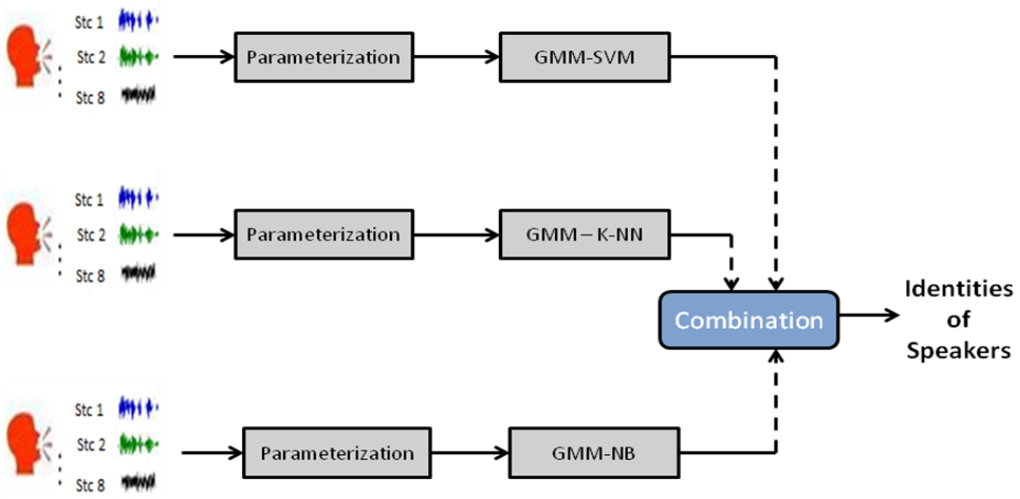

Figure 4.10. Hybrid systems combination architecture.

4.8.2. Evaluation of results

The table 4.15. present the results given by the different possibilities of the implemented combination systems.

Table 4.15. Results of different strategies of hybrid systems combination.

Systems		Identification rate (IR) (%)
System 1	GMM-NB + GMM-K-NN	97%
System 2	GMM-KNN + GMM-SVM	96%
System 3	GMM-SVM + GMM-NB	98%
System 4	GMM-KNN+ GMM-NB + GMM-SVM	100%

From Table 4.15. we notice that identification rates vary between 96% and 100%. Those different strategies of combination give good results that can even reach 100% of identification rate, like strategies 4.

4.9. Appraisal And Synthesis Of The Results

Figure 4.11. shows a set of curves representing the different possible combinations results of various hybrid identification systems. We used 12 MFCC for all combination systems.

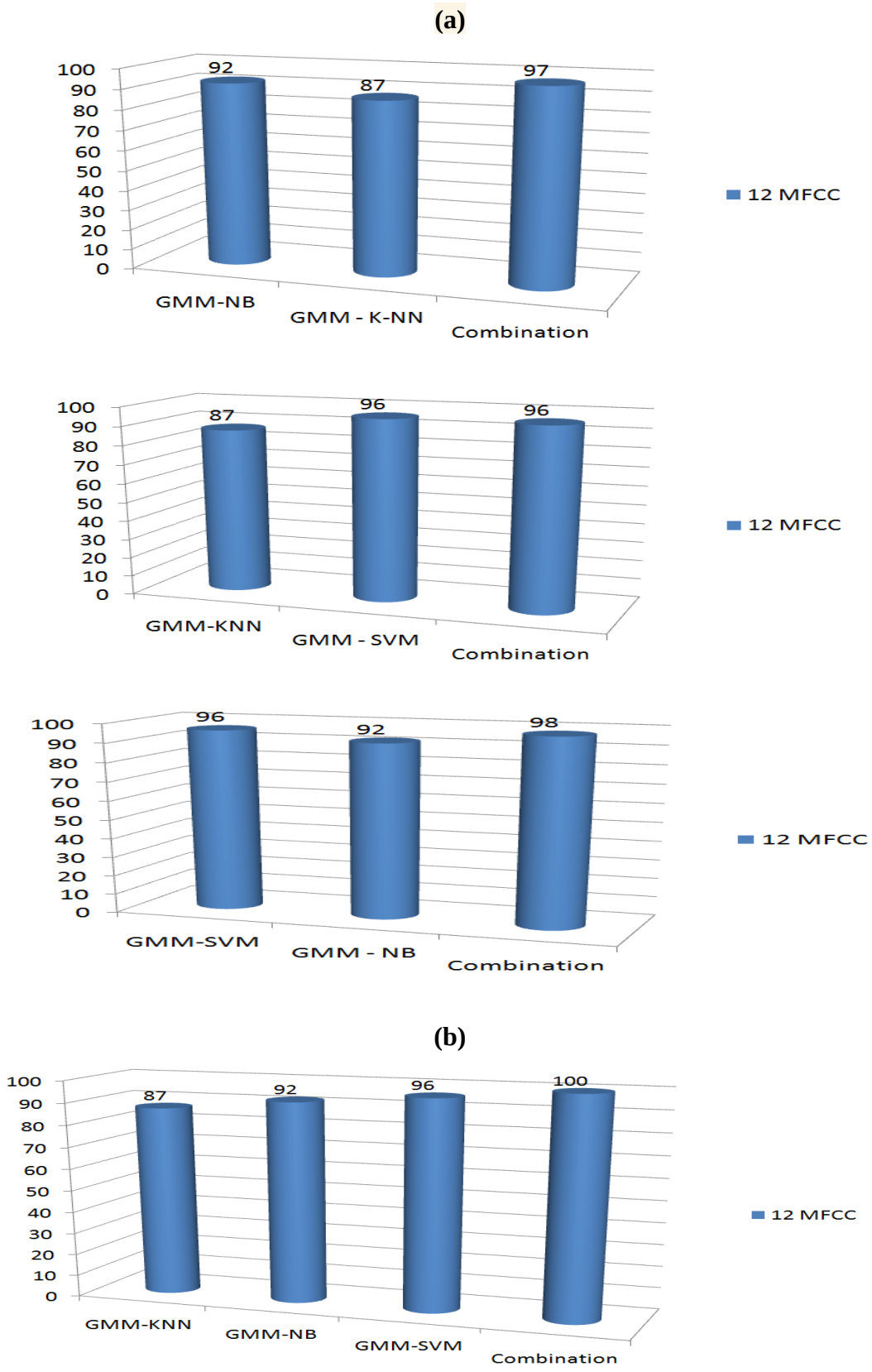

Figure 4.11. Results of hybrid systems combination by two systems(a) and by three systems (b).

By comparing the different hybrid systems studied with different possibilities of their combinations, we notice an improvement of identification rate level in all cases of combinations, which has been able to reach 100%. The minimum combination rate is 96% given by system 2: GMM-K-NN + GMM-SVM. The best identification rates of hybrid systems are those given by GMM-SVM and GMM-NB; it's for that reason all the combination including SVM and NB could reach an identification rate of 98%.

4.10. Robustness Of The Combination Of Hybrid Systems

After testing the hybrid systems with noisy data, we evaluate the combination of these hybrid systems with original data as is shown by the table 4.15 and then we tested the robustness of this combination with noisy data. The following figure 4.12. represents the resistance to noise of the combination of hybrid systems.

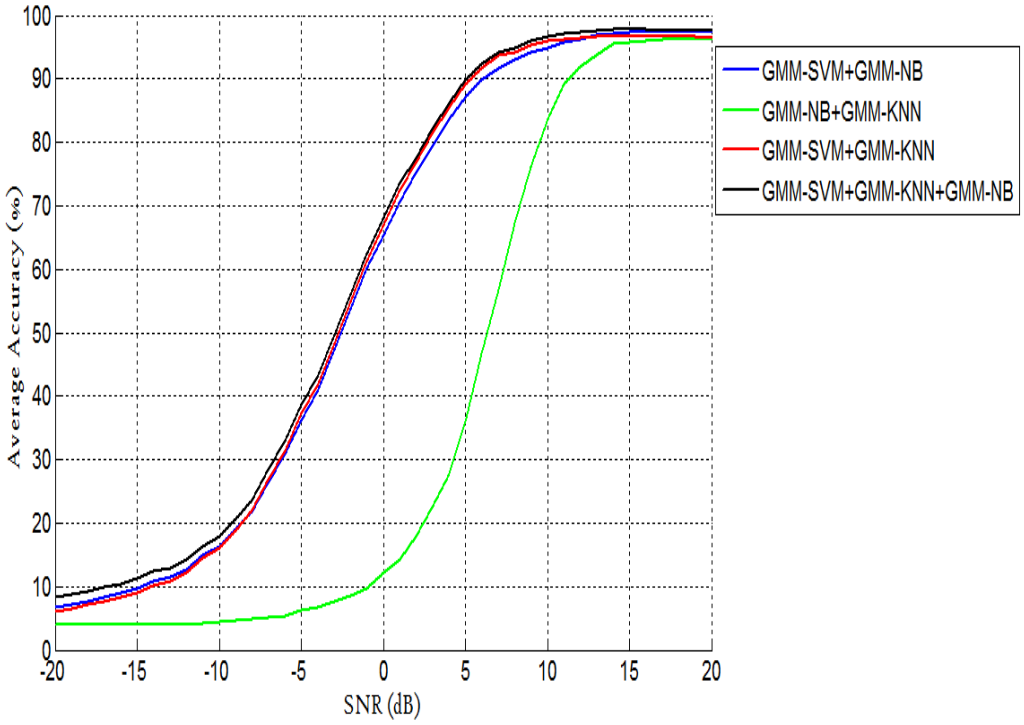

Figure 4.12. Performance of the combination of hybrid systems with noisy data.

By looking the figure 4.12, we can conclude that the combination of hybrid systems resist to the noise by comparing them to the single hybrid system.

As the GMM-SVM hybrid system proved its robustness to noisy date a combination also including this hybrid system gives good results. For example, with 5dB of SNR the combination of GMM-SVM+GMM-NB, GMM-SVM+GMM-KNN, GMM-SVM+GMM-KNN+GMM-NB hybrid systems give respectively the identification rates of 87%, 89%, and 90% instead of GMM-NB+GMM-KNN which gives 37% of identification rate.

4.11. Conclusion

In this chapter, we introduced the general tests which we adopted in order to evaluate the performances of our different systems. We started by evaluating SVM , K-NN and NB systems. Subsequently we evaluated these hybrid systems GMM-SVM, GMM-K-NN, GMM-NB. We studied the combination of our hybrids systems and we ended the chapter by testing the robustness of our hybrid systems and their combination with noisy data.

CHAPTER 5. GENERAL CONCLUSION

In this work of masters, we were interested by the design and implementation of most used discriminative classification methods. We also performed a combination of those methods based on generative modeling approach GMM, and this in order to enhance the robustness of the implemented systems.

In first part, the identification of speakers is implemented by three discriminative applications which are: SVM, K-NN, NB and by also studying the impact of some different technical parameters on speaker identification rate like the impact of FMCC coefficients number, impact of adding dynamics parameters and other impacts according to the selected classifier.

In second part, to enhance the robustness of these different systems, various strategies of hybridization have been implemented: GMM-SVM, GMM-K-NN and GMM-NB which are characterized by their capacities of multi-Gaussian modeling (GMM) and the effectiveness of decision. The implementation of such systems requires an initialization phase of acoustic parameters, an optimization phase of parameters generated by EM algorithm and an adaptation phase. A dictionary representing all GMM speakers supervectors is generated. The classifier selected is used to evaluate the similarity between a supervector characterizing the unknown speaker with all dictionary supervectors.

In third part, we studied the different possible combinations of those hybrid systems based on parallel combination of their decisions.

In fourth part, we introduced random noise in our database in order to test the robustness of our implemented hybrids and combination systems.

The study showed that the identification systems without modeling gave weak results with the identification rates of 3% for SVM, 27% for K-NN and 11% for NB. Also, we find that hybridization of discriminative and generative approaches is an effective method for identification systems and gave the following identification rates: 96% for SVM, 87% for K-NN and 92% for NB. In addition, the combination strategies implemented were also interesting and promising with an identification rate which can achieve 100% for some combination cases. With noise data our hybrids and combination systems proved their resistance to noise by giving an interested average identification rate.

The perspectives of this work are:

1. Evaluation of our systems on other corpus such that NTIMIT to detect the level of degradation.
2. Using other types of acoustical parameters.
3. Study of Gaussians number in the implemented hybrid systems.
4. Integration of one or several modalities (such as lips movement, face picture, etc) to the speech and merge them to characteristic parameters.

REFERENCES

[1] J.KHAROUBI., Etude de Techniques de Classement Machines à Vecteurs Supports pour la Véri cation Automatique du Locuteur. Thèse de l'Ecole Nationale Supérieure des Télécommunications, 2002.

[2] M.SCHMIDT, Identifying With Support Vector Networks. Computing Science and Statistics: 305-316, 1996.

[3] J. E.FIX, Discrimantory Analysis: NonParametric Discrimination: Consistency Properties. International Statistical Review(Revue Internationale de Statistique): 238-247, 1995.

[4] D.Meuwly, Reconnaissance de Locuteur par K Plus Proche Voisin. Université de Lausanne, Faculté de droit, Institut de police scientifique et de criminologie 2001.

[5] A. J. L.TOTH., On Naïve bayes in Speech Recognition. Int. J. Appl. Math. Comput. Sci., 15(2): 287–294, 2005.

[6] P. B. J. Zeljkovic, GMM/SVM N-Best Speaker Identification Under Mismatch Channel Conditions. In Acoustics, Speech and Signal Processing, IEEE International Conference on pp. 4129-4132, 2008.

[7] I.TRABELSI., Vers un Système Robuste en Identification Automatique du Locuteur par Supports de Vecteurs Machines. Mémoire de Mastere de l'Institut Supérieure d'Informatique, 2011.

[8] M. L. L. LAZLI., Nouvelle Méthode de Fusion de Données pour l'Apprentissage des Systèmes Hybrides MMC/RNA. ARIMA, CARI'04: 125-170, 2005.

[9] T. H. R.BOITE., Traitement de la Parole. PPUR presses polytechniques, 2000.

[10] D. A. REYNOLDS., Speaker identification and verification using gaussian mixture speaker models. Speech communication, 17(1):91-108, 1995.

[11] B. S. ATAL., Automatic Recognition of Speakers from Their Voices. Proceedings of the IEEE, 64(4):460-475, 1976.

[12] D.CHARLET., Authentification Vocale par Téléphone en Mode Dépendant du Texte. Thèse de Doctorat de l'Ecole Nationale Supérieure des Télécommunications, 1997.

[13] J.B.PIERROT., Elaboration et Validation d'Approches en Vérification du Locuteur. Thèse de Doctorat de l'Ecole Nationale Supérieure des Télécommunications, 1998.

[14] L.BESACIER, Un Modèle Parallèle pour la Reconnaissance Automatique du Locuteur. Thèse de Doctorat de l'Université d'Avignon et des Pays Vaucluse, 1998.

[15] G. M.M.HOMAYOUNPOUR., A Comparison of Some Relevant Parametric Representations for Speaker Verification. In ESCA Workshop on Automatic Speaker Recognition and Verification: 185-188, 1994.

[16] D.A.REYNOLDS., A Gaussian Mixture Modelling Approach to Text-Independent Speaker Identification. PhD thesis from Georgia Institute Technology, 1992.

[17] D.A.REYNOLDS., Experimental Evaluation of Features for Robust Speaker Identification. Speech and Audio Processing, IEEE Transactions on, 2(4): 639-643, 1994.

[18] S. VUUREN., Comparison of Text Independent Speaker Recognition Methods on Telephone Speech With Acoustic Mismatch. Proceedings., Fourth International Conference on. IEEE: 1788-1791, 1996.

[19] B. A. R.VERGIN., Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker Independent Continuous-Speech Recognition. Speech and Audio Processing, IEEE Transactions on, 7(5):525-532, 1999.

[20] P. S. B. DAVIS., Comparison of Parametric Representations for Mono Syllabic Word Recognition in Continuously Spoken Sentences. Acoustics, Speech and Signal Processing, IEEE Transactions on, 28(4): 357-366, 1980.

[21] A.PRETI., Surveillance de Réseaux Professionnels de Communication par la Reconnaissance de Locuteurs. Thèse de Doctorat de l'Université d'Avignon et des Pays, 2008.

[22] R.G.LEONARD., A Data base for Speaker-Independent Digit Recognition. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84. IEEE :328-331, 1984.

[23] L. W. J. D. E. N. . J.S.GAROLFO, The DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CDROM. US Dept. of Commerce, NIST, Gaithersburg, MD, Feburary, 1993.

[24] A. S. C.JANKOWSKI., NTIMIT: A Phonetically Balanced, Continous Speech Telephone Bandwidth Speech. In Acoustics, Speech, and Signal Processing. International Conference on IEEE,:109-112, 1990.

[25] W.CAMPBELL., Testing With the Yoho CD-ROM Voice Verification Corpus. In Acoustics. Speech, and Signal Processing. International Conference on IEEE,:341-344, 1995.

[26] F. M. A. A. E. M. M.ALGHAMDI, Saudi Accented Arabic Voice Bank. Final report, Computer and Electronics Research Institute, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia. 2003.

[27] D. P.JOSEPH., Corpora for The Evaluation of Speaker Recognition Systems. In Acoustics, Speech, and Signal Processing. International Conference on IEEE,: 829-832, 1999.

[28] V.VAPNIK., The Nature of Statistical Learning Theory. NewYork:Wiley, 1998.

[29] I. V. B.BOSER., A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM,: 144-152, 1992.

[30] H. M.SCHMIDT., Speaker Identification Via Support Vector Classifiers. In Acoustics, Speech, and Signal Processing. International Conference on IEEE,: 105-108, 1996.

[31] J. R. S.FINE., Enhancing GMM Score Using SVM Hints. In INTERSPEECH,: 1757-1760, 2001.

[32] J.RACHEDI., Reconnaissance et Classification de Phonèmes. Rapport de Mastère. Université de IRCAM, Paris, 2005.

[33] N. P. E. P. R.DEHAK., Linear and Non Linear kernel GMM SuperVector Machine for Speaker Verification. In INTERSPEECH,: 302-305, 2007.

[34] M.J.CARATY., La Reconnaissance Vocale et son Mentor : l'Evaluation. Habilitation à Diriger des Recherches. Université Pierre et Marie, 1999.

[35] D.SPIEGELHALTER., Speaker Probabilistic Prediction in Patient Management and Clinical Trials. Statistics in Medecine, 10(6): 925-937, 1991.

[36] N. M. I. SEBE., Emotion Recognition Using a Cauchy Naive Bayes Classifier. In Pattern Recognition. 16th International Conference on IEEE,: 17-20, 2002.

[37] J. A. L.ZHOU., Applying the Naive Bayes Classifier to Assist User in Detectind Speech Recognition Errors. Proceedings of the 38th Annual Hawaii International Conference on IEEE,: 183b-183b, 2005.

[38] M. H. R.DJEMILI, A Hybrid GMM/SVM System for Text Independent Speaker Identification. International Journal of Computer and Information Science and Engineering IJCSE, 2007.

[39] J. D. E., P. C., W.CAMPBELL., Support Vector Machines for Speaker and Language Recognition. Computer Speech & Language, 20(2): 210-229, 2006.

[40] T., R. B., D. A., REYNOLDS., Speaker Verification Using Adapted Gaussian Mixture Models. Digital signal processing, 10(1): 19-41, 2000.

[41] M. J., E.S., Parris., Speaker Verification Using Connected Words. PROCEEDINGS-INSTITUTE OF ACOUSTICS, 14:p95-p95, 1992.

[42] J.C.BEZDEK., Pattern Recognition With Fuzzy Objective Function Algorithms. Springer Science & Business Media, 2013.

[43] N. D. A.DEMPSTER., Maximum likelihood from Incomplete Data via the EM Algorithm. Journal of the royal statistical society. Series B (methodological),: 1-38, 1977.

[44] J.ZHANG, The Mean Field Theory in EM Algorithm Procedure for Markov Random Fields. Signal Processing, IEEE Transactions on, 40(10): 2570-2583, 1992.

[45] J. P.C., REYNOLDS, D. A., E., W.CAMPBELL, Support Vector Machines for Speaker and Language Recognition. Computer Speech & Language, 20(2): 210-229, 2006.

[46] I.BLOCH, Fusion d'Informations en Traitement du Signal et des Images. Hermes Science Publications, 2003.

[47] M., F.KIMURA, Handwritten Numeral Recognition Based on Multiple Algorithms. Pattern recognition, 24(10): 969-983, 1991.

[48] F.LEFEVRE, Estimation de Probabilité non Paramétrique pour la Reconnaissance Markovienne de la Parole. Thèse de l'Université Pierre et Marie Curie, 2000.

[49] T. DIDÉ, Réalisation d'un Framework pour la Reconnaissance de la Parole. Rapport de Stage de l'Ecole Polytechnique de Toulouse, 2007.

[50] D. B., R. A., N., AMAMI, An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel recognition. IJIIP: International Journal of Intelligent Information Processing, 3(1): 63-70, 2012.

[51] H. G., M., SIU, Evaluation of word confidence for speech recognition systems. Computer Speech & Language, 13(4): 299-319, 2000.

[52] K. E., K. A., O. M. N., K. MASOUD MALEKI, A new method for selection optimum k value in k-NN classification algorithm. In Signal Processing and Communications Applications Conference (SIU). IEEE,: 1-4, 2013.

[53] Y. S., A. Y. N., A. M., R. Raina, Classification with hybrid generative/discriminative models. In Advances in Neural Information Processing Systems, 2003.

[54] T. J., D. Haussler, Exploiting generative models in discriminative classifiers. Advances in neural information processing systems,: 487-493, 1998.

[55] L. L., C. Y. Suen., Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 27(5): 553-568, 1997.

[56] J.C.Bezdek., Models for Pattern Recognition. In Pattern Recognition With Fuzzy Objective Function Algorithm. Springer US,: 1-13, 1981.

RESUME

Yussouf NAHAYO was born in Burundi 1984. He finished primary school in 1998 at Ngozi II Primary School in Burundi and he continued the high school in the same city at Technical High School Alexandro Rossi where he finished in 2006 with A_0 level in Electro-electric section. in 2007 after passing national exam in Burundi he got a scholarship in Rwanda at National University of Rwanda, in Faculty of Applied Science, Computer Science and Systems Department. He has been interested with database management and programming, he graduated with distinction in 2012. February 2012 He started working in Burundi as assistant at Hope University in Faculty of Computer and Communication and after 4 months he got the Turkish scholarship at Sakarya University where he started learning Turkish language in Sakarya Tömer and finished it with C1 level in 2013.