PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار – عنابة
Year 2013/2014

FACULTY OF ENGINEERING SCIENCE
COMPUTER SCIENCE DEPARTEMENT

THESIS

Submitted In Fulfillment of the Requirement of the Degree of **Doctor** of Computer ***Science***

THEME

Contribution To The Automatic Speech Recognition of Arabic Language And its Applications

---

Option

Pattern Recognition and artificial intelligence

By

HAMMAMI Nacereddine

Thesis Director:	**Pr. FARAH Nadir**	Badji Mokhtar-Annaba University
Co-director:	**Pr. BEDDA Mouldi**	Aljouf University Arabie Saoudite

Jury

President:	**Pr. Khadir Med Tarek**	University of Annaba
Examiners:		
	Pr. Bahi Halima	University of Annaba
	Pr. Seridi Hamid	University of Guelma
	Pr. Moussaoui Abdelouaheb	University of Setif
	Pr. Boukabou Abdelkrim	University of Jijel

University year: 2013/2014

Acknowledgments

Praise be to Allah and thanks to Him, who facilitated the accomplishment of the thesis.

I express my gratitude to my supervisor, **Mr Nadir Farah**, professor at the University of Badji Mokhtar - Annaba and my co-supervisor Mr**. Mouldi Bedda**, professor at the University of Aljouf (Saudi Arabia) for their trust and for supervising my research group. I am grateful for their moral and scientific support, guidance and accurate feedback, which contributed to the accomplishment of this work. I express my gratitude for their personal and scientific ethics, which made me proud to be one of their students.

I thank Professor **Khadir Med Tarek**, from the University of Annaba for accepting the chairmanship of my thesis committee.

I offer many thanks and much gratitude to **Pr. Bahi Halima** from the university of Annaba, **Pr. Seridi Hamid** from the University of Guelma, **Pr. Moussaoui Abdelouaheb** from the university of Setif and **Pr. Boukabou Abdelkrim** from the university of Jijel for evaluating and discussing this work.

I thank my parents - may God bless them – my wife and all my family members for their caring, help and support during the achievement of this work.

I thank my beloved country Algeria and its universities, which provide me the opportunity to serve the nation, the world and humanity.

Abstract

Automatic speech recognition (ASR) techniques are evolving in manufacturing and public use, in which the techniques of voice recognition have been implemented in many electronic multimedia devices for daily use or in other fields including manufacturing, the military, and medical science. These techniques and their applications are rapidly advancing; automatic speech recognition is becoming one of the most important means of communication between humans and machines. In spite of tremendous progress in theoretical and applied technology for ASR), it is limited to English and some other languages, whereas ASR research and application in the Arabic language is limited. Many devices that support speech recognition techniques do not support Arabic, although it is an international language used by more than 400 million people in more than 22 countries and an official language of the United Nations; it is an important language for over one billion people because of its relationship to the Islamic religion. This thesis is an attempt to enrich the field of Arabic speech recognition in which two new methods in the field of speech recognition have been suggested and adopted. The first method is a statistical method depending on the tree probability model, and the second method depends on copula, an advanced statistical tool. These methods were suggested and used for the first time and are an addition to ASR in Arabic and other languages. The methods have produced excellent results that compete with other known models in the field of speech recognition. After this thesis, a case study was conducted and an application using automatic speech recognition has been used- with the aid of speech specialists - to develop a speech database to diagnose and treat speech disorders in children. The database is unique, and we hope that it could be a reference for research and in the use of Arabic automatic speech recognition. The results of automated diagnosis for accent disorders among children using this method have been encouraging, given the nature of the database, which contains noise; the database has encouraged investment in this research for the manufacture of devices that could serve children, and it is dependent on the techniques developed in this thesis.

ملخص

إن تقنيات التعرف الآلي على الأصوات تشهد في العصر الحالي ثورة حقيقية في الجانب التصنيعي والاستعمال العام حيث ضُمنت تقنيات التعرف على الكلام في كثير من الأجهزة الإلكترونية ذات الاستخدام اليومي سواء في الأجهزة المتعددة الوسائط أو غيرها من المجالات الصناعية والعسكرية والطبية . . .إلخ، بل إن هذه التقنيات وتوظيفها تسير بالبشرية و بخطى حثيثة لتصبح وسيلة التواصل الأولى بين الإنسان والآلة هي التخاطب الصوتي. لكن رغم هذا التقدم الهائل في الجانبين النظري والتطبيقي لتقنية التعرف الآلي على الكلام فإنه مقتصر على اللغة الإنجليزية وبعض اللغات الأخرى في حين تبقى اللغة العربية تعاني من نقص وشح واضح في الأبحاث و التطبيقات العملية لتقنيات التعرف الآلي على الكلام العربي ، بل لم يعد مستغرب أن تجد كثير من الأجهزة التي توظف تقنيات التعرف على الكلام لا تعتمد اللغة العربية ! وهذا رغم كونها لغة عالمية يتحدث بها أكثر من 400 مليون شخص عبر أكثر من 22 دولة كما أنها لغة رسمية بالأمم المتحدة وهي لغة مهمة عند ما يفوق المليار شخص بحكم علاقتها بالدين الإسلامي. هذه الأطروحة هي محاولة لإثراء مجال التعرف على الكلام العربي حيث تم اقتراح واستثمار طريقتين جديدتين في مجال التعرف على الكلام ؛ الأولى طريقة إحصائية تعتمد على النموذج الشجري الاحتمالي و الثانية طريقة تعتمد على أحد طرق الإحصائيات المتقدمة والمسماة الرابطة (Copula) وهذه الطرق فضلا على كونها تطرح وتستخدم لأول مرة فإنها تعتبر إضافة ليست خاصة فقط للغة العربية بل للتعرف الآلي على الكلام في بقية اللغات أيضا وقد أعطت نتائج جد مشجعة ومنافسة للنماذج والطرق المشهورة والمعروفة في مجال التعرف الآلي على الكلام . في آخر هذه الأطروحة تم دراسة حالة وإجراء عمل تطبيقي حقيقي يوظف تقنيات التعرف الآلي العربي في المجال الحياتي، حيث تم وبالتعاون مع أخصائيين في النطق إنشاء قاعدة بيانات صوتية لتشخيص ومعالجة الاضطرابات النطقية لدى الأطفال وكانت النتائج مشجعة وتفتح الباب لمجال استثمار هذه الأبحاث في تصنيع أجهزة تخدم هذه الفئة من الأطفال وتعتمد على التقنيات المطورة في هذه الأطروحة.

Résumé

Les techniques de reconnaissance automatique de la parole (RAP) sont en train de connaitre une grande évolution non seulement dans les domaines industriels et publiques où ces techniques intègrent les appareils électroniques utilisés au quotidien, mais aussi dans d'autres domaines tels que l'industriel, le médical, le militaire, etc. ces techniques et leur mise en pratique sont en train de conduire rapidement l'humanité vers une ère où le système de RAP pourrait devenir parmi les moyens de communication les plus importants entre l'homme et la machine. en dépit de l'énorme progrès qu'on observe dans les deux domaines théorique et pratique, les systèmes de RAP restent limités à l'anglais et à quelques autres langues. L'arabe reste cependant parmi les langues qui ne bénéficient pas beaucoup des travaux de recherche portant sur la pratique des techniques du RAP, il n'est pas surprenant alors de constater aujourd'hui que beaucoup d'appareils électroniques ne prennent pas en charge la reconnaissance la parole arabe ! ceci, malgré le fait que l'arabe soit l'une des langues officielles des nations unies car parlée par plus de 400 millions de personnes dans plus de 22 pays; sans oublier le fait qu'elle reste une langue d'une grande importance pour plus d'un milliard de personnes à travers le monde, notamment de par son lien étroit avec l'islam. cette thèse a pour objectif d'enrichir les systèmes de reconnaissance vocale pour la langue arabe, nous proposons deux nouvelles techniques de RAP. la première fait recours à une méthode statistique basée sur un modèle de probabilité arborisant, tandis que la deuxième se base sur des techniques de statistique avancée appelées (coupules). L'utilisation de ces techniques est une première dans le domaine de la RAP, et ceci non seulement pour la langue arabe, mais aussi pour les autres types de reconnaissance de la parole (d'autres langues) pour lesquelles notre méthode peut constituer une réelle alternative, d'autant plus qu'elle a donné d'excellents résultats qui rivalisent même avec ceux des modelés les plus connus dans le domaine de la reconnaissance automatique de la parole. A la fin de cette thèse, une étude de cas a été réalisée et on a procédé à l'utilisation en temps réel d'une application disposant d'un système de reconnaissance vocale – avec l'aide d'experts en linguistique -. durant le processus, une base de données de paroles a été mise en place pour analyser et traiter les troubles de la parole chez les enfants. Cette application est unique en son genre, et nous espérons qu'elle deviendra une référence pour les chercheurs dans le domaine de la reconnaissance automatique de la parole arabe et son utilisation. les résultats des analyses automatisées pour les troubles de la parole chez certaines catégories d'enfants en utilisant ce principe étaient très encourageants, en considérant la nature même de la base de données, qui contient quelques facteurs gênants, ils plaident largement à la faveur de l'exploitation de nos travaux de recherche basées sur les techniques développées dans cette thèse, pour la création d'appareils qui pourraient servir à ces catégories et l'ouverture de nouvelles perspectives de recherche dans ce domaine et pourra en plus être étendues à d'autres applications.

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
CDF	Cumulative Distribution Function
CMU	Carnegie Mellon University
CSR	Command Success Rate
CV	Cross Validation
DARPA	Defense Advanced Research Projects Agency
DHMM	Discrete Hidden Markov Model
DTW	Dynamic Time Warping
EARS	Effective Affordable Reusable Speech-to-text
EM	Expectation-Maximization
GCEMD	Gaussian Copula based on the Empirical Marginal Distributions
GCGMM	Gaussian Copula based on Gaussian Mixtures Marginal
GDP	Gross Domestic Product
GMM	Gaussian Mixture Model
GMMD	Gaussian Mixtures Marginal Distribution
HMM	Hidden Markov Model
ICA	Independent Component Analysis
JV	Japanese Vowels
KDE	Kernel Density Estimator
LDA	Linear Discriminant Analysis

LPC	Linear Predictive Coding
LPCM	Linear Pulse Code Modulation
MFCC	Mel-Frequency Spaced Cepstral Co-Efficient
MSA	Modern Standard Arabic
MWST	Maximum Weight-Spanning Tree
NLP	Natural Language Processing
PCA	Principal Component Analysis
PDF	Probability Density Function
PGM	Probabilistic Graphical Models
RDI	Research & Development International
/R/-DTDW	/R/-Disorder Types on the Beginning of the Words
/R/-DTEW	/R/-Disorder Types at the End of the Words
/R/-DTMW	/R/-Discord Types on the of the Word
/R/-LDD	/R/-Letter Disorder Diagnosis
SAD	Spoken Arabic Digits
SVM	Support Vector Machine
TDA	Tree Distributions Approximation
UN	United Nations
VIVO	Voice-In/Voice-Out
VODER	Voice Operating DEmonstratoR
WER	Word Error Rate

Contents

Part I : Automatic Speech Recognition : State of art

Part II: Proposed Contributions

Part III Application : Automatic Diagnosis and Rehabilitation Of Language Disorders For Children

Chapter I

Introduction

I. INTRODUCTION

This chapter provides the thesis introduction and emphasises the importance, usages, incentives and subject motivation as well as addressing the Automatic speech recognition (ASR) problems for the Arabic language. In this chapter, the outlines, general framework, key centres and desired targets of the thesis are clarified. The chapter briefly reviews the solutions that are suggested and developed in the thesis and provides a synopsis of the content of the remaining sections. The international publications within this thesis are found at the end.

I.1 MOTIVATION AND OVERVIEW

Speech is the primary means of communication between people. Automatic speech recognition (ASR) is the independent, computer-driven transcription of spoken language into readable text in real time [1]. ASR is a technique that automatically translates incoming speech signals into their contextual information via a sequence of words or other linguistic units by means of an algorithm implemented as a computer program. Having a machine understands fluently spoken speech has driven speech research for more than 50 years; however, the desire for automation of a simple task has existed for one hundred years. Based on major advances in statistical modelling of speech in the 1980s, automatic speech recognition systems today are frequently considered a key technology for human–machine communication and are incorporated in numerous applications. The features of these systems have many applications and domains that have useful applications. In the speech and telephone communication sector, such systems could be implemented in telephone network assistance, e.g., directory enquiries without operator assistance.

In the education sector, automatic speech recognition systems have a myriad of uses. They are utilised in teaching students of foreign languages to pronounce vocabulary correctly, and similarly in teaching international students to pronounce English correctly. They are utilised on behalf of students who are physically handicapped and unable to use a keyboard, enabling them to enter text verbally. In the higher education sector, the systems are used for the automatic generation of transcripts in narrative oriented research. This use would eliminate the time cost of generating the transcript manually and the

human error involved in manual transcription.

Speech recognition systems have recreational applications in computer games, video games and gambling as well as functional applications in precision surgery. In the domestic sector, the benefits of these systems could be applied to appliances including ovens, refrigerators, dishwashers and washing machines.

Speech recognition systems have a myriad of potential uses in the military sector including utilisation in high performance fighter aircraft, helicopters, battle management, air traffic controller training, and telephony as well as assistance to people with disabilities.

In the artificial intelligence sector, ASR could be applied to robotics; in medical fields, in could benefit healthcare and the generation of medical transcriptions (digital speech to text).

General areas of application for automatic speech recognition include automated transcription, telematics, air traffic control, multimodal interacting, court reporting and the grocery industry. Such systems could be useful to physically disabled people with limited use of their arms and legs and for those with sight problems.

"It understands what you say, it knows what you mean". This slogan represents the new trend in the field of ASR, especially in the omnipresent multimedia technology that exists in human interaction with electronic devices. Much effort has been made in recent decades to develop different ASR models. Because of extensive research, many models in recent years have offered a formal mathematical yet highly flexible method for solving many speech and language processing problems, namely acoustic phonetic, statistical pattern recognition and artificial intelligence approaches. Inflected languages such as Arabic pose new challenges for automatic speech recognition and related topics because of its rich morphology.

Although there are many diverse ASR applications in the commercial domain, applications in the Arabic language are very limited. Arabic is non-existent in many software applications that depend on ASR and natural language processing (NPL) and interest a large group of users. Such applications include multimedia voice applications (Siri and S-voice programs on smart phones) [1] and automotive functions that integrate speech recognition into select voice control systems.

[1] Siri is a voice personal assistant and knowledge navigator which works as an application for Apple Inc.'s iOS. S-voice is the Samsung Inc.'s Android version.

Arabic countries comprise one of the fastest growing regions in the world in terms of population, gross domestic product (GDP) growth and purchasing power, and the region is key focus area of governmental and developmental agencies.

The central motivation of this thesis is to conduct research into practical aspects of automatic Arabic speech recognition, to support the Arabic language in the new visionary digital age and to provide the most modern Arabic language translation solutions that apply across the full range of operational tasks. The tasks include quickly deciphering manuscript volumes, facilitating real-time translations on mobile devices and providing agility, flexibility and innovation to support the demands associated with the communication needs of today.

I.2 AIMS AND SCOPE

Automatic speech recognition for the English language is an emerging and very broad topic of research whereas a minimal amount of work has been conducted in the Arabic language. The existing research on Arabic language ASR has several major limitations that cause a lack of Arabic software and in sufficient applications in addition to the limitations resulting from the industrial recession. This thesis aims to address these limitations and build on the existing body of research by extending the current approaches to enable practical deployment of Arabic ASR applications and to investigate new methods by which enhancement systems could be optimised to improve Arabic ASR performance.

We hope that the Automatic Rehabilitation of Language Disorders for Children application developed in this thesis will encourage more research to create applications that serve Arabic-speaking communities.

The general research objectives are as follows:

(i) To review the current state of research in ASR and Arabic ASR.

(ii) To require attention for the extraction and estimation of speech signal features and parameters for ASR models because recognition performance depends heavily on this phase.

(iii) To propose novel approaches in ASR.

(iv) To consider the implementation of the proposed models within real data sets.

(v) To quantify the word accuracy performance of ASR systems with and without noise.

(vi) To assess each of the proposed techniques and report their performance based on speech recognition accuracy using real data.

(vii) Design applications in the Arabic language.

The scope of this program is the machine-learning algorithm of pattern recognition that underlies speech recognition. Specifically, this research focused on isolated Arabic word speech recognition and the proposal of real applications (The Automatic Rehabilitation Of Language Disorders For Children application was developed).

For this thesis to have a tight focus in its aims and objectives, the scope of research is limited to the following field and applications.

- ***Arabic Word Recognition***: This objective is based on "isolated word recognition", which operates on single words at a time and requires a pause between saying each word.
- ***Small and Medium-Size Vocabulary:***
  The size of the vocabulary in a speech recognition system affects the complexity, processing requirements and accuracy of the system. The task provides a better indication in the practical application of the performing system while limiting the complexity of the ASR system.
- ***Real World Noisy Environment*:** Our goal is to develop programs and applications the benefit from Arab community and it is appropriate to develop systems that work in real conditions and to accept the challenge of achieving the best results from the database that was collected.
- ***A Speaker Independent System*:** This system is developed to operate for any speaker. Speaker independent recognition is more difficult because the internal representation of speech must be sufficiently global to cover all types of voices and all possible pronunciations of words.
- ***The Statistical Model For ASR****:* Statistical models are the classical approach for speech recognition, and they are the major contributors and fundamentally employed methods for many recent applications of pattern recognition.
- ***Working On Real Datasets For Life Application.***

I.3 ORIGINAL CONTRIBUTIONS

In this thesis, a number of original contributions were made to the field of Arabic speech recognition and to ASR in general. These contributions are summarised as follows:

- Two new efficient models for automatic speech recognition were developed. The first one is for discrete speech and is based on a graphic model (Trees) (Chapter III). The second is for continuous speech and is based on the concept of copulas (Chapter IV).
- Experiments on different languages and a real benchmark database.
- Comparison with several basic models for ASR.
- Joint fieldwork with specialists in the treatment of childhood speech disorders and the development of a database for use in the diagnosis of pronunciation in children and of automatic programs for the rehabilitation process, in which speech was recorded in the natural environment of the children (schools, home, street, yard games) (Chapter V & VI).

I.4 Thesis Structures

The remainder of the thesis is organised as follows:

Part I: Automatic Speech Recognition: The State of the Art

Chapter II provides a general literature review of the digital speech processing field and the fundamentals of ASR with a historical review of the efforts to develop automatic recognition systems. Historical information is provided, and the basic properties of human speech production and perception systems are explained. We discuss the main components required for building a generic state-of-the-art ASR system. We also discuss several of the current ASR models, including the feature extraction step. This chapter presents the Arabic language and the basic properties of Arabic phonology, including the problems in Arabic speech recognition. We provide an overview of current ASR with a significant focus on the research with the Arabic language.

Part II: Proposed Contributions

Chapter III details a new discrete speech recognition method that investigates the capability of the graphic models based on tree distributions, which are widely used in many optimisation areas. A novel spanning tree structure that utilises the temporal nature of the speech signal is presented. We evaluated this method experimentally using a Japanese and Arabic speech database and the proposed approaches were compared to the conventional discrete hidden Markov model (DHMM).

Chapter IV presents a new technique for automatic speech recognition by employing a full measure of statistical dependence among random variables, which is known as copulas. We present the definition of copulas and the details of a novel probabilistic classifier that combines finite Gaussian mixture modelling for the marginal distribution function and the Gaussian copula. We present the experimental results and discussions on the use of different features and classification methods.

Part III Applications: Automatic Diagnosis of Language Disorders in Children

Chapter V presents speech disorders in children and discusses an overview of the types of and most common disorders in speakers of Arabic this chapter details the collaboration with experts in the field to determine the most common problems and create a database that contributes to the creation of an automated system for diagnosis and rehabilitation therapy for children.

Chapter VI presents the registration process and details the database (including the recording conditions, preliminary processing and speech file databases). This chapter presents the experimental results and discussions on the use of different classifiers using the developed datasets for the Automatic Diagnosis of Language Disorders in Children.

Chapter VII concludes the dissertation with a summary of the contributions of this research and suggests further directions for continuing research in and applications of Arabic speech recognition.

I.5 PUBLICATIONS

I.5.1 PUBLICATION RESULTING FROM DISSERTATION RESEARCH

- **International Journal Publications**

[1] N. Hammami, M. Bedda, and N. Farah, " Tree distributions approximation model for robust discrete speech recognition" ,International Journal of Speech Technology, Springer, 2012.

[2] N. Hammami, M. Bedda, and N. Farah, "Copula Function: A New Approach for Speech Recognition with Application to Spoken Arabic Digits", (submitted to Elsevier journal).

- **International conference publications**

[3] N.Hammami, M.Bedda, N.Farah, R.Lakehal-Ayat3 "Spoken Arabic Digits recognition based on (GMM) for e-Quran voice browsing: Application for blind category", Proc.IEEE (ICAITHQ'13), Medina, KSA, Sep 2013.

[4] N.Hammami, M.Bedda, N.Farah, "Probabilistic Classification Based on Gaussian Copula for Speech Recognition: Application to Spoken Arabic Digits", Proc.IEEE (SPA'13), Signal Processing, Poland, Sep 2013.

[5] N.Hammami, M.Bedda, N.Farah, "Spoken Arabic Digits recognition using MFCC based on GMM," Proc. IEEE, (STUDENT'12) Sustainable Utilization and Development in Engineering and Technology, pp.160,163,Malysia, 6-9 Oct. 2012.

[6] N. Hammami, M. Bedda, N.Farah ," Spoken Arabic Digits recognition Using MFCC based on GMM ", Accepted in. IEEE STUDENT'12 Conference, Malysia, 2012.

[7] N. Hammami, M. Bedda, N.Farah "The second-order derivatives of MFCC for improving spoken Arabic digits recognition using Tree distributions approximation model and HMMs," Communications and Information Technology (ICCIT), 2012 International Conference on , vol., no., pp.1,5, Tunisia, 26-28 June 2012.

[8] N. Hammami, M. Bedda, N.Farah "HMM parameters estimation based on cross-validation for Spoken Arabic Digits recognition," Communications, Computing and Control Applications (CCCA), 2011 International Conference on , vol., no., pp.1,4,Tunisia, 3-5 March 2011.

[9] N.Hammami, M.Bedda, N.Farah, " إختيار عناصر الإدخال في نموذج ماروكفي ، الدورة السابعة "مخفي بإستعمال التحقق المتقاطع للتعرف على الأعداد العربية المنطوقة الرياض المملكة .ICCA, 2011 المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية، (نسخة معربة)العربية السعودية

[10] N. Hammami, M. Bedda "Improved tree model for arabic speech recognition," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.5, no., pp.521,526,China, 9-11 July 2010.

[11] N. Hammami, M. Bedda, Adaptation and preprocessing of the Repository dataset:"spoken Arabic digit", http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit.

I.5.2 OTHER PUBLICATIONS

[12] Hammami Nacereddine, Mohamed Goudjil and Alruily Meshrif, "*Probabilistic Classification Based on Copula for Text categorization: An Overview*", Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.

[13] Alruily Meshrif , Hammami Nacereddine, Goudjil Mohamed, "*Using Transitivity for developing Arabic Text Summarization System*", Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.

[14] Mohamed Goudjil , Mouloud Koudil , Nacereddine Hammami ,Mouldi Bedda and Meshrif Alruily; "*Arabic Text Categorization Using SVM Active Learning Technique : An Overview*"; Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.

[15] Khelifa, M.A.; Boukabou, A.; Hammami, N., "Data Transmission Based on Chaotic Synchronization System," Computer Applications Technology (ICCAT), 2013 International Conference on , vol., no., pp.1,2, Tunisia, 20-22 Jan. 2013

[16] N. Hammami, M. Sellami "Tree distribution classifier for automatic spoken Arabic digit recognition," Internet Technology and Secured Transactions International Conference for , vol., no., pp.1,4,UK, 9-12 Nov. 2009.

[17] A.Douzal, N. Hammami,C.gabry, "Local Analysis of multivariate time series". ISI 2007 (International Statistical Institute), Portugal 2007.

Part I

Automatic Speech Recognition:

State of art

Chapter I

An Overview of Automatic Speech Recognition

II.1 INTRODUCTION

The aim of this chapter is to provide background information on automatic speech recognition and its components as well as on human speech production and perception. An historic perspective on the key inventions that have enabled progress in speech recognition is presented. We describe in this chapter the structure of ASR and its performance evaluation and feature extraction methods, and we explore the ASR approaches. The Arabic language and ASR for Arabic are discussed.

II.2 THE HUMAN SPEECH AND BITS OF HISTORY

In this section, we present briefly human speech production and perception, detail can be found on [2][3].

II.2.1 SPEECH PRODUCTION

The production of speech sounds starts in the brain. After the creation of the message and the lexico-grammatical structure in our mind, we need a representation of the sound sequence and a number of commands which will be executed by our speech organs to produce the utterance.

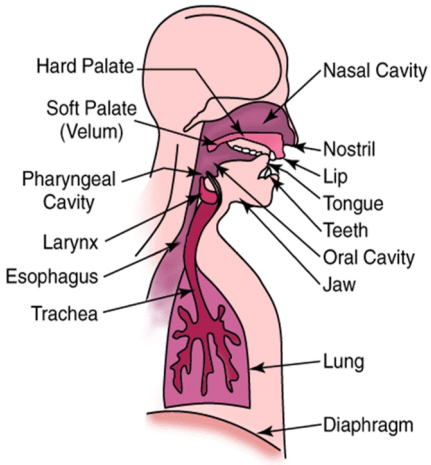


Figure II.1. PRODUCTION OF SPEECH SOUNDS [3]

So, we need a phonetic plan of and a motor plan (Belinchón, Igoa y Rivière, 1994: 590). Then the physical production of sounds starts. It starts when air is expelled from the lungs by muscular force providing the source of energy. Then the airflow is modulated in various ways to produce different speech sounds.

The modulation is performed in the vocal tract, through movements of many articulators, such as the velum, teeth, lips, and tongue. The movements of the articulators modify the shape of the vocal tract, which creates many resonant frequencies and, consequently, different speech sounds (Figure II.1). Finally, the articulation process which takes place in the mouth and it is the process through which we can differentiate most speech sounds. In the mouth we can distinguish between the oral cavity, which acts as a resonator, and the articulators, which can be active or passive: upper and lower lips, upper and lower teeth, tongue (tip, blade, front, back) and roof of the mouth (alveolar ridge, palate and velum). Speech sounds are distinguished from one another in terms of the place where and the manner how they are articulated.

II.2.2 SPEECH PERCEPTION

This section gives a brief overview of the speech perception by focusing on the physical aspects of the speech perception used for speech recognition.

II.2.2.1 THE AUDITORY SYSTEM

The performance of an automatic speech recognition system could be expected to improve by using knowledge about the human auditory system. The paragraph below presents briefly the human auditory system.

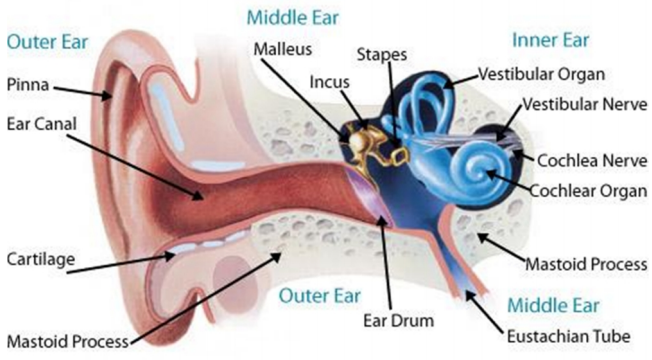


Figure II.2. STRUCTURE OF THE HUMAN EAR[1]

At the linguistic level of communication first the idea is formed in the mind of the speaker. The idea is then transformed to words, phrases and sentences according to the grammatical rules of the language.

[1] http://www.hearingclinic.net.au/mhc/content/the_ear.php

At the physiological level of communication the brain creates electric signals that move along the motor nerves.

The auditory system is anatomically and functionally divided into three areas: the outer ear, the middle ear, and the inner ear (fig II.2). The outer ear is composed of the pinna (external ear, the part we can see) and the external canal (or meatus). The pinna modifies the incoming sound (particularly high frequencies) and directs it to the external canal. The filtering effect of the human pinna preferentially selects sounds in the frequency range of human speech. It also adds directional information to the sound. The middle ear is an air-filled cavity (tympanic cavity) that couples sound from the air to the fluids via oval window in the cochlea (fig II.3).

The inner ear consists of a bony labyrinth filled with fluid that has two main functional parts: the vestibular system and the cochlea.

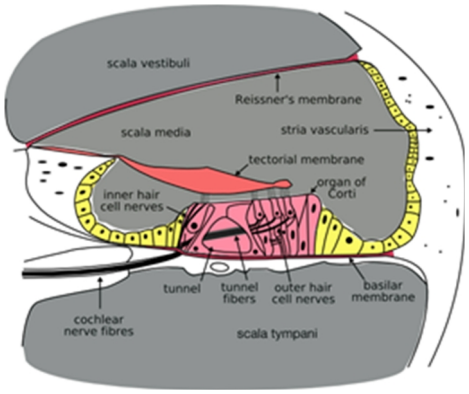


Figure II.3. CROSS SECTION OF THE COCHLEA[2]

II.3 BITS OF HISTORY

This section presents a brief chronological overview on Speech Recognition and discusses the important themes and advances made in the past two centuries of research, this section summarizes speech recognition system evolution presented on [4][5].

IN THE 18th C: The early interest was on creating a speaking machine, perhaps due to the readily available knowledge of acoustic resonance tubes which were used to approximate the human vocal tract.

[2] http://en.wikipedia.org/wiki/Cochlea

IN 1773: The Russian scientist Christian Kratzenstein, a professor of physiology in Copenhagen, succeeded in producing vowel sounds using resonance tubes connected to organ pipes. Later,

IN 1791: Wolfgang Von Kempelen, in Vienna, constructed an "Acoustic-Mechanical Speech Machine"

IN THE MID-1800'S: Charles Wheatstone built a version of von Kempelen's speaking machine using resonators made of leather, the configuration of which could be altered or controlled with a hand to produce different speech-like sounds, as shown in Figure II.4.

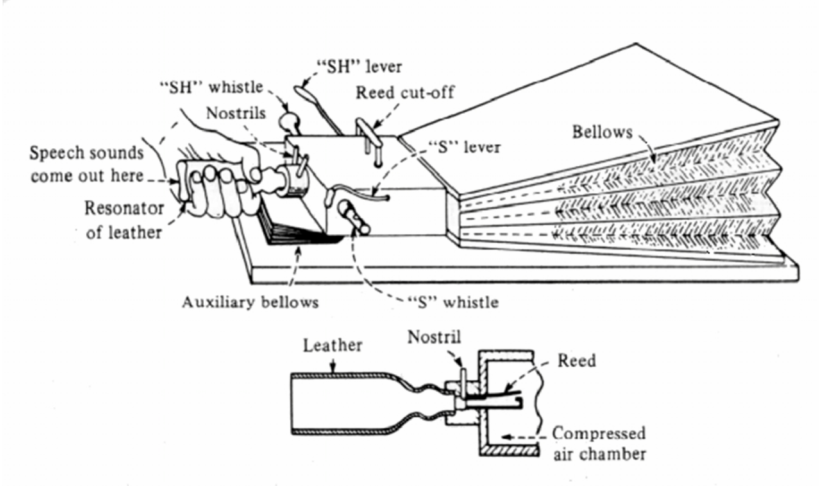


Figure II.4. WHEATSTONE'S VERSION OF VON KEMPELEN'S SPEAKING MACHINE[6]

EARLY 20th CENTURY: Fletcher and others at Bell Laboratories documented the relationship between a given speech spectrums band its sound characteristics as well as its intelligibility, as perceived by a human listener.

IN 1930: Homer Dudley developed a speech synthesizer called the VODER (Voice Operating Demonstrator). An electrical equivalent (with mechanical control) of Wheatstone's mechanical speaking machine (Fig II.5). The VODER was showcased at the World Fair in New York City in 1939. It was considered as an important milestone in the evolution of speaking machines.

IN THE 1950s and 1960s:

- **In 1952**, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker, using the formant frequencies measured/estimated during vowel regions of each digit.

- **In 1956**, at RCA Laboratories, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words
- **In 1959**, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants. This work represents the first use of statistical syntax in automatic speech recognition.
- **In 1959**, at MIT Lincoln Laboratories, Forgie and Forgie devised a system which was able to recognize 10 vowels embedded in a /b/ - vowel -/t/ format in a speaker independent manner.
- **In the 1960s**, at the Radio Research Lab in Japan, special-purpose hardware was designed. Suzuki and Nakata built a hardware vowel recognizer
- **In 1962:** Sakai and Doshita at Kyoto University built a hardware phoneme recognizer using a hardware speech segmenter and a zero-crossing analysis of different regions of the input utterance.
- **In 1963**: Nagata and his colleagues at NEC Laboratories built a hardware digit recognizer.

IN THE 1970S

- **IBM Labs**: A study on large vocabulary speech recognition for three distinct tasks, namely the New Raleigh language for simple database queries, the laser patent text language for transcribing laser patents, and the office correspondence task, called Tangora, for dictation of simple memos, was carried out.
- **AT&T Bell Labs**: Experiments was held aiming at making speaker-independent speech- recognition systems. A wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population.
- DARPA program: An ambitious speech understanding project was funded by the Defense Advanced Research Projects Agency (DARPA), which led to many seminal systems and technologies.

IN THE 1980s

The focus of the researches in the 1980s was mainly on problems of creating a robust system capable of recognizing a fluently spoken string of connected word.

Statistical modeling: Speech recognition research was characterized by a shift in methodology from the more intuitive template-based approach towards a more rigorous statistical modeling framework.

- **HMM:** The hidden Markov model (HMM). It is a doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations.
- **Δcepstrum**: Furui proposed to use the combination of instantaneous cepstral coefficients and their first and second- order polynomial coefficients, now called Δ and ΔΔcepstral coefficients, as fundamental spectral features for speech recognition.
- **N-gram:** it defined the probability of occurrence of an ordered sequence of n words. It has become indispensable in large- vocabulary speech recognition systems.
- **Neural net:** Neural networks were first introduced in the 1950s, but they did not prove useful because of practical problems. Later in the 80s, a deeper understanding of the strengths and limitations of the technology was achieved.

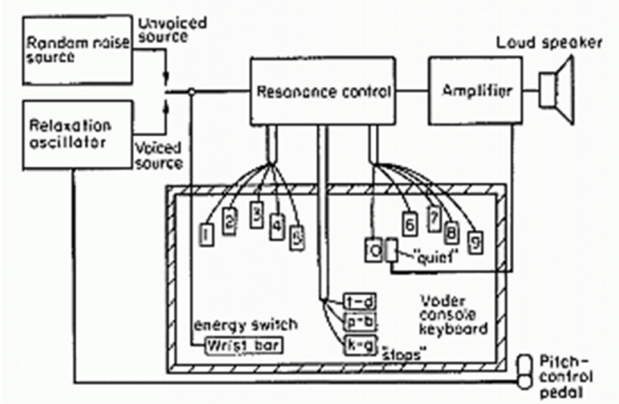


Figure II.5. A BLOCK SCHEMATIC OF HOMER DUDLEY'S VODER [7]

IN THE 1990s

Many innovations took place in the 1990s in the field of pattern recognition. This fundamental change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory becomes inapplicable under these circumstances.

IN THE 2000s

DARPA program: The Effective Affordable Reusable Speech-to-Text (EARS) program was conducted to develop speech-to-text (automatic transcription) technology in order to achieve substantially richer and much more accurate output. The program was focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages (Figure II.6).

Spontaneous speech recognition: Aiming at increasing recognition performance for spontaneous speech, several projects have been conducted such as in Japan, a 5-year national project "Spontaneous Speech: Corpus and Processing Technology" was conducted.

Robust speech recognition: In order to further increase the robustness of speech recognition systems, especially for spontaneous speech, utterance verification and confidence measures are being intensively investigated. The confidence measure serves as a reference guide for a dialogue system to provide an appropriate response to its users.

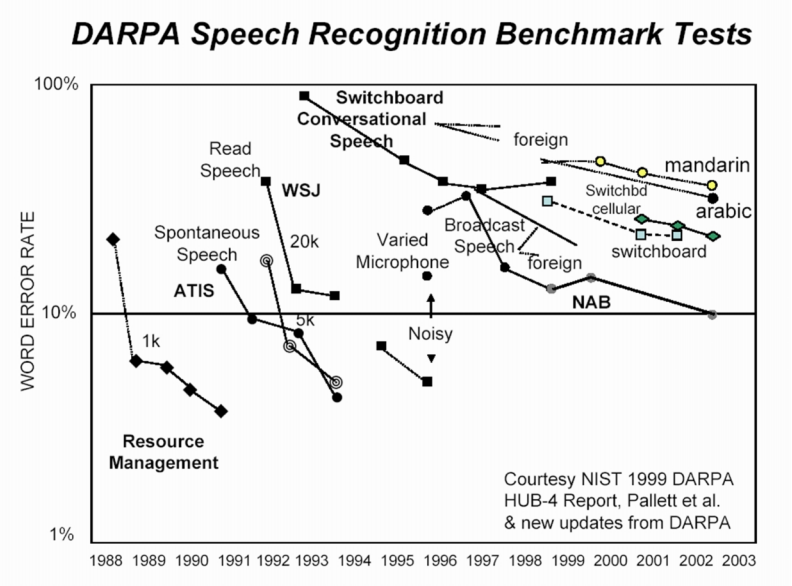


Figure II.6. DARPA BENCHMARK EVALUATION OF SPEECH RECOGNITION FOR A NUMBER OF TASKS[4]

Multimodal speech recognition: The use of the visual face information in speech recognition has been investigated, and results show that using both types of information gives better recognition performances than using only the audio or only the visual information.

In Figure II.7 represents a timeline of progress in speech recognition and understanding technology over the past four decades.

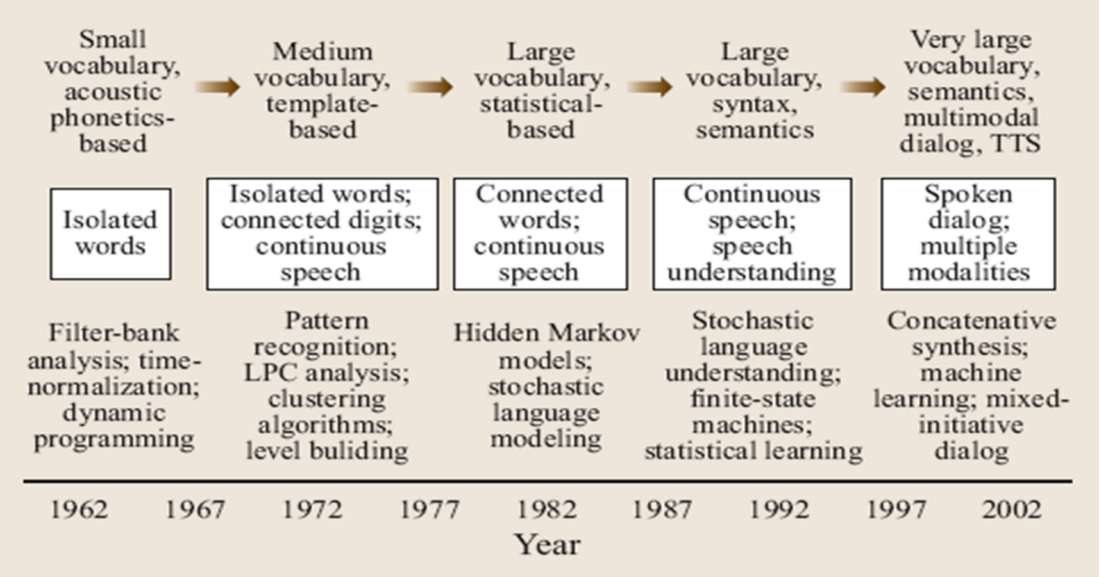


Figure II.7. IMPORTANT STAGES IN SPEECH RECOGNITION AND UNDERSTANDING TECHNOLOGY OVER THE PAST 4 DECADES [4].

II.4 AUTOMATIC SPEECH RECOGNITION

II.4.1 SOUND GENERATION AND SPEECH SIGNAL

Acoustic speech output is commonly considered to result from a combination of a source of sound energy (e.g. the larynx) modulated by a transfer (filter) function determined by the shape of the supra-laryngeal vocal tract. This combination results in a shaped spectrum with broadband energy peaks.

Other sound sources are created by turbulence at obstacles to the air-flow. Noise sources caused by the turbulence have broad continuous spectra, varying from about 2 to 6 kHz depending on the exact place and shape of constriction.

Normally, noise sources have a single broad frequency peak, rolling off at lower and high frequencies, as shown at the bottom in the left column in figure II.8. When sound goes out of the lips and nostrils, its frequency shaping is modified again which helps differentiate the signals. In a long period, speech signals are non-stationary but in a short interval between 5 and 100ms.

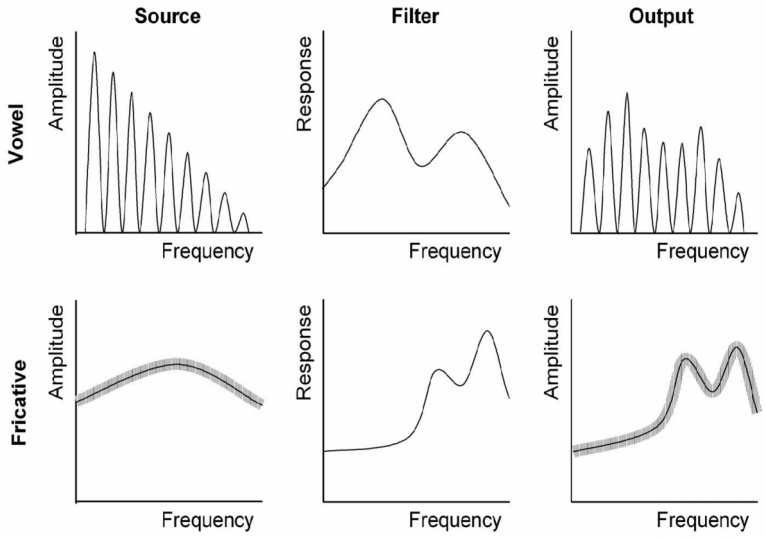


Figure II.8. FREQUENCY DOMAIN DIAGRAM OF THE SOURCE-fiLTER EXPLANATION OF THE ACOUSTICS OF A VOWEL (VOICED) AND A FRICATIVE (VOICELESS). THE SOURCE SPECTRUM (LEFT),THE VOCAL TRACT TRANSFER FUNCTION (MIDDLE), AND THE OUTPUT SPECTRUM (RIGHT), AFTER DELLWO [8]

The basic mechanism involved in transforming a speech waveform into a sequence of parameter vectors is illustrated in figure II.9. The frames overlap by setting the frame period smaller than the window size. Each frame is then investigated to extract parameters. This process results in a sequence of parameter blocks. SOURCERATE and TARGETRATE in the following figure are the number of samples of the wave source and the number of extracted feature vectors, respectively.

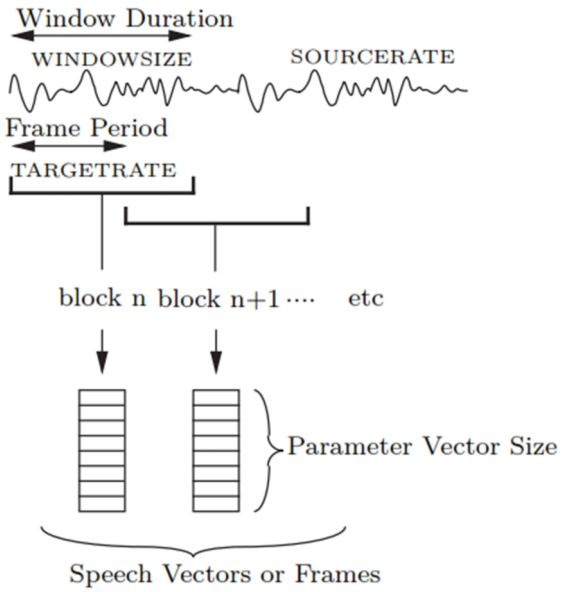


Figure II.9. SPEECH ENCODING PROCESS, AFTER YOUNG [9]

The frame of speech to be the product of a shifted window with the speech sequence[19]:

$$f_s(n;m) = s(n)w(m-n) \quad \text{(II.1)}$$

Where $s(n)$ is the speech signal and $w(m-n)$ is a window of length N ending at sample m. There are some simple pre-processing operations that can be applied before the actual signal analysis. At first, the DC mean (the mean amplitude of the waveform) can be removed from the source waveform. This is useful when the original analogue-digital conversion has added a DC offset to the signal. Second, the signal is usually pre-emphasized by applying the first order difference equation[9]:

$$s'(n) = s(n) - ks(n-1) \quad \text{(II.2)}$$

to the samples $s(n), n = 1,\ldots,N$ in each window. Where k in the range $0 \leq k < 1$ is the pre-emphasis coeffcient. Finally, the samples in each window usually apply a window with smooth truncations so that discontinuities at the window edges are attenuated. Some of the commonly used windows with smooth truncations are Kaiser, Hamming, Hanning and Blackman. These windows have the benefit of less abrupt truncations at the boundaries. For Hamming window, the $samples\, s(n), n = 0,\ldots,N$ in each window apply the following transformation"[10]:

$$w_n = \begin{Bmatrix} 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) & 0 \leq n \leq N \\ 0 & otherwise \end{Bmatrix} \quad \text{(II.3)}$$

II.4.2 AUTOMATIC SPEECH RECOGNITION STRUCTURE

The recognition problem on (ASR) consists of two fundamental modules:

A) Front-end speech parameterization stage: This stage is for converting the speech into format machine can both use and understand.

B) Back-end stage: Consist of training and deployment. Training is undertaken so the system will learn to associate model with known transcriptions. Once the back-end recognizer can be deployed to transcribe unknown speech samples. Figure II.10 shows an overview of a typical ASR structure.

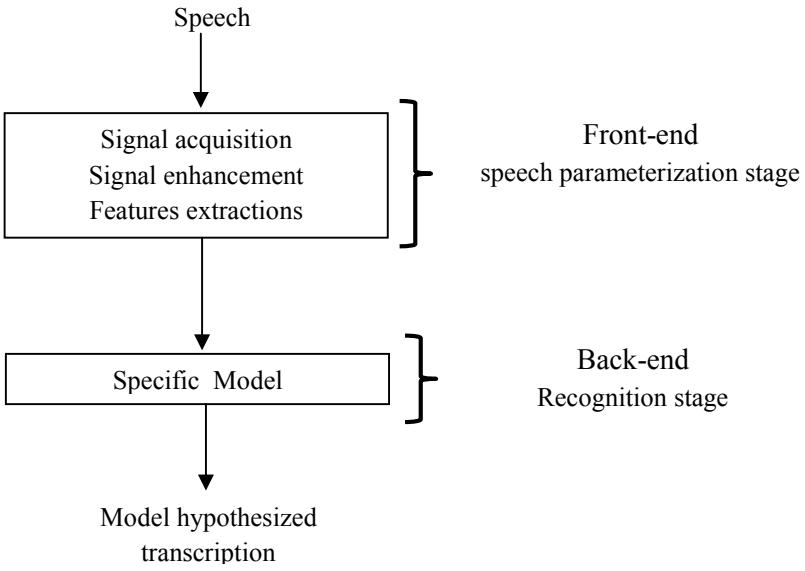


Figure II.10. GENERIC ASR FRAMEWORK WITH FRONT-END AND BACK-END

II.4.3 MATHEMATICAL REPRESENTATION OF ASR

The classic approach to speech recognition using statistical methods [11] formulates the problem as follows. The goal of the speech recognizer is to find the word sequence $\widehat{W}$ that maximizes P(W|O), which is the probability of the word sequence given the acoustic evidence that is extracted for the observed acoustic waveform.

Using Baye's rule it can be written as:

$$\widehat{W} = argmax_w \ P(W|O)$$
$$= \ argmax_w \ P(O|W) \ P(W)$$

since P(O) does not depend on W. This decomposition defines the basic structure and components of all of today's speech recognition systems [12].The first component is a feature extractor that produces O, the second component is the acoustic model P(W|O).

II.4.4 PERFORMANCE EVALUATION OF SPEECH RECOGNITION SYSTEMS

The recognition performance evaluation of an ASR system must be measured on a corpus of data different from the training corpus. The performance of speech recognition system is usually specified in terms of accuracy and speed.

Accuracy is computed by word error rate, whereas speed is measured with the real time factor. Other measures of accuracy include single word error rate and command success rate (CSR). Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system.

WER can then be computed as:

$$WER = (S + D + I)/N \quad \text{(II.4)}$$

Where;

S is the number of substitutions,

D is the number of deletions,

I is the number of insertions,

N is the number of words in the reference.

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used:

$$WRR = 1 - WER = 1 - (S + D + I)/N = (H - I)/N \quad \text{(II.5)}$$

where H is N-(S+D) the number of correctly recognized words.

II.5 FEATURE EXTRACTION

Because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc. Many such features have been investigated in the literature [13]-[16]. Table II.1 shows broadly various methods for feature extraction in speech recognition.

Method	Property	Comments
Principal Component Analysis(PCA)	Non linear feature extraction method, Linear map; fast; eigenvector-based	Traditional, eigenvector based method, also known as karhuneu-Loeve expansion; good for Gaussian data.
Linear Discriminant Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast; eigenvector-based	Better than PCA for classification;
Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non- Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
Linear Predictive coding	Static feature extraction method,10 to 16 lower order co- efficient,	
Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope
Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel-frequency scale.
Filter bank analysis	Filters tuned required frequencies	
Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	
Kernel based feature extraction method	Non linear transformations,	Dimensionality reduction leads to better classification and it is used to remove noisy and redundant features, and improvement in classification error
Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	
Spectral subtraction	Robust Feature extraction method	
Cepstral mean subtraction	Robust Feature extraction	
RASTA filtering	For Noisy speech	
Integrated Phoneme subspace method	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existing methods

Table II.1. FEATURE EXTRACTION METHODS [30]

III.5.1 SPECTRAL FEATURES [17]

Over the past two decades spectral based features, most typically derived by direct application of the Fourier Transform, have become popular. These features are Mel-Frequency spaced Cepstral Co-efficient (MFCCs) and their success arises From the use of perceptually based Mel spaced filter Bank processing of the Fourier Transform And the particular robustness (to the environment) and flexibility that can be achieved using cepstral analysis .

A. FRAMES

The most fundamental process that is common to all forms of speaker and speech recognition systems is that of ex- tracing vectors of features uniformly spaced across time from the time-domain sampled acoustic waveform. with reference to Fig II.11, proceeds as follows (the numerical parameter values mentioned are those typi- cally adopted in practice)"[17]:

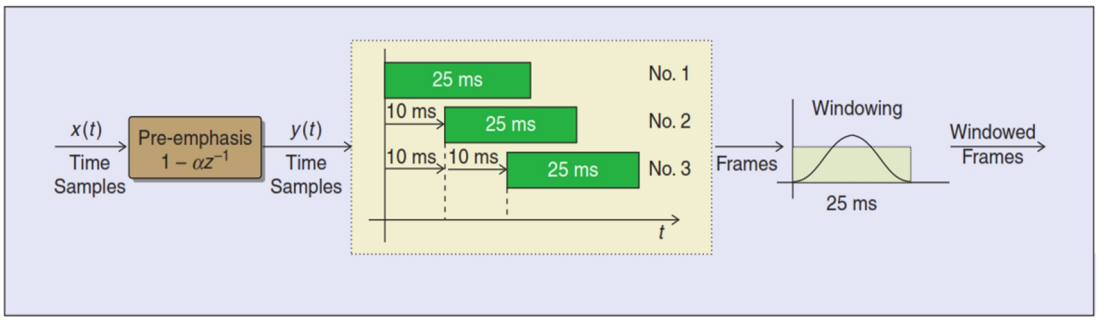


Figure II.11. ANALYSIS BLOCK DIAGRAM FOR FRAMING [17]

B. MEL-FREQUENCY SPACED CEPSTRAL CO-EFFICIENTS FEATURES (NFCCS)

The filterbank which is an array of band-pass filters that separates the input signal into multiple components, models the ability of the human ear to resolve frequencies nonlinearly across the audio spectrum and decreases with higher frequencies, as shows fig II.12, MFCC features are derived as follows:

a) Fourier Transform,

b) Mel- spaced filter bank values,

c) Cepstral analysis, and finally.

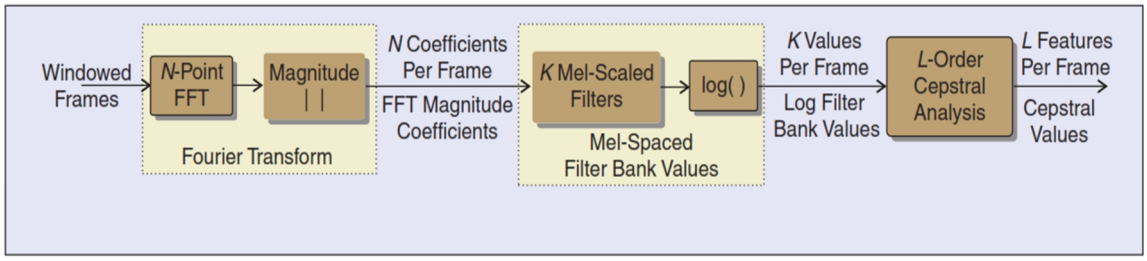


Figure II.12. ANALYSIS BLOCK DIAGRAM FOR MFCC[17].

C. CONCATENATION WITH TEMPORAL DERIVATIVES

In speech recognition the dynamic information also plays a role in helping to identify speaking styles and durations (albeit in a very simplis- tic fashion compared to prosodic cues). Define $\vec{c}_t$ as the compensated MFCC feature vector at frame time index t. The first-order derivative or "delta" feature is approximated by [18]:

$$\vec{d}_t = \frac{\sum_{P=1}^{P} p\left(\vec{c}_{t+p} - \vec{c}_{t-p}\right)}{2\sum_{P=1}^{P} p^2} \qquad \text{(II.6)}$$

Where typically P=2. By replacing $\vec{c}_t$ by $\vec{d}_t$ one can similarly derive the second-order delta-delta or "ac-celeration" parameters, $\vec{a}_t$.

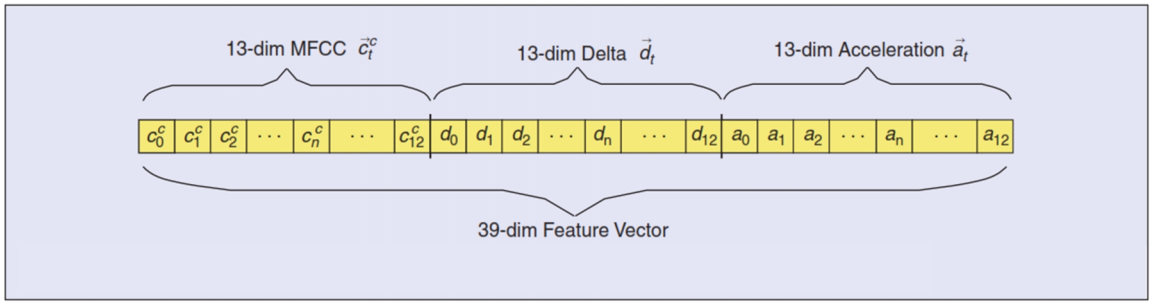


Figure II.13. CONCATENATION OF MFCC [17].

These temporal derivatives are concatenated with the original MFCC co-efficient to yield an augmented feature vectors .This feature parameterization is elucidated in fig II.13"[17].

II.5.2 PROSODIC FEATURES

Generally prosodic features are related to changes in pitch, intensity, and duration. Since they spread over more than one phoneme segment, prosodic features are super segmental. The creation of prosodic features depends on source factors or vocal-tract shaping factors [19]. The source factors are changes in the speech breathing muscles and vocal folds, and the vocal-tract shaping factors relate to the upper articulators movements [10].

II.5.3 VOICE QUALITY FEATURES

Voice Quality features include jitter, shimmer and harmonics-to-noise ratio. Jitter and shimmer are micro fluctuations in vocal fold frequency and amplitude. They are correlated to rough or hoarse voice quality [20].

II.5.4 STATIC FEATURES [10]

Static feature vectors are derived per speaker turn by a projection of each uni-variate time series X onto a scalar feature x of real value (R^1) independent of the length of the turn [21].

$$F : X \rightarrow x \in R^1$$

Functional F includes statistical functional, regression coefficients and transformations are applied to each contour on the turn-level [22][23].

II.6 SPEECH RECOGNITION APPROACH

Basically there exist three approaches of speech recognition. (Fig II.14) [14]:

1. Acoustic Phonetic Approach
2. Pattern Recognition Approach
3. Artificial Intelligence Approach

- The acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time.

- The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm.

- The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic Phonetic and Pattern Recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram.

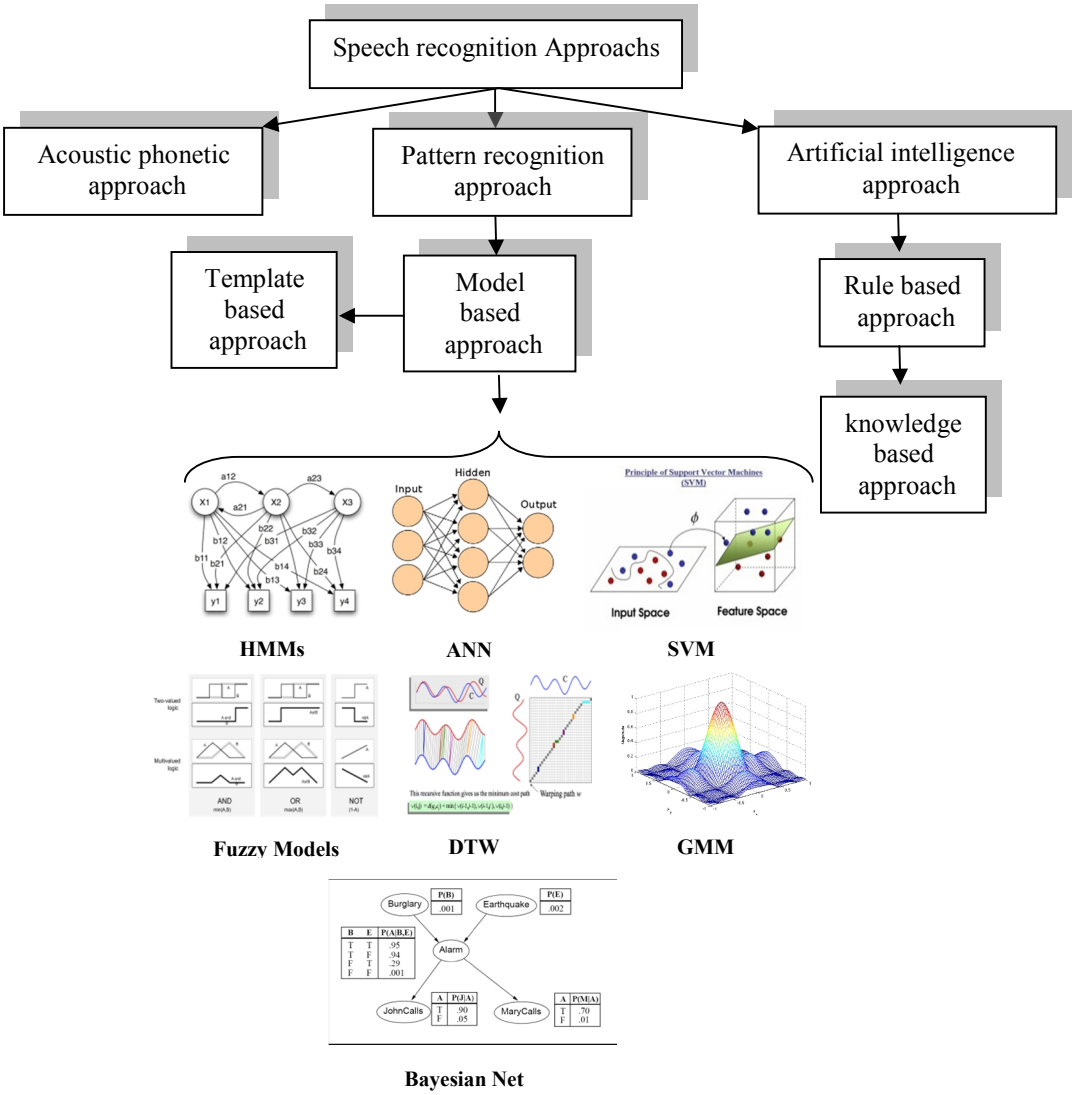


Figure II.14. TAXONOMY OF SPEECH RECOGNITION

II.7 CLASSIFIER COMBINATION

Several reasons for combining multiple classifiers to solve a given ASR problem. Some of them are listed below [24]:

- "Different classifiers trained on the same data may not only differ in their global performances, but they also may show strong local differences.
- Some classifiers such as neural networks show different results with different initializations due to the randomness inherent in the training procedure. Instead of selecting the best network and discarding the others, one can combine various networks, thereby taking advantage of all the attempts to learn from the data"[24].

In summary, we may have different feature sets, different training sets, different classification methods or different training sessions, all resulting in a set of classifiers whose outputs may be combined, with the hope of improving the overall classification accuracy. A large number of combination schemes have been proposed for pattern recognition approach in the literature [25].

Various schemes for combining multiple classifiers can be grouped into three main categories according to their architecture: 1) parallel, 2) cascading (or serial combination), and 3) hierarchical (tree-like) [jean]. Using these three basic architectures, we can build even more complicated classifier combination systems.

II.8 ARABIC SPEECH RECOGNITION

II.8.1 THE ARABIC LANGUAGE

The Arabic language had limited number of research efforts, although it is one of the oldest languages in the world and the fifth widely used language nowadays [26] with an estimated number of 350 million speakers covering a large geographical area. Arabic is a Semitic language and one of the six official languages in the United Nations (UN) and is one of the most widely spoken languages in the world. Statistics show that it is the first language (mother-tongue) of more than 300 million native speakers ranked as fourth after Mandarin, Spanish and English [27]. Since it is also the language of religious instruction in Islam, many more speakers have at least a basic knowledge of Arabic language. "Arabic" not a single linguistic variety; rather, it is a collection of different dialects, as shown in (figure II.15). Modern Standard Arabic (MSA) is a version of Classical Arabic with a

modernized vocabulary. MSA is a formal standard common to all Arabic-speaking countries [28]. The Arabic script evolved from the Nabataean Aramaic script[1].

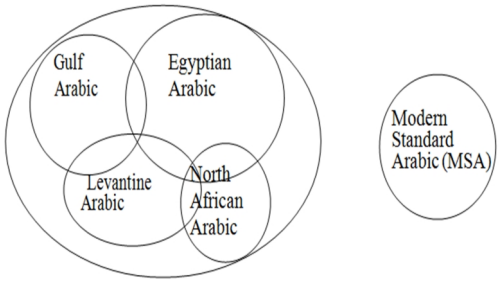


Figure II.15. ARABIC LINGUISTIC VARIETIES[28].

Arabic language have the bellow notable features:

- Type of writing system: *abjad.*
- Direction of writing: horizontal lines from right to left.
- Numerals are written from left to right.
- Number of letters: 28-29 letters .
- Most letters change form depending on whether they appear at the beginning, middle or end of a word.
- Vowel diacritics, which are used to mark short vowels, and other special symbols appear generally in the Qur'an.

The Appendix A presents a resume about the phonology and phonetic characteristics.

II.8.2 PROBLEMS FOR ARABIC AUTOMATIC SPEECH RECOGNITION

Many aspects of Arabic, such as the phonology and the syntax, do not present problems for automatic speech recognition. Standard, language-independent techniques for acoustic and pronunciation modeling, such as context-dependent phones, can easily be applied to model the acoustic-phonetic properties of Arabic. The most difficult problems in developing high-accuracy speech recognition systems for Arabic are the predominance of non-diacritized text material, the enormous dialectal variety, and the morphological complexity [28].

[1]http://www.omniglot.com/writing/arabic.htm

In particular, the complexity of Arabic morphology is well known for presenting a problems in language modeling, because of this high number of prefixes and suffixes that can be grafted to a root which leads an explosion of forms that can be associated with a word [19].

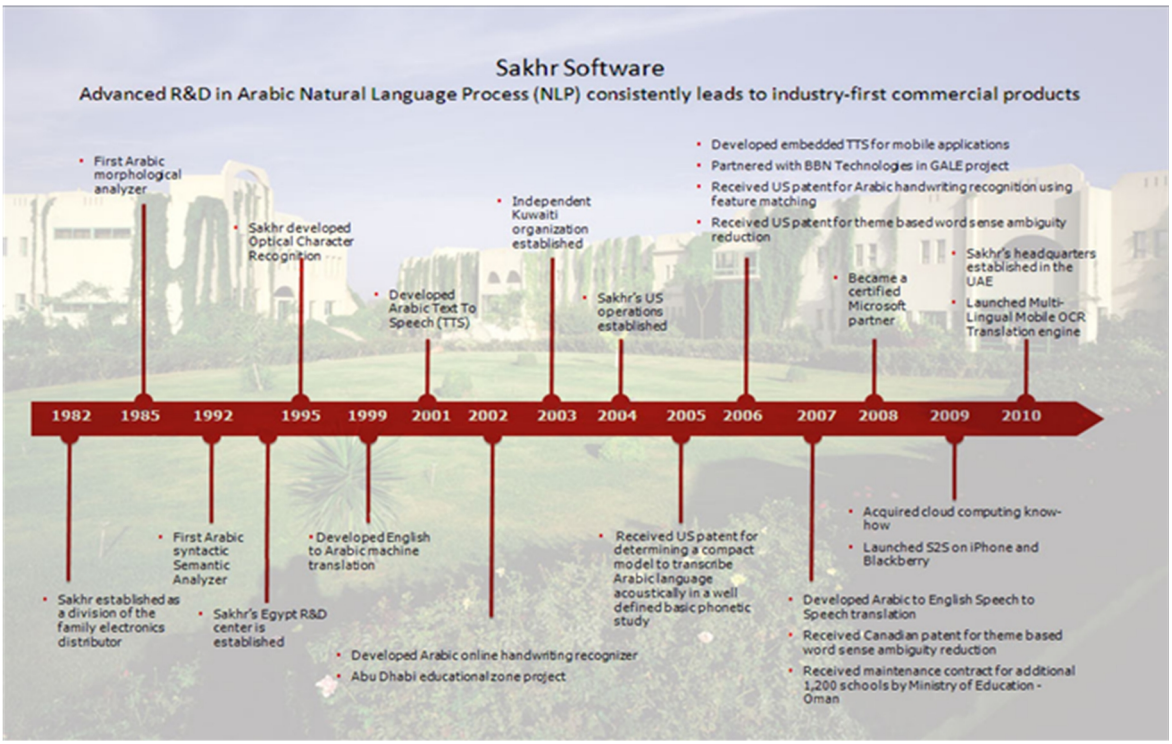


Figure II.16. SAKHR COMMERCIAL SOFTWARE [30].

II.8.3 PREVIOUS WORK IN ARABIC SPEECH RECOGNITION

Almost of work in Arabic ASR has focused on Modern Standard Arabic. We emphasize in particular the work of [29] [31], which focuses on acoustic , the syntactic appearance, and structural investigations of Arabic sounds. The work given in [32] enabled the realization of recognition system MARS at the University of Poitiers. In [33], a morphology-based language model was investigated . In [28], the discrepancies between dialectal and formal Arabic in a speech recognition was investigated by the. In [34], the authors reported the feasibility of using the automatic diacritizing Arabic text in acoustic model training for ASR. In [35], an attempted to use Carnegie Mellon University (CMU) Sphinx speech recognition system, to develop an extension useful for Arabic language. In [36]-[46] some recent works on ASR for Arabic.

In the ASR applications side we finds very diverse commercial domain for author languages, but it is in the Arabic language remains limited and narrow exception of certain

applications, such as the applications developed but Sakhr company as shown on (Figure II.16) [30] . In addition to some of the applications that take care of the Quran like "Hafss" software developed by RDI's[1] Speech Processing Technologies.

The goal of Hafss is to afford an effective assistance in the self-learning of spoken language. Hafss which has proved effective in helping people learn Tajweed of the glorious Quran is also helpful in learning spoken Arabic and other languages as well.

II.9 CONCLUSION

In this chapter we provide a general literature review of the digital speech processing field and the fundamentals of ASR. We discuss the main components required for building an ASR system. The chapter presents the Arabic language and the basic properties of Arabic phonology and a current overview of ASR with significant focus given to the research effort made regarding the Arabic language. Feature extraction techniques are presented. LPC was an efficient method for coding of speech; however, MFCCs based on human perceptual and auditory processing is simple and relatively fast to compute. It has become the standard feature and the choice for many speech recognition applications despite same negative characteristics such as non-direct dependence on the obtained feature space in the speech data and the performance degradation on the noisy speech signal. The positive and negative aspects of the feature extraction for Arabic speech are the same as for other languages. Other methods have been developed to extract features of Arabic speech; however, none of them has surpassed the MFCCs method. The work [47] reviews and compares several methods that could be applied to Arabic speech. The basic approaches in speech recognition are discussed briefly. Concerning the methods for automatic speech recognition, the emphasis on one model and the assertion of its absolute advantage over the others is contested and is the subject of controversy and assertions. In cases in which a method is superior to others in the benefits it offers or excels in one particular aspect, it frequently fails in others, according to the nature of the speech database and its application .

This effect has perhaps been the motivation for a noticeable shift toward a hybrid method that combines more than one way and more than one rationale (Section 2.2).

[1]RDI : Research & Development International [http://www.rdi-eg.com/technologies/speech.htm]

Most of the methods used currently provide satisfactory results and have advanced practical and industrial applications in the field of speech recognition. Statistical models comprise the classical approach to speech recognition and are the major contributor to and fundamentally employed method for many recent pattern recognition applications. They have not had an effect in the field of automatic speech recognition. New approaches in this field are investigated for ASR in the second part of this thesis.

Part II

Proposed Contributions

Chapter III

Tree Distributions Approximation Model (TDA)

III.1 INTRODUCTION

"*A graph is a two-dimensional visual formalism that can be used to describe many different phenomena. Graphs are used in a wide variety of fields, including computer science, data and control flow, entity relationships and social networks, Petri and neural networks, software/hardware visualization, and parallel computation. The popularity of graphs is in large part due to their ability represent complex situations in an intuitive and visually appealing way*." *(Bilmes & Bartels)* [48].

This chapter introduces a new discrete speech recognition method that investigates the capability of graphic models based on tree distributions, which are widely used in many areas of optimisation. A novel spanning tree structure that utilises the temporal nature of speech signalling is proposed.

III .2 GRAPHICAL MODEL

An exciting development over the last decade has been the gradually widespread adoption of probabilistic graphical models (PGMs) in many areas of computer vision and pattern recognition [66]. The graphical model is a simple way to efficiently express a novel complicated idea in an intuitive, concise, mathematically precise way, and to speedily and visually communicate that idea. Moreover, with the right software, it is possible to readily prototype that idea on a standard desktop workstation [48].

The same approaches modeling the emission probability distributions attempt to obtain an accurate and efficient way of capturing complex inter-feature dependencies. For the discrete case, the process of modeling multivariate distributions is a very interesting and complex problem, especially when dealing with a large number of variables and their complex interactions. In the last decades, much effort has been made in the graphical model framework. In recent years and due to a wide range of research, graphical models have offered a mathematically formal approach yet a highly flexible means for solving many of the problems in speech and language processing [48]-[51].

One of the techniques within the graphical model framework that has received special attention is the probabilistic model based on tree structures that provides a powerful and intuitive framework for formulating several problems in machine learning and its application[52]-[56]. Large storage and high computational capacity are needed for small and medium vocabularies and to achieve high accuracy of recognition.

An indeed more critical aspect of current ASR systems is robustness which is investigated through a probabilistic model based on tree distributions approximation (TDA) in discrete speech recognition introduced in [57]. This chapter demonstrates the capability of TDA models while using a proposed tree structure to improve the learning and the inference step.

III.3 MODELING WITH TREE DISTRIBUTION APPROXIMATION

An attractive feature of the graphical model is that it captures and represents the relations among the data quite efficiently [48]. These include direct graphs (Bayesian network or belief networks), indirect graphs (Markov random fields or Markov networks), and a combination between the previous models, which is called a chain graph. A critical choice to be made when selecting a probabilistic model is in regard to its complexity. In practice, most structure learning methods are heuristic methods that perform local searches by starting with a given graph and improving it by adding or deleting one edge at a time.

There is an important case in which both parameter learning and structure learning

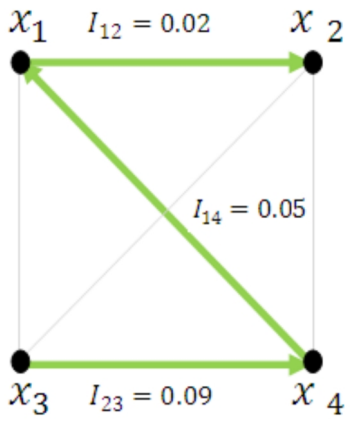


Figure III.1. THE MAXIMUM WEIGHT-SPANNING TREE OF FOUR RANDOM VARIABLES USING MUTUAL INFORMATION AS WEIGHT.

are tractable, namely, the case of graphical models in the form of a tree distribution. Chow and Liu [58] introduced an algorithm for fitting a multivariate distribution with a tree, i.e., a density model, which assumes that there are only pair-wise dependencies between variables and that the graph of these dependencies is a spanning tree (Fig III.1).

Chow and Liu demonstrated that the tree distribution that maximizes the likelihood of a set of observations on M nodes as well as the parameters of the tree can be found in

time quadratic in the number of variables in the domain. Graphical probability models of tree distribution enjoys many properties that make them attractive as modeling tools: they are intuitively appealing, have low complexity and yet a flexible topology, the sampling and computing likelihoods of trees are in linear time and they have existing efficient and simple algorithms for marginalizing and conditioning [59] .

To deal with the aforementioned issues, this chapter presents an automatic method for discrete speech recognition using the TDA model. We will discuss another approach where we benefit from the temporal nature of the speech signal to propose a tree structure that provides fast learning by reducing the complexity.

III.3.1 TREE DISTRIBUTION

Let V denote a set of n discrete random variables of interest. For each random variable $v \in V$, let $\delta(v)$ represent its range, $x_v \in \delta(v)$ a particular value.
$x = (x_1, \dots, x_n)$ denotes an assignment to the variables in V . Let's consider a complete non oriented graph $G(V, E)$ corresponding to the n variables, where E is a set of edges. Two neighbor vertices u and v are noted $u \sim v$.

- ***Proposition***

If graph G was a tree (a connected graph without loops) parameterized in the following way:

For $u, v \in V$ and $(u, v) \in E$, let T_{uv} denote a joint probability distribution of u and v. We require these distributions to be consistent with respect to marginalization, denoted by $T_u(x_u)$ the marginal of $T_{uv}(x_u, x_v)$, or $T_{vu}(x_v, x_u)$, with respect to x_v for any $v \neq u$. We now assign a distribution T to the graph $G(V, E)$ as follows [60]:

$$T(x) \approx \prod_{(u \sim v) \in E} \frac{T_{uv}(x_u, x_v)}{T_u(x_u)\, T_v(x_v)} \prod_{u \in V} T_u(x_u) \qquad (III.1)$$

The distribution itself will be called a tree when no confusion is possible.

III.3.2 LEARNING OF TREE DISTRIBUTION BASED ON MWST

The learning problem is formulated as follows: given a set of observations $X = \left(x^{(1)}, \dots, x^{(N)}\right)$, where each $x^{(i)}$ represents a vector of observations for all the variables in V.

We want to find the optimal tree distribution T^* that satisfies:

$$T^* = argmax \sum_{i=1}^{N} log\, T^*(x^i) \qquad \text{(III.2)}$$

This problem is a special case of a more general task of fitting a tree to a distribution P by minimizing the Kullback-Liebler (KL) divergence ;

$$KL(P||T) = \sum_{x} P(x) log \frac{T(x)}{P(x)} \qquad \text{(III.3)}$$

Where P is an empirical distribution obtained from an iid set of data.

We learn the model by maximizing the log-likelihood for X. Chow and Liu showed that the maximum weight-spanning tree (MWST) using mutual information I_{uv} as the weight for the edge $u \sim v$, maximizes the likelihood over tree distributions T for X.

The algorithm is summarized in Table III.1.

III.3.3 INFERENCE

Chow and Liu Algorithm ($\approx$)
Input: Distribution $\overline{\text{rithm}}$ er domain V, procedure MWST (Weights) that outputs a maxi$_{\text{num}}$ weight-spanning tree over V. 1.Compute marginal distributions T_u, T_{uv} for $u, v \in V$ 2.Compute mutual information values I_{uv} for $u, v \in V$ 3.T= $MWST(\{I_{uv}\})$ 4.Set $T_{uv} \equiv P_{uv}\, for\, u, v \in V$ 5.**Output**: T

Table III.1
THE CHOW AND LIU ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATION OF TREE STRUCTURE AND PARAMETERS

Import a source given an M set of observations :

$\left\{ X_1 = \left(x_1^{(1)}, \dots, x_1^{(N_1)}\right), \dots, X_M = \left(x_M^{(1)}, \dots, x_M^{(N_2)}\right)\right\}$ that represent M classes, for each class j, $j = 1, \dots, M$ T_j represent the correspondent distribution probability obtained by (TDA).

We denote the class of a given vector $x = (x_1, \dots, x_n)$, The quantity $P(C_d = j|x_d)$ quantifies the belief for the appurtenance of x to the class $C_d = j, j = 1, \dots, M$.

The expected classification error can be minimized by choosing $Argmax_j\big(P(C_d = j|x)\big)$.

According to Bayes' theorem:

$$P(C_d = j|x) = \frac{P(x|C_d = j)P(C_d = j)}{P(x)} \qquad (III.4)$$

Moreover,

$$P(x) = \sum_{j=1}^{j=M} P(x|C_d = j)P(C_d = j) \qquad (III.5)$$

In which

$$P(x|C_d = j) \approx \prod_{(u\sim v)\in T_j} \frac{T_{j_{uv}}(x_u, x_v)}{T_{j_u}(x_u)\, T_{j_v}(x_v)} \prod_{u\in V} T_{j_u}(x_u) \quad = T_j(x) \qquad (III.6)$$

Suppose

$$\forall\, i, j = (1, \dots, M)\ P(C_d = j) \approx P(C_d = i)$$

Therefore

$$P(C_d = j|x) \approx \frac{T_j(x)\, P(C_d = j)}{P(x)} \qquad (III.7)$$

All the elements of "(III.5)" and "(III.6)" are previously computed in the learning setup.

III.4 TREE DISTRIBUTION APPROXIMATION MODEL FOR ASR

In this section, we introduce an adapted TDA for the task of speech recognition. One of the main drawbacks of the Chow-Liu algorithm is the enormous computational complexity needed to obtain the optimal spanning tree. Computing marginal from data is a computationally expensive step and takes $O(n^2N)$ operations. Further, to find the tree structure, given by its set of edges E, one has to compute the mutual information between each pair of variables in V under the target distribution P.

$$I_{uv} = \sum_{x_u x_v} T_{uv}(x_u, x_v) log \frac{T_{uv}(x_u, x_v)}{T_u(x_u)\, T_v(x_v)}$$

$$u, v \in V, u \neq v \qquad \text{(III.8)}$$

As V has n variables, there are $\mathrm{n(n-1)/2}$ mutual information to be computed and n^{n-2} trees with n vertices.

Next, the optimal tree structure E is found by a maximum weight spanning tree (MWST) algorithm using I_{uv} as the weight for edge $(\mathrm{u,v})\ \forall\, u, v \in V$.Though there are

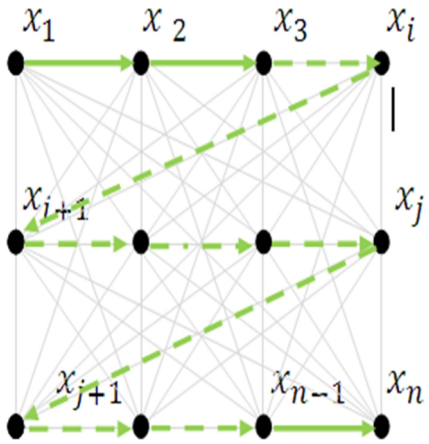


Figure III.2. PROPOSED TREE STRUCTURE FROM N^{N-2} POSSIBLE TREES. TO LEARNING THE TDA MODEL FOR SPEECH RECOGNITION.

several ways to compute MWST, we have used the Kruskal algorithm [61]. An assumption that helps to develop an accelerated algorithm has been proposed in [59], achieving speed-ups of up to 2-3 orders of magnitude in the experiments.

As a natural way to alleviate the aforementioned drawback, we propose a new graphical tree structure inspired from the temporal nature of the speech signal, in which,

only the linear dependencies between the features are considered, see Fig III.2.

In this case, the "(III.8)" can be simplified to the following;

$$q(x) \approx \prod_{i=1}^{n-1} \frac{q_{i\,i+1}(x_i, x_{i+1})}{q_i(x_i)\, q_{i+1}(x_{i+1})} \prod_{i=1}^{n} q_i(x_i) \qquad (III.9)$$

Where i is the index of a vertices on the proposed tree (Fig III.2), $q_i(x_i)$ is the marginal of $q_{i\,i+1}(x_i, x_{i+1})$.

In contrast to the Chow-Liu algorithm, the proposed method doesn't require any computation of mutual information to find the MWST.

III.5 EXPERIMENTAL DESIGN

III.5.1 DATASET DESCRIPTION

- **Spoken Arabic digits (SAD)**

The data included Arabic digit corpus collected by the Laboratory of Automatic and Signals, University of Badji-Mokhtar - Annaba, Algeria[62]. Dataset from 8800(10 digits x 10 repetitions x 88 speakers) time series of 13 frequency cepstral coefficients (MFCCs) were taken from 44 male and 44 female Arabic native speakers between ages 18 and 40 to represent ten spoken Arabic digit.

- **Japanese Vowels (JV)**

The data set of Japanese Vowels is introduced in [63]. It is a speaker recognition problem. Nine male speakers uttered two Japanese vowels /ae/ successively. For each utterance, a 12-degree linear prediction analysis was applied to obtain discrete-time series with 12 LPC cepstrum coefficients. This means that one utterance by a speaker forms a time series whose length is in the range of 7-29 and each point of a time series is of 12 features (12 coefficients).

III.5.2 FEATURE EXTRACTION AND PARAMETER ESTIMATE

In the signal analysis phase, the input speech signal is transformed into feature vectors containing spectral and/or temporal information using Mel frequency cepstral coefficients (MFCCs) for the Arabic data set and 12-degree linear prediction (LPC) cepstrum coefficients for the Japanese data-set (Sec II.5.1).

Parameter	Value
Sampling rate	11025 Hz, 16 bits
Pre-emphasis	0.97
Window type	Hamming

Table III.2. MFCC SYSTEM PARAMETERS FOR ASD DATASET

Tables III.2 and III.3 show several system parameters adopted for such a task. The result of the feature extraction is a series of vectors, characteristic of the time-varying spectral properties of the speech signal.

Parameter	Value
Sampling rate	10 KHz
Frame length	25.6 ms
Shift length	6.4ms

Table III.3. LPC CEPSTRUM COEFFICIENT SYSTEM PARAMETERS FOR JV DATASET

The properties can be mapped into discrete vectors by quantizing them using vector quantification (VQ)[64] . A discrete HMM or TDA speaker independent isolated word recognition system can be described as a two-step modeling process. In the first step, vector quantization (VQ), is used to classify the speech signal space into U regions, where U is the codebook size or the number of models generated in this step. Each region is represented by a typical vector, usually the centroid vector for that region. The codebook is then composed of these typical vectors. The second step is used to produce a set of reference models that represent the possible sequences of the quantized observation vectors from the codebook generated by the first step.

This model requires a scalar sequence *X.* vector quantization divides the *D*-dimensional Euclidian space R^D into unempty subsets $V_i, i = 1, \dots, U$.

$$\bigcup_{i=1}^{U} V_i = R^D, V_i \cap V_j = \emptyset \ (i \neq j), \qquad (III.10)$$

Where $\emptyset$ is an empty set.

The operation that maps an input vector $x \in V_i$ into a predetermined vector Y_i in subset V_i is called VQ.

Denoting the operation of VQ as Q,

$$Q(\boldsymbol{x}) = y_i \; if \;\; x \in V_i, \qquad (III.11)$$

where y_i is called code, $\boldsymbol{C} = \{y_1, \dots, y_{|C|}\}$ is called codebook, and ($|\boldsymbol{C}| = U$) is the codebook size. Codebook is pre-designed and stored in the system memory. To this end, we propose to adopt the well-known *k-means* clustering algorithm; the subset V_i in "(III.10)" is obtained by the following nearest neighbor condition,

$$V_i = \{x \in R^D, d(x, y_i) \leq d(x, y_j) \; ; j = 1, \dots, |C|\} \quad \text{(III.12)}$$

Where $d(.)$ denotes the Euclidian distance. Further, we try to build an HMM model. To put the initial model parameters, we have to be careful as one might easily slip into divergence with bad model initialization.

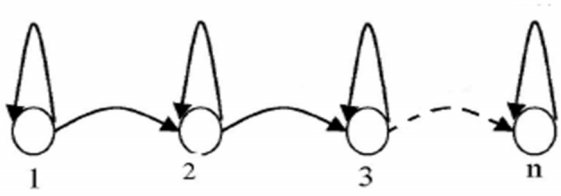


Figure III.3. N-STATE LEFT-TO-RIGHT SEQUENTIAL HMM STRUCTURE

The problem with discrete observation HMMs is that they are less effective given that we can initialize the model parameters with random values. In continuous density HMMs (CHMMs), the problem is more serious and the parameters should be judiciously selected to get rid of the divergence fate. The N-state constrained left-to-right sequential HMMs structure was employed as shown in Fig III.3.

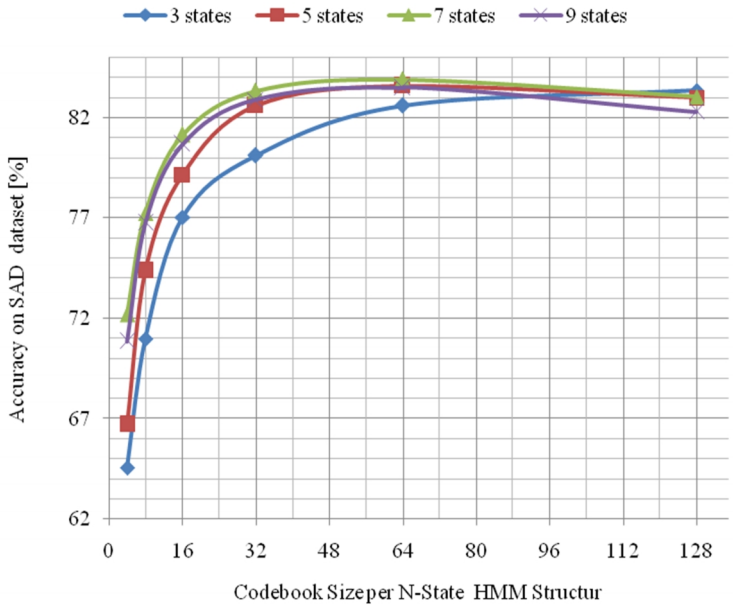


Figure III.4. CV ACCURACIES FOR DHMM PARAMETERS [STATES AND CODEBOOKE SIZE] ON SPOKEN ARABIC DIGIT DATASET

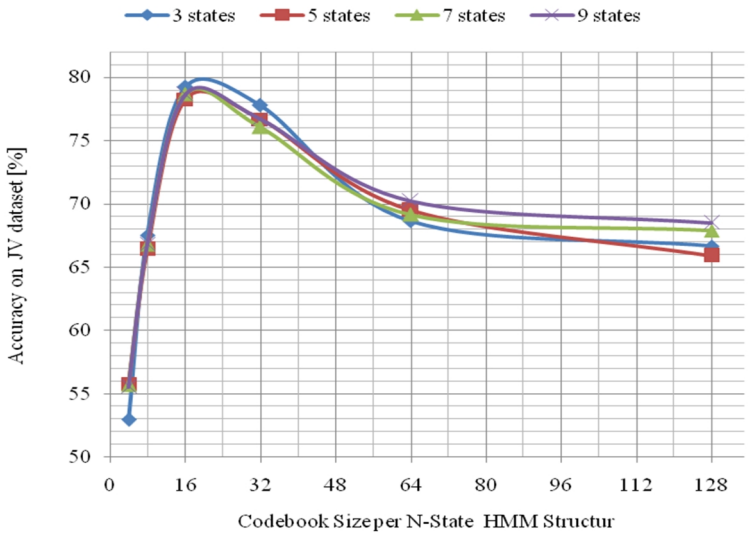


Figure III.5. CV ACCURACIES FOR DHMM PARAMETERS [STATES AND CODEBOOK SIZE] ON JAPANESE VOWEL DATASET

We propose the cross validation (CV) method to estimate the optimal number of clusters (initial set of code words in the codebook), the number of states, and the number of mixtures for CHMM. The capability of CV was previously presented in [65].

The procedure for the m-fold cross validation is as follows:

1) Split the training data into m roughly equal-sized parts.
2) For the i^{th} part, train the classifier model to the other $(m - 1)$ parts of the data using k clusters, and test the classifier for the i^{th} part of the data.
3) Continue to do so for $i = 1, \cdots, m$, and take the average of m results.

As shown in Fig III.4, the best parameter for the DHMM is 64 as codebook size and 7 states for the Arabic data base (SAD), while for the Japanese data-set (JV) the best parameters are 16 as codebook and 3 states, Fig III.5. For TDA models, the best parameters are 32 as codebook for SPA and 128 for JV. The local optimum size of the codebook obtained using the 3-fold cross validation on the training set was employed for discretization of data.

Figs.III.6 and III.7 show the accuracy results for various parameters of the CHMM, 7 stats and 2 mixtures are selected for the SAD data set and 5 states and 4 mixtures for the JV dataset.

III.6 RESULTS AND DISCUSSIONS

III.6.1 TREE DISTRIBUTION APPROXIMATION AND HMMS RESULTS

The algorithms developed (TDA, HMM) were applied to data sets SAD and JV, and the results summary are given in Tables III.4 and III.5. The analyses of results show that the recognition performance of TDA (MWST) and TDA (proposed structure) give better results than DHMM. This latter is a reference in discrete speech recognition and recognized as commonly used, not merely for its reasonable performance and potential low computation cost, but also because it is a parametric model of the speech signal that can model various events (phonemes, syllables, etc.) in the speech signal [67].The improvement achieves 2.54% for spoken Arabic digit and 12% for Japanese vowels.

Japanese Vowels Classes	Success Rate [%]			
	TDA (MWST)	TDA(proposed Structure)	DHMM	CHMM
Speaker_1	95.62	96.77	86.02	96.77
Speaker_2	93.33	94.28	95.24	93.38
Speaker_3	93.83	94.32	86.74	97.35
Speaker_4	97.72	100.0	70.46	98.49
Speaker_5	100.0	96.55	82.76	100.0
Speaker_6	100.0	100.0	94.44	100.0
Speaker_7	96.65	95.00	84.17	98.17
Speaker_8	94.00	88.00	66.67	100.0
Speaker_9	86.20	93.10	88.51	96.40
Average	**95.26**	**95.34**	**83.89**	**97.84**

Table III.4
RECOGNITION RESULTS OF JV SPEAKERS, CLASSIFICATION BY TDA AND HMMS

Arabic Digits Classes	Success Rate [%]			
	TDA (MWST)	**TDA**(proposed structure)	DHMM	CHMM
Syfr - 0	85.55	84.04	90.69	93.28
Wahid- 1	98.36	98.27	96.92	99.95
Ethnan- 2	92.91	92.55	90.86	90.19
Thalath'a- 3	94.09	94.09	84.59	92.16
Arb'a- 4	90.00	91.19	94.29	94.59
Khams'a- 5	94.00	94.45	94.04	97.62
Syt'a- 6	94.01	94.18	94.75	95.35
Sab'a-7	90.18	89.45	86.99	89.27
Thamany'a- 8	99.00	99.00	88.18	92.98
Tes'a- 9	93.50	93.72	86.61	95.51
Average	**93.16**	**93.09**	**90.79**	**94.09**

Table III.5
RECOGNITION RESULTS OF SPOKEN ARABIC DIGITS , CLASSIFICATION BY TDA AND HMMS

Although the proposed model deals with discrete data, experimental results indicate considerable competitive with CHMM through the closed overall recognition accuracies (93.16 % vs 94.09% for SAD and 95.26% vs 97.84% for JV).

	TDA (MWST)		TDA(proposed structure)		DHMM		CHMM
	Learning Time [Second]	Code book size	Learning Time [Second]	Code book size	Learning Time [Second]	Code book size	Learning Time [Second]
SAD Data set	**480.36**	**32**	**19.12**	**32**	**540.15**	**64**	**2485.51**
JV Data set	**156.5**	**128**	**10.05**	**128**	**12.35**	**16**	**49.92**

Table III.6. LEARNING TIME BY TDA AND HMMS

However, the proposed model (TDA with the proposed structure), while being quite easy to implement, gives performance recognition comparable to the results given by TDA with MWST, which are complex and computationally expensive.

It is worth mentioning that from Table III.5 there is a noticeable difference in the results of the recognition of the digit (zero) by the TDA (proposed structure) and CHMM (84.04% and 93.28%).

This difference can be explained by the nature of the pronunciation of this digit in the Arabic language (spoken Arabic digits are polysyllabic, except the 0 (/sifr/)).

III.6.2 COMPUTATIONAL COMPLEXITY

The proposed approach algorithm was implemented in Matlab and the run time using an Intel(R) Core (TM) 2 CPU 2.00-GHz processor was measured for different parameter choices.

	TDA (MWST)		TDA (proposed structure)		DHMM		CHMM
	Inference Time [Second]	Code book size	Inference Time [Second]	Code book size	Inference Time [Second]	Code book size	Inference Time [Second]
SAD Data set	**41.78**	**32**	**5.11**	**32**	**35.25**	**64**	**71.08**
JV Data set	**3.01**	**128**	**1.23**	**128**	**2.86**	**16**	**9.94**

Table III.7. INFERENCE TIME BY TDA AND HMMS

The resulting real learning time factors in Table III.6 show the impressive gain in time computation for the proposed TDA to speech recognition. This letter was capable of achieving speed-ups of up to 25 orders of magnitude in the experiment (19.80 seconds for the proposed technique against 480.32 seconds obtained by Chow Liu in the Arabic dataset), and it massively surpassed even HMMs that are considered relatively rapid.

This positive contribution of the proposed model solves the problem of expanding the learning data in real time.

Table III.7 shows that TDA model with the proposed tree structure is too fast and achieves 3 to 8 fold speed improvement compared to HMMs for JV data base and 7 to14 folds for SAD data base although the codebook size for the TDA is much larger then that for DHMM.

This advantage is very useful even on relatively high end embedded devices.

III.7 MFCCS CONCATENATION EFFECT (ΔΔMFCC)

III.7.1 ΔΔMFCC EXTRACTIONS AND PARAMETER ESTIMATE

Mel Frequency Cepstral Coefficients (MFCCs) are the most popularly used speech features in many speech and speaker recognition applications. In this Section, we study the effect of the second-order derivatives of MFCC on the recognition of the Spoken Arabic digits. In contrast of the precedent work on section (III.5.2) we use second (ΔΔ)- order derivatives of MFCC (sec II.5.1, 'C') approximately given by :

$$\Delta\Delta\mathrm{MFCC}_l(i) = [\Delta\mathrm{MFCC}_{l+1}(i) - \Delta\mathrm{MFCC}_{l-1}(i)] \qquad \text{(III.13)}$$

$\Delta\mathrm{MFCC}_l(i)$ is the first order derivative of MFCC

$$\Delta\mathrm{MFCC}_l(i) = G\sum_{k=-K}^{K} k(\Delta\mathrm{MFCC}_{l-k}(i)) \qquad \text{(III.14)}$$

Where k and l are frames indexes, i the MFCC component and G a gain factor selected as 0.375.Hence, the j^{th} frame of the digit D (i) is represented by an acoustic vector X_{ij}.

$$X_{ij} = \{\mathrm{MFCC}, \Delta\mathrm{MFCC}, \Delta(\Delta\mathrm{MFCC}), \log(\mathrm{Energy})\}$$

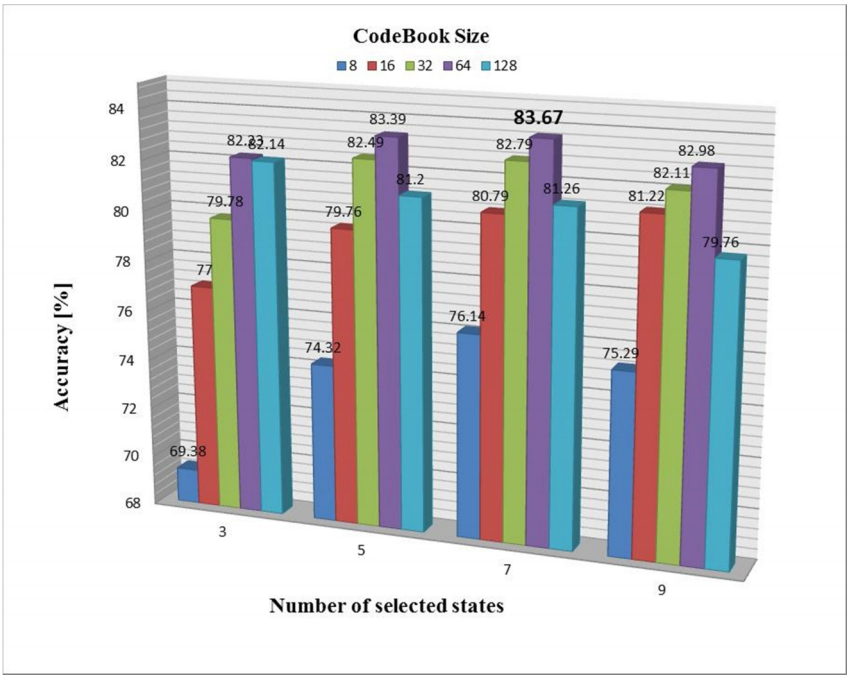


Figure III.6. CV accuracies for DHMM parameters [states and codebook size]

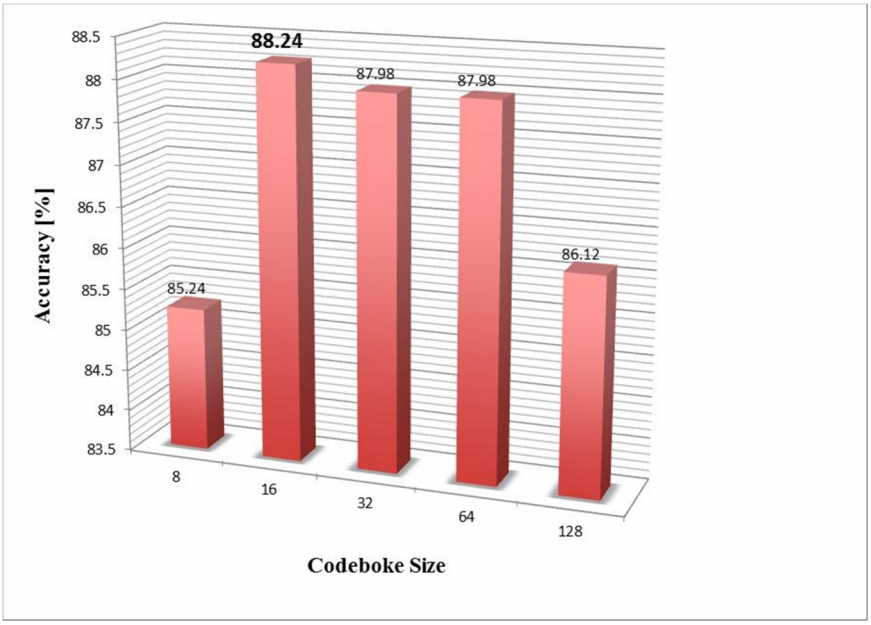


Figure III.7. CV ACCURACIES FOR TREE MODEL PARAMETERS [CODEBOKE SIZE]

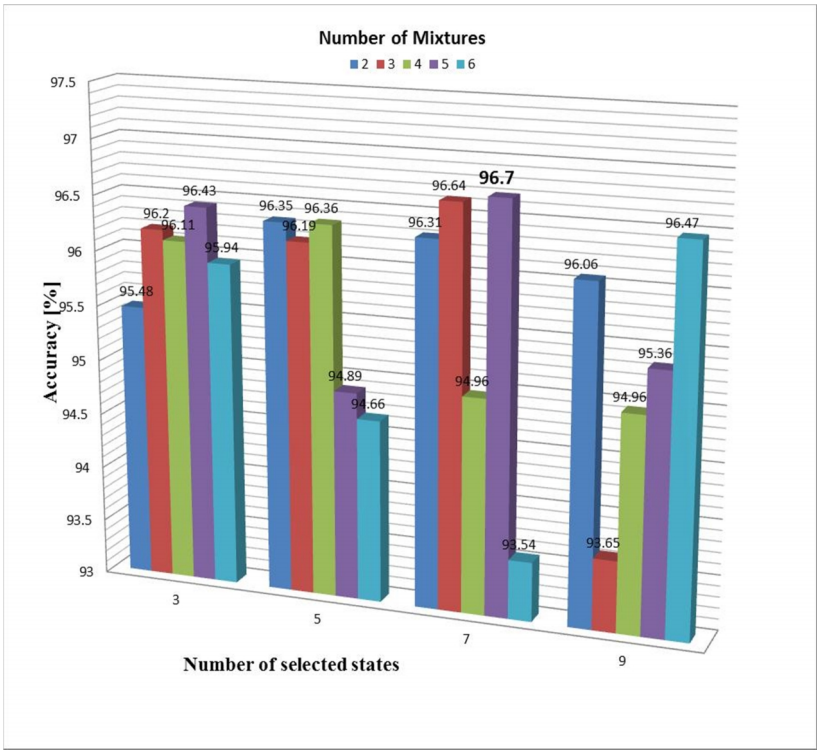


Figure III.8. CV ACCURACIES FOR CHMM PARAMETERS [STATES AND GAUSSIAN]

Arabic Digits Classes	Success Rate %	
	Dependence Tree Model using MFCC Parameters	Dependence Tree Model using $\Delta(\Delta^{MFCC})$
Syfr - 0	84.04	85.00
Wahid- 1	98.27	98.18
Ethnan- 2	92.55	92.72
Thalath'a- 3	94.09	94.75
Arb'a- 4	91.19	93.85
Khams'a- 5	94.45	92.73
Syt'a- 6	94,18	94.09
Sab'a-7	89.45	86.51
Thamany'a- 8	99.00	98.35
Tes'a- 9	93.72	95.00
Average	**93.09**	**93.11**

Table III.8. RECOGNITION BY TDA MODEL

The result of the feature extraction is a series of vectors, characteristic of the time varying spectral properties of the speech signal.

Doing the same experiment presented on section (III.5.2), The result of the clustering is a codebook of 16 for the tree model (optimum size obtained using 3-fold cross validation on the training set). Figure III.7 shows the accuracy results for different values of k.

Arabic Digits Classes	Success Rate %			
	DHMM using MFCC Parameters	D_{DHMM} using $\Delta(\Delta MFCC)$	CHMM using MFCC parameters	C_{HMM} using $\Delta(\Delta_{MFCC})$ Parameters
Syfr - 0	90.69	88.27	93.28	95.05
Wahid- 1	96.92	95.90	99.95	99.55
Ethnan- 2	90.86	89.54	90.19	100.00
Thalath'a- 3	84.59	85.05	92.16	99.54
Arb'a- 4	94.29	92.27	94.59	98.18
Khams'a- 5	94.04	92.12	97.62	100.00
Syt'a- 6	94.75	96.81	95.35	100.00
Sab'a-7	86.99	89.54	89.27	95.45
Thamany'a- 8	88.18	86.81	92.98	96.82
Tes'a- 9	86.61	92.27	95.51	99.55
Average	**90.79**	**90.86**	**94.09**	**98.41**

Table III.9 RECOGNITION BY HMMS

As shown in Fig III.6, the best parameter for the DHMM is 64 as codebook size and 7 states. Figures.III.8 shows the accuracy results for different parameters of the CHMM, 7 states and 5 mixtures are selected.

III.7.2 RESULTS AND DISCUSSIONS

The algorithms developed (TDA, HMMs) were applied to data sets SAD, and the results summary are given in Tables III.8 and III.9.

The analyses of results show that the impact of the Concatenation parameters of MFCC ($\Delta(\Delta MFCC)$) is not very important and do not affect the recognition performance for discrete speech (TDA and DHMM) , while the result for CHMM, the system using $\Delta(\Delta MFCC)$ obtains higher recognition rate than one employing the simple MFCC by 4.60%.

III.8 CONCLUSION

This chapter presents a new approach to discrete automatic speech recognition that accelerates graphic model learning based on tree distribution approximation using the proposed structure, which accurately and efficiently describes the interactions and dependencies among the acoustic features. We studied the effect of concatenation with temporal derivatives for the model. Different databases (Arabic and non-Arabic) have shown that for the task of isolated word recognition, accuracy is significantly improved compared to that of the DHMM. The proposed approach is of moderate computational complexity and suitable for online applications.

Chapter IV

Copula Function for Speech Recognition

IV.1 INTRODUCTION

Copula is a Latin noun that means "a link, tie, bond" (Cassell's Latin dictionary) and is used in grammar and logic to de- scribe "that part of a proposition which connects the subject and predicate" (Oxford English Dictionary) [78].
This chapter introduces a novel statistical approach, based on the copula function, for the problems of speech recognition in general and Arabic speech in particular. A general overview of the theory of copulas is presented in this chapter1, and we introduce the Gaussian copula with Gaussian mixture marginal distribution instead of the empirical one. We demonstrate the capability of the copula as an advanced statistical tool for modelling and estimating the multivariate probability density function of speech data. To evaluate the validity of the proposed models, we perform several experiments designed for the recognition of benchmark datasets and compare the results obtained to those of the hidden Markov model (HMM), a well-known statistical tool for modelling speech recognition. The classifier combines finite Gaussian mixture modelling for marginal distribution and the Gaussian copula. The experiment and results are presented.

IV.2 COPULA CONCEPT AND ASR

Introduced by A. Sklar in 1959 [68], copulas are functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas are multivariate distribution functions whose one-dimensional margins are uniform on the interval [0,1], they began to play an important role in probability and mathematical statistics. Although Copula has been known for a long time (Sklar, 1959) [68], the copula has only relatively recently been rediscovered in applied work, notably, in finance, insurance, and economics. It also recently started to appear in the field of image processing [69]-[76]. The copula was introduced by (Sklar, 1959) as a result of a question about the relationship between a multidimensional probability function and lower dimensional margins. It is a way of formalizing dependence structures of random vectors. A copula is a distribution function with the implicit capacity to model nonlinear dependencies via concordance measures, such as Kendall's τ.

[1]Basic tutorial and comprehensive overview of the theory of copula are given in [25]

For the ASR there are several approaches; namely, acoustic phonetic, statistical pattern recognition, and artificial intelligence approaches. The statistical approach is considered the major contributor and is the fundamentally employed method [67]. To the best of our knowledge, our dissertation is the first to use this technique for speech recognition following our seminar overview [77]. In other classification tasks, the efforts to concede the dependence on speech data are often carried out within a parametric modelling framework, in which the observed data are assumed to follow specific models such as a Gaussian probability distribution. Obviously, a multivariate Gaussian distribution assumes that the marginal distribution is univariate Gaussian distributed. However, this assumption is clearly impertinent in a large number of statistical learning problems. In multivariate distribution estimation settings, copulas have emerged as a useful tool to separate the joint dependence structure from the marginal distribution. This elegant separation also facilitates the creation of arbitrary distribution functions with known joint dependence structures, without imposing restrictions on the marginal distributions. Consequently, we can combine multivariate distributions to obtain a joint distribution with a particular dependence structure. We can construct any multivariate distribution from its marginal distributions and a copula [78].

IV.3 THE THEORY OF COPULA

IV.3.1 COPULA FUNCTION

A copula, C, is a joint distribution function of standard uniform random variables. That is,

$$\mathrm{C}(u_1, \dots, u_d) = P(U_1 \leq u_1, \dots, U_d \leq u_d), \qquad (IV.1)$$

where $U_i \sim U(0,1)\ for\ i = 1, \dots, d$

The copula density, $c(u_1, \dots, u_d)$, can be regarded as the joint probability density function (PDF) of multivariate standard uniform random variables $(U_1, \dots, U_d)$. Most copulas are exchangeable, thus implying that $c(u, v)$ is symmetric.

$c(u, v)$ must satisfy the following four properties:

1) $c(u, v) \geq 0, for\ [u, v]\ \in\ [0, 1]^2$; $(IV.2)$

2) $\int_0^1 c(u, v)du = 1, for\ 0 \leq v \leq 1$; $(IV.3)$

3) $\int_0^1 c(u, v)dv = 1, for\ 0 \leq u \leq 1$; $(IV.4)$

4) $c(u, v) = c(v, u)$. $(IV.5)$

A multivariate copula, $C(u_1, \ldots, u_d)$, defined on $[0, 1]$, is a multivariate cumulative distribution function (CDF) with univariate standard uniform margins:

$$C(u_1, \ldots, u\ ) = \int_0^{u_1} \ldots \int_0^{u_d} c(s_1, \ldots, s\ )\, ds\ \ldots\ ds_1 \qquad (IV.6)$$

Copulas do not always have densities, but when they do, copula density is given as follows:

$$c(u_1, \ldots, u\ ) = \frac{\partial}{\partial_{u_1} \ldots \partial_{u_d}} C(u_1, \ldots, u\ ) \qquad (IV.7)$$

Sklar's theorem states that the joint CDF, $F(x_1, \ldots, x_d)$, of a multivariate random variable, (X_1,… X), with marginal CDFs F_i, *i=1,…,d* can be written as:

$$F(x_1, \ldots, x\ ) = C(F_1(x_1),\ldots,F\ (x\ )), \qquad (IV.8)$$

where copula C is the joint CDF of

$$(U_1, \ldots, U\ ) = (F_1(X_1),\ldots,F\ (X\ )). \qquad (IV.9)$$

This indicates that a copula connects the marginal distributions to the joint distribution and justifies the use of copulas for building multivariate distributions.

If the joint distribution, F, is absolutely continuous with strictly increasing and continuous marginal (dfs); $F_1, \ldots, F_d$; an important consequence of Sklar's theorem is that the d-dimensional joint density f and the marginal densities f_i are also related:

$$f(x_1, \ldots, x\ ) = c(F_1(x_1),\ldots,F\ (x\ ))\ \prod_{i=1}^{d} f_i\ (x_i) \qquad (IV.10)$$

Equation (IV.9) shows that the product of marginal densities and copula densities build a d-dimensional joint density.

VI.3.2 WHY USE COPULA?

There are numerous advantages of using copula to model dependencies in multivariate data. First, note the decomposition of the joint distribution into the dependency structure (copula) and the constituent marginal distributions. This separation allows each variate to be described by an entirely different distribution (e.g. Gaussian, Pareto, Gamma, etc.)[81].

Further, unlike the correlation coefficient which measures co-variations up to the second order, copula functions capture the complete dependence structure. This feature is especially important when considering speech signal.

IV.3.3 CHOOSING A COPULA

Copulas exist in various shapes and forms. The most commonly used copulas are the parametric copulas, namely, the Gaussian, the Student's t copula (also called the t copula), and the Gumbel copula. Figure (IV.1) depicts scatter plots of various copula models, while (Fig. IV.2) Shows the copula density surface of different copula models from 1000 bivariate random variables generated from standard Gaussian distribution with $p = 0.4$.

Frequently, the choice of copula is based on the usual criteria of familiarity, ease of use, and analytical tractability. Gumbel copula can be used for extreme distributions, the Gaussian copula for linear correlation, and the Archimedean copula and the t-copula for dependence in the tail [78].

In chapter, we use the Gaussian copula, which is useful for its easy simulation method and generalization to multi-dimensions.

The Gaussian copula is derived from the multivariate Gaussian or Normal distribution. Other methods of construction may use geometry and the definition above to construct copula functions, such as the Frank copula.

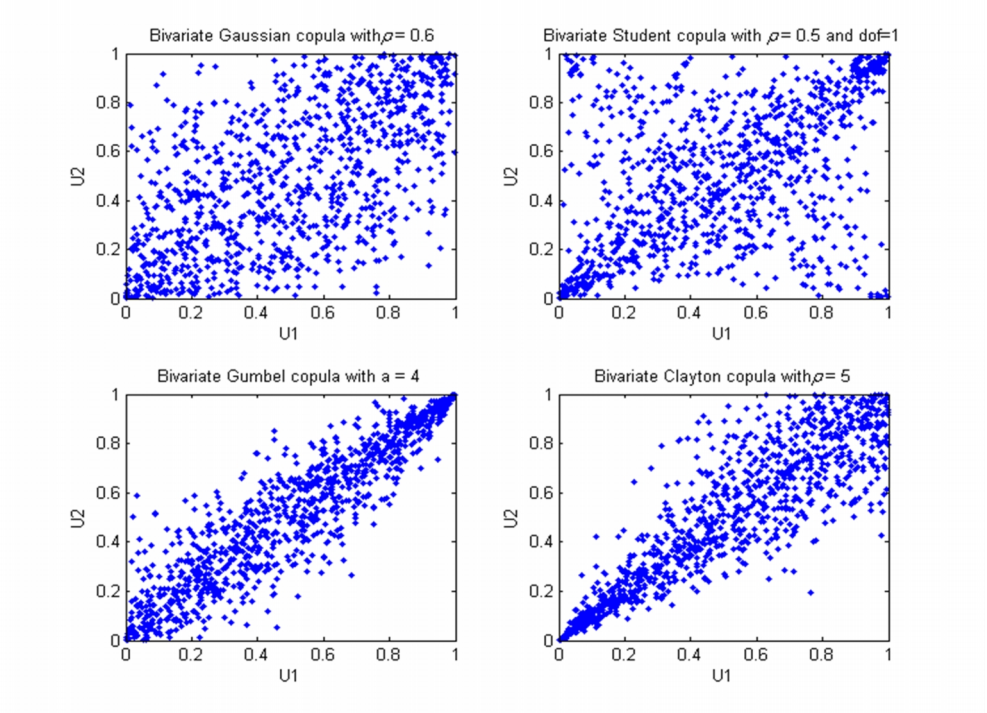


Figure IV.1. SCATTER PLOTS OF VARIOUS COPULA MODELS

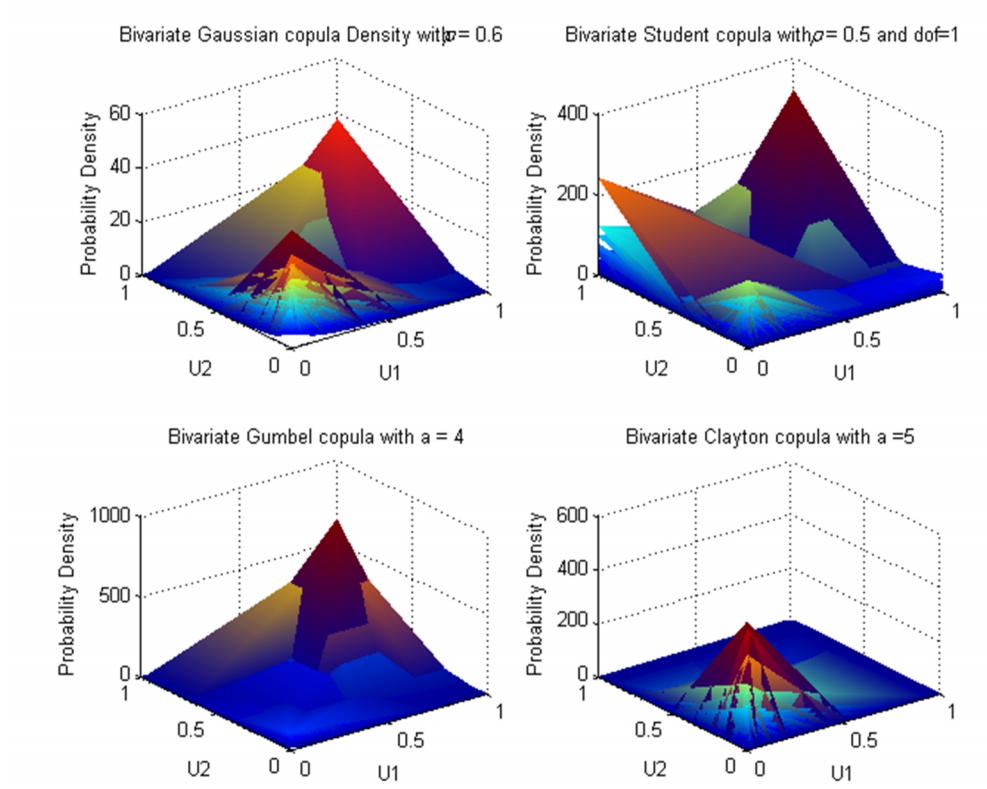


Figure IV.2. COPULA DENSITY SURFACE OF VARIOUS COPULA MODELS FROM 1000 BIVARIATE RANDOM VARIABLES GENERATED FROM STANDARD GAUSSIAN DISTRIBUTION WITH P = 0.4.

IV.4 Gaussian copula function

The Gaussian copula is derived from the multivariate Gaussian or normal distribution. These copulas are defined as follows:

$$C(u_1, \dots, u_d) = \varphi_\rho(\varphi^{-1}(u_1), \dots, \varphi^{-1}(u_d)) \qquad (IV.11)$$

$$= \int_{-\infty}^{\varphi^{-1}(u_1)} \dots \int_{-\infty}^{\varphi^{-1}(u_d)} \frac{e^{-\frac{1}{2}\gamma'\rho^{-1}\gamma}}{(2\pi)^{\frac{d}{2}}|\rho|^{1/2}} d\gamma_d \dots d\gamma_1, \qquad (IV.12)$$

where ρ is a symmetric, positive definite matrix with $diag(\rho) = 1$, φ_ρ is the standardized multivariate normal distribution with correlation matrix ρ, $\varphi^{-1}(u)$ denotes the inverse of the normal CDF, and $\gamma = \left(\varphi^{-1}(u_1), \dots, \varphi^{-1}(u_d)\right)'$.

From Eq. (IV.11), the d-dimensional Gaussian copula density can be calculated as,

$$c(u_1, \dots, u\ ; \rho) = \frac{1}{|\rho|^{\frac{1}{2}}} e^{-\frac{1}{2}\gamma'(\rho^{-1}-I)\gamma} \qquad (IV.13)$$

Figure (IV.3) shows the density and cumulative distribution of the Gaussian copula with p_{12} = 0.6.

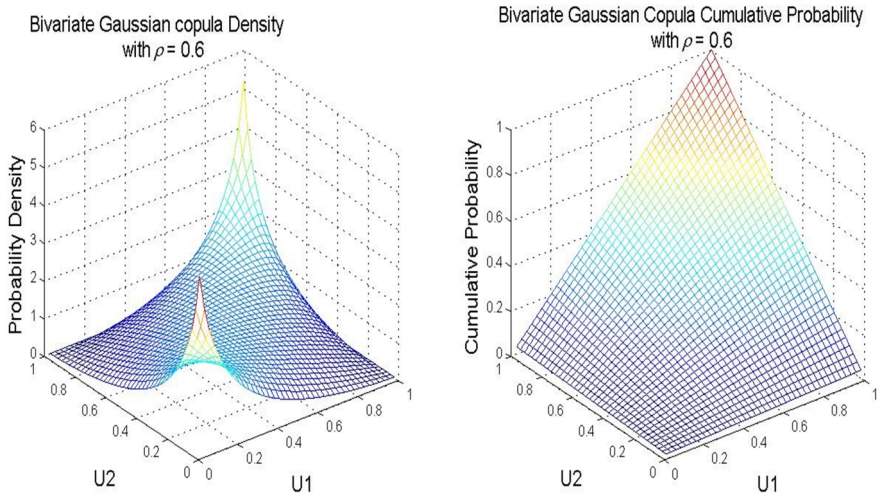


Figure IV.3. DENSITY AND CUMULATIVE DISTRIBUTION OF GAUSSIAN COPULA WITH P_{12} = 0.6

IV.5 MARGINAL DISTRIBUTION SELECTION AND COPULA ESTIMATION

Let $\{Y_i = (x_1^i, \dots, x_d^i)\}_{i=1}^N \forall\, Y \in R^d$, represent N training samples assumed random from unknown distribution F of $(X_1, \dots, X_d)$. When the marginal distributions are continuous, the copula density $c(u_1, \dots, u)$ is the unique multivariate density of $(U_1, \dots, U) = (F_1(X_1), \dots, F(X))$, as implied by Sklar's theorem.

Because copulas are not directly observable, a nonparametric copula density estimator has to be formed in two stages: First, obtain the observations for $(U_1, \dots, U)$, and then estimate the copula density on the basis of these observations. An appropriate marginal distribution in the copula function specification is crucial. In principle, the copula function models impose no restriction on the choice of marginal distribution. Thus, any parametric marginal distribution can be used, even an empirical one. Ultimately, the data can be the guide for such a selection.

In this contribution, we use a in the first step of the simple empirical CDFS (ECDFS) to estimate the marginal distribution of X_j features:

$$\widehat{u_j^i} = \widehat{F_j}(x_j^i = t) = \frac{1}{N}\sum_{i=1}^{N} x_j^i \le t \qquad (IV.14)$$

We also use a normal kernel density estimation for marginal densities, $\hat{f}$:

$$\hat{f}_j(x_j^i = t) = \frac{1}{N}\sum_{i=1}^{N} K(t - x_j^i), \qquad (IV.15)$$

Where K(•) is the normal kernel.

In the second step, we use Gaussian mixtures marginal distribution (GMMD). However, in order to do that, we need to model each marginal X_j by the GMM model first. The probability density function (pdf) drawn from the GMM is a weighted sum of M component densities, see (Eq. IV.16).

$$\hat{f}_j(x) = \sum_{i=1}^{M} p_i b_i(x) \qquad (IV.16)$$

where x is the variable from X_j, $b_i(x)$ are the component densities $(i = 1, \dots, M)$ shown in (Eq. IV.17), and p_i are the mixture weights.

$$b_i(x) = \left(\frac{1}{\delta_i\sqrt{2\pi}}\right)e^{-\frac{(x-\mu_i)^2}{2{\delta_i}^2}} \quad \text{(IV.17)}$$

In (Eq. *IV*.16), $\sum_{i=1}^{M} p_i = 1$,

where μ_i is the mean vector and δ_i is the standard deviation.

Each marginal distribution is represented by a mixture model and is referred to by the model $\Phi = \{p_i, \mu_i, \delta_i\}$. We can easily calculate the GMD function by integrating Eq. (*IV*.14) to obtain,

$$\widehat{F}_j(x) = \int_{-\infty}^{x}\sum_{i=1}^{M} p_i b_i(t)\, dt \quad (IV.18)$$

$$= \sum_{i=1}^{M} p_i \int_{-\infty}^{x} b_i(t)\, dt$$

$$= \sum_{i=1}^{M} p_i B_i(x)$$

$$\widehat{u_j^i} = \widehat{F_j}\left(x_j^i\right) = \sum_{i=1}^{M} p_i B_i\left(x_j^i\right) \quad (IV.19)$$

Here, B is the normal CDF with the parameters μ_i, δ_i. The GMM parameters are estimated from training data using the iterative expectation-maximization (EM) algorithm [79], and its mean-vectors are randomly initialized. In the second stage, we estimate the copula density based on the estimated observation $(\widehat{U}_1, \dots, \widehat{U}_d)$. A direct maximization of the log-likelihood function of the copula density with respect to the parameters can be employed. Using the exact maximum likelihood method, the estimated parameters for the Gaussian copula is the correlation matrix ρ that has the following closed form:

$$\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \gamma^i\ \gamma^{i'} \quad (IV.20)$$

Where,

$$\widehat{\gamma^i} = \left(\varphi^{-1}\left(\widehat{u_1^i}\right), \dots, \varphi^{-1}\left(\widehat{u_d^i}\right)\right)'$$

IV.6 THE PROBABILISTIC CLASSIFIER

The aim of this chapter is, in fact, to introduce the use of the copula function in speech recognition. According to Sklar's theorem, we can use a copula function in a probabilistic classifier, such as a Bayesian classifier. The Bayes theorem states the following:

$$P(K = k\backslash E = e) = \frac{P(E=e\backslash K=k)P(k=k)}{P(E=e)}, \qquad (IV.21)$$

Where $P(K = k\backslash E = e)$ is the posterior probability, $P(E = e\backslash K = k)$ is the function, $P(K = k)$is the prior probability, and $P(E = e)$ is the data probability.

The probabilistic classifier can be designed by comparing the posterior probability that an object (speech signal) belongs to class K given its attributes E. The object is then assigned to the class with the highest posterior probability. For practical reasons, the data probability $P(E)$ does not need to be evaluated in the comparison of the posterior probabilities. Further, the prior probability $P(K)$ can be substituted by a uniform distribution if we do not have an informative distribution. A copula function can be used to model the dependence structure in the likelihood function. In this case, Bayes' theorem can be written as:

$$P(K = k\backslash E = e) = \frac{C(F_1(e_1), \dots, F_d(e_d)) \prod_{i=1}^{d} f_1(e_i\backslash c)\, P(K = k)}{f(e_1, \dots, e_d)}, \qquad (IV.22)$$

Where $F_i$is the marginal distribution function that can be approximated by parametric marginal distributions or nonparametric marginal distributions, $\widehat{F}$, based on empirical cumulative distribution functions and $f_i$marginal densities of attributes that can be approximated by parametric or nonparametric marginal densities, $\hat{f}$ based on a histogram.

The function c is a d-dimensional copula density defined by Eq. (*IV*.13). As can be seen in Eq. (*IV*.22), each category determines a likelihood function. The general procedure used by the proposed speech recognizer is depicted in (Fig. *IV*.4).

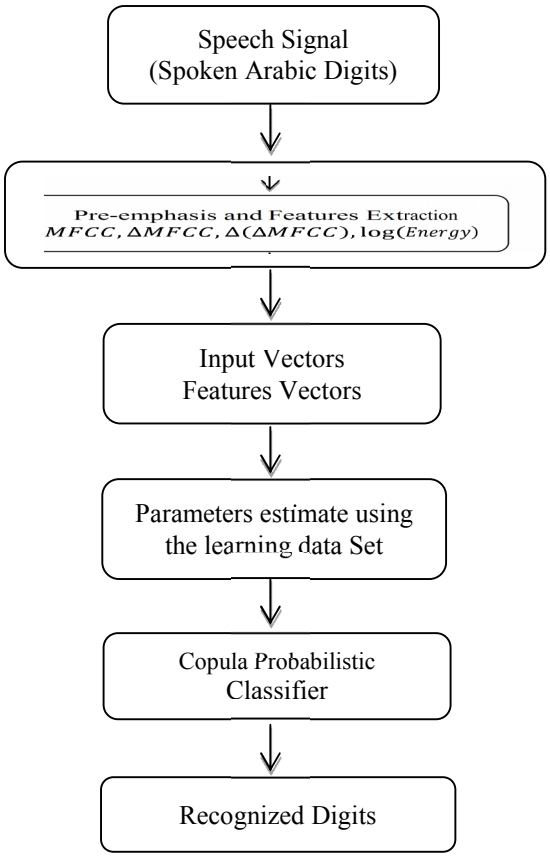


Figure IV.4. GENERAL PROCEDURE USED BY THE PROPOSED SPEECH RECOGNIZER

IV.7 EXPERIMENTAL EVALUATION

We conducted all our experiments using an Intel(R) Xeon(R) 3.33 GHz CPU computer. This section is for using the SAD database Chapter III(section III.5.1).The feature parameters are also based on Concatenation of MFCCs (ΔΔMFCCs) (section III.7.1).

- **PARAMETER ESTIMATION**

We compared our proposed method with the method most commonly used in the literature as HMMs. The optimal number of clusters (initial set of code words in the codebook, for DHMM), the number of states, and the number of mixtures for CHMM can be estimated using (CV), which proved its effectiveness with this database and HMMs (Hammami et al. 2011)[80].

An N-state constrained left-to-right sequential HMMs structure (Section III.5.2) was employed.

A) **ESTIMATED PARAMETERS FOR DHMMS**

The result of the clustering is a codebook of 64 DHMM (optimum size obtained using three-fold cross validation on the training set) and the best number of states is seven.

B) ESTIMATED PARAMETERS FOR CHMMS

In continuous density HMMs (CHMMs), the problem of parameter estimation is more serious and the parameters need to be judiciously selected to omit the divergence fate

IV.8 RESULTS AND DISCUSSIONS

IV.8.1 GAUSSIAN COPULA USING THE EMPIRICAL MARGINAL RESULTS VERSUS COPULA-BASED GAUSSIAN MIXTURES MARGINAL DISTRIBUTION

This experimental section covers two major issues: First, it covers the evaluation of the Gaussian copula based on the empirical marginal distributions (GCEM). Secondly, it covers the Gaussian mixtures marginal (GCGMM) distribution, where each experiment was repeated several times to improve the effect of some initialization parameters of GMM. The results are given in Tables IV.1 and IV.2, while the summary is given in Table IV.3.

Arabic Digits Classes	Success Rate (%) by GCEM
Syfr - 0	95.81
Wahid- 1	100.00
Ethnan- 2	92.63
Thalath'a- 3	98.43
Arb'a- 4	98.54
Khams'a- 5	98.18
Syt'a- 6	95.45
Sab'a-7	93.18
Thamany'a- 8	99.90
Tes'a- 9	96.82
Average	**96.89**

Table IV.1 RECOGNITION RESULTS FOR SPOKEN ARABIC DIGITS, CLASSIFICATION BY GCEM

Table *IV*.2 shows that the accuracy can virtually reach the empirical marginal distribution 100% for the words "Wahid" and "Thamany`a". An average of 96.89% is reached for all words.

Arabic Digits Classes	GCGMM Success Rate (%), Classification by Gaussian Densities				
	2	**3**	**5**	**7**	**9**
Syfr - 0	95.91	97.27	**98.64**	97.27	97.27
Wahid- 1	94.55	99.55	**100.00**	99.55	100.00
Ethnan- 2	93.64	93.64	**93.18**	93.18	93.54
Thalath'a- 3	88.18	94.09	**98.18**	99.09	99.09
Arb'a- 4	100.00	99.09	**98.18**	97.27	96.36
Khams'a- 5	97.27	97.27	**99.09**	99.09	99.55
Syt'a- 6	93.64	97.73	**98.64**	96.81	96.82
Sab'a-7	85.91	89.55	**96.36**	95.90	95.91
Thamany'a- 8	99.09	100.00	**100.00**	99.09	99.55
Tes'a- 9	95.45	97.27	**97.27**	97.73	96.72
Average	**94.36**	**96.55**	**97.95**	**97.50**	**97.48**

Table IV.2. RECOGNITION RESULTS SPOKEN FOR ARABIC DIGITS, CLASSIFICATION BY GCGMM

Table *IV*.3 shows the accuracy for five of nine Gaussian mixtures. The result given for the word "Wahid" is 100% and the word "Thamany`a", which has 99.99% for the empirical marginal distributions becomes 100% by GCGMM; the average accuracy is 97.95%.

Arabic Digits Classes	Success Rate (%) using Δ(ΔMFCC)	
	GCEM	GCGM
Syfr - 0	95.81	98.64
Wahid- 1	100.00	100.00
Ethnan- 2	92.63	93.18
Thalath'a- 3	98.43	98.18
Arb'a- 4	98.54	98.18
Khams'a- 5	98.18	99.09
Syt'a- 6	95.45	98.64
Sab'a-7	93.18	96.36
Thamany'a- 8	99.90	100.00
Tes'a- 9	96.82	97.27
Average	**96.89**	**97.95**

Table IV.3. Recognition results for spoken Arabic digits, classification by GCEM and GCGMM

In Table *IV*.4, it is remarkable that the difference between the two proposed methods is 1.06%. This means that the GCGMM gives a better result than the GCEM.

In the next section, we compare the results for our proposed copula based on GMM distribution to those for HMMs.

IV.8.2 RESULTS FOR COPULA BASED ON GMM DISTRIBUTION AND HMMS

The results of the algorithms developed (GCGMM and HMM) are summarized in Table IV.4. Comparing the results given by DHMM to that given by GCGMM, an improvement of 8.75% can be seen. Although the proposed model deals with speech data, experimental results indicate that it is very competitive with CHMM through the close overall recognition accuracies (98.41% versus 97.95%). However, the proposed model, while being quite easy to implement, gives performance recognition results that are competitive to the results given by HMMs.

Arabic Digits Classes	Success Rate (%) using $\Delta(\Delta^{MFCC})$		
	DHMM	CHMM	GCGM
Syfr - 0	88.27	95.05	98.64
Wahid- 1	95.90	99.55	100.00
Ethnan- 2	89.54	100.00	93.18
Thalath'a- 3	85.05	99.54	98.18
Arb'a- 4	92.27	98.18	98.18
Khams'a- 5	92.12	100.00	99.09
Syt'a- 6	96.81	100.00	98.64
Sab'a-7	89.54	95.45	96.36
Thamany'a- 8	86.81	96.82	100.00
Tes'a- 9	92.27	99.55	97.27
Average	**90.86**	**98.41**	**97.95**

Table IV.4. RECOGNITION RESULTS SPOKEN ARABIC DIGITS, CLASSIFICATION BY GCGMM AND HMMS

IV.8.3 COMPUTATIONAL COMPLEXITY

The resulting real learning time factors in Table IV.5 shows the impressive gain in computation time for our proposed method, GCGMM, for speech recognition. This method was capable of achieving speed-ups of up to three orders of magnitude in the experiment (174.30 s for the proposed technique versus 572.62 s obtained by GCEM), and speed-ups of up to 81 orders (174.30 s versus 14202.12 s obtained by HMM).

This positive contribution of the proposed model solves the problem of expanding the learning data in real time.

The GCGMM inference time is extremely fast and achieves a 323-fold speed improvement compared to GCEM. This advantage is very useful even on relatively high-end embedded devices.

	DHMM	CHMM	GCEM	GCGMM
LEARNING TIME (S)	170.01	**14202.12**	**572.62**	**174.30**
INFERENCE TIME (S)	29.75	**526.69**	**11936.03**	**34.73**
ACCURACIES (%)	**90.86**	**98.41**	**96.89**	**97.95**

Table IV.5.
LEARNING AND INFERENCE TIME BY GCEM AND GCGMM

Although the inference time of GCGMM inference is close to that given by DHMM, relatively high recognition accuracy with the minimum requirement of calculation power is produced. Thus, we have demonstrated that GCGMM reduces the learning calculation time from 14202.12 s given by HMM to 174.30 s, and the inference time from 526.69 s to 34.73 s with only a 0.48% drop in recognition accuracy.

IV.9 CONCLUSION

This chapter proposes a novel statistical approach based on the copula function, namely GCGMM, for problem speech recognition, in general, and in Arabic speech, in particular. This proposed approach enables a new level of flexibility in modelling the joint distribution of speech parameters. The results of the experimental evaluations conducted on a spoken Arabic digital database indicate that the proposed method has good classification accuracy compared to the HMMs. There is a significant gain in the learning and inference time, without a significant loss in the recognition rate. To fully exploit the generality of the copula-based models, possible directions for further development of this work include the development of other marginal distribution modelling methods for various copulas.

Part III

Application : Automatic Diagnosis Of Language Disorders For Children

Chapter V

Children's Speech Disorders /Impairment

V.1 INTRODUCTION

Speech and language impairment is a modern special education that appeared in the early 1960s. Communication through speaking and language is a very complicated process, and it is a development that naturally occurs in people without language impairment. Diagnosis and treatment are required as soon as an impairment is detected.

Speech impairment predominantly occurs in children as a result of sound output errors in pronouncing letters that are not properly formed. In this chapter, we address the theoretical status of childhood speech disorders by defining the natural status as well as presenting the most prevalent disorders and the treatment options. We address diagnosis methods according to a theoretical point of view based notably on books including *"Logopedics"* by Dr. Mustafa Fahmi [82] and *Speech and Language Disorders* by Dr. Faisal Al-Afif [83] as well as on field visits to speech disorder treatment centres and discussions with specialists.

This chapter facilitates a general idea of the collected speech database that would contribute to diagnosing impairment and its treatment without recourse to a specialist. The treatment of one of the most functional disorders includes promoting and supporting communication as well as the psyche of the child. These are not always available from a specialist and could be provided with high professionalism with interactive computer programs intended for children.

V.2 TYPES OF SPEECH AND LANGUAGE DISORDERS

Traditionally researchers divide it into two main categories [83]:

A) Flaws etiology for Organic factors

B) Flaws etiology for functional reasons

The first type of these flaws was caused by one of the following factors:

- A defect in the Articulation or hearing system or a deformation of any organ.
- Lack of the child's innate abilities (IQ)

The second type is not about any organic deformations but it is about the child's ability to pronounce properly which is influenced by factors other than organic.

Speech disorders can be divided regardless the flaw, but based on the outward appearance of the flaw, such as:

- Aphiasis
- Articulation
- Spastic Speech
- Voice disorders
- Fluency disorders
- Organic disorders

In this part, the research will mainly focus on inorganic speech disorders, which is usually treatable through adjusting the linguistic behavior as well as exercise: Faulty phonological processes. Disorders related to organic factors such as hearing disabilities, perceptual causes, or kinetic problems, oral-motor difficulties, Dysarthria, and Oral-Structural deviation are treated by clinical surgeries or psychologically, a filed far from the aim of this research which mainly ficus on the diagnostic part, phonological treatment and which can effectively contribute in the automatic identification on human voices.

- **Faulty phonological processes [83]**

Faulty phonological process is one of the functional factors. Stampe (1973) theory stands for the fact that children are producing language at an early age is not a random act but it happens according to specific rules. He also mentioned that a child is born with some organic, mental and learning disabilities that prevent him/her form producing a language the way adults do. This disability disappears with age evolution. Dunn (1982) calls these procedures as Phonological Process. It is what the child use while trying to pronounce adult's language with a simple way. It is mostly linked to Children's Speech disorders.

It is very important to mention that the children's produced language is judged by comparing it to adult's language, which is considered as the linguistic basic standard. Within the same idea, Shriberg and Kwiatkowski (1980) mentioned that the pronunciation process of sounds is based upon a set of rules basically 3 trials:

- The production of speech sounds.
- Rules of organizing and distributing the voice
- Rules of sound changes

Such disorder takes 4 known aspects, such as:

V.2.1 DISTORTION

Distortion includes voice pronunciation in a way that is closer to the regular voice, however, it is not quite the same but contains some errors. And often appears in the sound of certain letters such as(س /s/), (ش/sh/), where s is accompanied by a sound of a long whistle, or Pronounce a sound of /sh/ from the side of the mouth and tongue. Some uses the term lisping (lisp) to refer to this type of speech disorders.

Example: madrassa (مدرسة)- pronounce – Madrtha(مدرثة)

Dabet(ضابط) - pronounce – thabet (ذابط)

This may occur as a result of loss of teeth, the tongue is not put in its proper position during pronunciation, deviation of teeth position or loss of teeth on both sides of the lower jaw, which makes the air goes to both sides of the jaw and therefore the child cannot Pronounce the sounds like (س/s/), (ز/z/).

To illustrate this disorder you can put the tongue behind the front teeth -up- without being touched, and then try to pronounce some words to include sounds like s / z such as: *Sami*, *Sahran*, *Zahran*, *Saher*, *Zahir*, *Zayed*.

V.2.2 OMISSION

In this type of defects of pronunciation the child omits the sound of a (letter) of the votes, which is contained in the word and then pronounce a part of the word only ,OMISSION may include multiple letters in a steady manner , in this case the child speech is not understood at all , even for people who are familiar with listening to him , like parents and others, omission defects is common for young children more than older children , these defects tend to appear in the pronunciation of consonants , which is located at the end of the word more than consonants at the beginning of the word or in the middle.

V.2.3 SUBSTITUTION

Substitution defects are appears in pronunciation when issuing inappropriate sound for the letter instead of the desired one, for example, the child may replace letter (س/s/) with (و/sh/) or replace the letter (ر/r/) with a letter (و/w/) , Substitution defects are more common in young children more than older children, this type of speech disorder leads to

reduce the ability of others to understand the words of the child when it happens frequently.

V.2.4 ADDITION

This disorder includes adding extra sound (letter) to the word; the sound was heard like the one repeated like good morning, "*SSSbah...".*

Note, in many cases distortion and substitution could be considered as one type of defects for example, pronouncing of *Rajol* as *Gajol* can be considered a distortion as the letter (/r/) was changed to /gh/ and if it occurs in every word contains /r/ then it is also considered as substitution.

V.3 SPEECH DISORDERS' CHARACTERISTICS

- These disorders are common among young children in early childhood
- Different disorders related to letters varies from age to another
- Substitution is common among children more than any other disorders.
- If the child seven years old and still suffers from these disorders, he needs medical treatment.
- Speech disorders vary in degree, severity of the child, from an age group to another, and from a position to another…
- Whenever speech disorders remain with the child despite age advancement it becomes harder to cure.

V.4 SPEECH DISORDERS' DIAGNOSE

V.4.1 ARTICULATION SCREENING

Causes are determined in public primary schools and kindergartens where focus is on specific letters such as [(ل /l/), (ر/r/)], [(س/s/), (ش/sh/)], [(ز/z/), (ذ/th/)], [(ك/k/), (ق/ca/)], and other letters which are commonly the disordered.

At this stage, it is done without focus on their causes or how to treat it. The child is asked to pronounce some sentences and words .

V.4.2 HEARING AND LISTENING TESTING

Testing Sample:

Sawt, toot, boot, foot, koot, mawt	صوت ، توت ، بوت ، فوت ، قوت ، موت
R'aa = raha, baraza, sabara, rajul, Meriam, Kabeer, Saghir	ر = راح ، برز ، صبر ، رجل ، مريم ، كبير ، صغير
L'am = lamaha, milh, jamal,	ل = لمح ، ملح ، جمل ، جمل
Q'af = qala, maqalah, khalaq, falaq	ق = قال ، مقلة ، خلق ، فلق
K'af = kabeer, Akbar, Araka, kabsah, akala, malaka	ك = كبير ، أكبر ، أراك ، كبسة ، أكل ، ملك
Z'ay = Za'er, azeer, oroz,	ز = زائر ، أزير ، أرز
Dh'al = dhi'ib, dhanab, yadhoob, kadhib	ذ = ذئب ، ذنب ، يذوب ، كذب
Sin = sara,yasar, mares	س = سار ، يسار ، مارس
Shin = shajar, ashraq, yaaroosh	ش = شجر ، أشرق ، يرش
Kh'aa = kharoof, mokhtalef, tookh, kharaj, bokhar, kookh	خ = خروف ، مختلف ، طوخ ، خرج ، بخار ، كوخ
Djim = jamal, yajri, kharaj, jameel, yajreh, faraj	ج = جمل ، يجرى ، خرج ، جميل ، يجرح ، فرج
Th'aa = tha'er, atahr, irth	ث = ثأر ، آثار ، إرث
F'aa =fa'er, firash, yafooz, manoof, anef, faza, forn, yafer,	ف = فأر، فراش ، يفوز ، منوف ، أنف ، فاز ، فرن ، يفر
H'aa = haroof, ha'ar, ahmar, dahraj, jarah, marah, farah	ح = حرف ، حار ، أحمر ، دحرج ، جرح ، مرح ، فرح.

V.4.3 ARTICULATION SYSTEM SCREENING

V.4.4 ARTICULATION INVENTORY

It is a tool to help the specialist to identify the position of the sound error in the word (omission, substitution, distortion and addition.)

V.4.5 ASSIMILABILITY TESTING

It is an important step in the evaluation of speech disorders, including determining the child's ability to pronounce the disordered sound properly in front of a specialist, when repeatedly displayed (acoustically and visually, by touching). Snow and Mililsen (1954) found that showing and repeating the sound in different ways induces the child to pronounce it correctly.

V.4 .6 DEEP TESTING

It is ways explained in books of specialization where it is rare to find a child who is suffering from disorders Pronounce (whether functional or organic) and cannot Pronounce sounds correctly, even by a simple in-depth during the test.

V.5 SPEECH DISORDERS' TREATMENT

Here we will go through the treatment approved by a specialist in conversation, especially non-organic disorders (Voice disorders) which is the ultimate goal of this survey

where theoretical research will lead to the production of computerized programs to automate the process of diagnosis and treatment instead of specialists.

One of the main features of the treatment in education or re-education voice, to understand the child fully what the trouble sound, who is suffering from it, and what caused, and what must be done to alleviate this disorder, which does not need to confirm the need to be available to the child's defense enough to change the sound is appropriate, and to have the desire to modify some ingrained habits, without it be a therapeutic program prone to failure. The possible role that the clinical specialist of pronunciation can play in the treatment of children's sound disorder represents a small proportion, which requires a cooperation with the specialist and the desire to get to know (the new sound) and getting used to it.

It follows that, the child needs a great deal of encouragement and support from the specialist, the parents, teachers and colleagues throughout the period of the sounds training program. Although the therapeutic instructions vary according to clinicians and different cases, voice therapy usually includes four fundamental aspects deserve attention:

- If it is clear that the voice disorder associated with inappropriate use, it becomes one of the main manifestations of treatment to identify the sources of misuse, and to avoid these sources. Because the clinical specialist cannot fully rely on verbal reports provided by the children themselves, the good ideas and useful to the specialist is observing the child in a number of positions in order to determine the way that the child used to pronounce sounds, the reports submitted by the parents the teachers are necessary to identify the children's voice habits. Now we will discuss the types involved due to misuse and its impact on the speech with the child himself, and then start planning ways that can ease or avoid the problem. The child's understanding and cooperation is essential as the specialist cannot be present with the child at every moment and permanently alert to voice bad habits, and ask him to correct it.
- The second aspect of therapeutic program for speech disorders is relaxation training in this case the child is trained on how to get votes in a manner characterized by relaxation and smoothness, especially if the child speaks in a tensional manner, although the results with young children is not successful, the physical relaxation training may be necessary particularly in areas of the face, mouth and throat.

- The third aspect of the treatment includes vocal exercises and direct audio exercises to output different sounds. Now there are special exercises available to improve the voice layers, and exercises to raise the voice layer that the child used to it, and exercises to reduce this layer, and exercises to increase the flexibility of voice layer. When the child recognizes the new voice , he needs to al lot of practice to distinguish this sound and use it in different situations, which include speech . breathing exercises are always the main aspect of the fourth therapeutic program for speech disorders , and the aim of this kind of training is usually to get the child used to breath more effectively as the breathing for the purposes of speech does not need to provide the air more than the normal breathing, but the breathing for the purposes of speech requires adjustment and control , there are now many exercises to improve the rate of speech and adjust the process of breathing during speech .
- After avoiding sources of misuse of voice , and after adjusting the new sound , the specialist faces a difficult task to let the child continue in the correct use of voices educated , to get the child used to the new voice for all situations is the most difficult stages of the remedial action . Perhaps for this reason the continuous success of the treatment requires a team work, which includes clinical specialist, child, teacher, parents and others who are closely related to the child. This last aspect is the most important motivations to think about an automatic system for the treatment process.

V.6 FIELD DIAGNOSTICS AND THE OBSERVED DISORDERS' CASES AMONG ARABIC CHILDREN SPEAKERS.

V.6.1 VISITS TO SPEECH DISORDERS THERAPY CENTERS AND IDENTIFYING THE LETTER WITH MORE WIDESPREAD IMPAIRMENT

Since the aim is to develop a program for automatic diagnosis of speech disorders in children and rehabilitation, the most important step is to create recorded speech data base to be the basis for the development of automatic methods for speech recognition.In this regard we communicated with more than one center that specialized in treatment of speech disorders in children and we focused on international centers where they can collect a great diversity of nationalities of the children, and after the supervision crew of the dissertation has contacted the departments of education in Saudi Arabia where there are schools with different nationalities and residence foreigners , we have received

facilities for visiting schools and got speech (voice) samples. But before this we had to sit with consultants in the treatment of speech disorders to determine the most prevalent disorders among children, and the result is the following letters as arranged in (Table V.1) and the recommended used words to determine the type of disorder.

Letter	Words that are used to define the disorder type and place (beginning, middle, end)								
Ra	Raha	Baraza	Sabara	Rajul	Meriam	Kabeer	Saghir		
Lam	Lamaha	Milh	jamal						
qaf	Qala	Moqlah	Khalaq	falaq					
kaf	Kabeer	Akbar	Arak	Kabsah	Akala	malaka			
Za'a	Za'er	Azeer	oroz						
Dha'al	Dhi'eb	Dhanab	Yadhoob	kadheb					
Sin	Sara	Yasar	mares						
Shin	Shajar	Ashraq	yarosh						
Kh'aa	Kharoof	Mokhtalif	Tookh	Kharaj	Boukhar	koukh			
Djim	Jamal	Yajri	Kharaj	Jameel	Yajreh	faraj			
Tha'a	Tha'er	Athar	irth						
Fa'aa	Fa'er	Firash	Yafooz	Manoof	Anef	Faza	Forn	Yafer	Yafrom
Ha'a	Harf	Har	Ahmar	Dahraj	Jarah	Marah	faraj		

كلمات منصوح بها للوقوف على نوع الإضطراب وتحديد موضعه (أول ، وسط ، آخر)									الحرف
		صغير	كبير	مريم	رجل	صبر	برز	راح	ر
						جمل	ملح	لمح	ل
					فلق	خلق	مقلة	قال	ق
			ملك	أكل	كبسة	أراك	أكبر	كبير	ك
						أرز	أزير	زائر	ز
					كذب	يذوب	ذنب	ذئب	ذ
						مارس	يسار	سار	س
						يرش	أشرق	شجر	ش
			كوخ	بخار	خرج	طوخ	مختلف	خروف	خ
			فرج	يجرح	جميل	خرج	يجرى	جمل	ج
						إرث	آثار	ثأر	ث
يفرم	يفر	فرن	فاز	أنف	منوف	يفوز	فراش	فأر	ف
		فرج.	مرح	جرح	دحرج	أحمر	حار	حرف	ح

Table V.1. THE RECOMMENDED WORDS TO DETERMINE THE MOST DISORDERED LETTERS

In this research, the focus was on the letters that are more disordered in children which is the letter (ر/r/) and as it was difficult to register the disorders among children who suffer from speech disorder as there were not a lot of them , so used some healthy children to simulate the disorder.

To determine how to pronounce the disordered letter we distributed questionnaires to more than one center to find out how to pronounce the disordered letters and then simulate it. Selected disorders were approved by all specialists. Table V.2 shows the questionnaire form and the answer of the specialist.

Common letter with disorder	Ra'aa Letter								
Disorder type	Omission			Addition			distortion		
Disorder place	begin	Middle	end	Begin	Mid	end	Begin	Mid	End
Words used	*rajul*	*meriam*	*kabeer*	*rajul*	*meriam*	*kabeer*	*rajul*	*meriam*	*kabeer*
Specialist point of view about the pronunciation	*jul*	Rare disorder	*kabee*	*Rrrjul* *-Rrjul*	*-Merrrriam* *-Merriam* *-Mermeriem*	Rare disorder	*-Ghajul* *-Lajul* *-Ajul*	*-Meghiam* *-mayam*	*-Kabeegh* *-Kabeel* *-kabeeh*

الحرف الذي يكثر فيه الاضطراب	الراء (ر)								
نوع الاضطراب	حذف			إضافة (تأتأة)			إبدال أو تحريف		
مكان الاضطراب	في الأول	في الوسط	في الأخير	في الأول	في الوسط	في الأخير	في الأول	في الوسط	في الأخير
الكلمة المستخدمة للكشف عنه	(رجلٌ)	(مريمْ)	(كبيرْ)	(رجلٌ)	(مريمْ)	(كبيرْ)	(رجلٌ)	(مريمْ)	(كبيرْ)
كيفية نطق الاضطراب من واقع خبرة المختص	جلٌ	اضطراب نادر	كبي	رررجل رررجل	مرررريم مرريم مرمريم	اضطراب نادر	غجل لجل آجل	مغيم مايم	كبيغ كبيل كبيه

Table V.2 QUESTIONNAIRE FORM AND THE EXAMPLE OF ANSWER OF THE SPECIALIST

Table (V.2) a questionnaire to explain how to pronounce disordered letters in words according to letter position (in the beginning - in the middle – in the end) by one from specialists.

Common letter with disorder	Ra'aa Letter									
Disorder place	Beginning			Middle			End			
Disorder type	Correct pronun-ciation	Substituti-on / distortion	omission	addition	Correct pronoun-ciation	Substituti-on / distortion	addition	Correct pronoun-nciation	Substituti-on / distortion	ommision
Specialist point of view about the pronounciation	**Rajul**	**Ghajul lajul**	**jul**	**Rrrjul**	**Meriam**	**Maghyam malyam**	**merrrriam**	**kabeer**	**Kabeegh Kabeel**	**Kabee**

الحرف الذي يكثر فيه الاضطراب	الراء (ر)									
مكان الاضطراب	اضطرابات في الأول			اضطرابات في الوسط			اضطرابات في الأخير			
نوع الاضطراب	نطق صحيح	إبدال وتحريف	حذف	إضافة (تأتأة)	نطق صحيح	إبدال أو تحريف	إضافة (تأتأة)	نطق صحيح	إبدال أو تحريف	حذف
الكلمة المستخدمة للكشف عنه	**(رَجلٌ)**	**غجلٌ لجلٌ**	**جلٌ**	**رَرَرْجل**	**(مَريمْ)**	**مَغيمْ مَليمْ**	**مَرَرَرْيم**	**(كَبيرْ)**	**كَبيغْ كَبيلْ**	**كبي**

Table V.3 APPROVED DISORDERED PRONUNCIATION LETTERS BY SPECIALISTS.

Table (V.3) how to pronounce disordered letters approved by specialists.

- Shown in table (V.3), that the process of diagnosis of the type of disorder in the letter /r/ includes three stages:

- First make the child pronounce the word (***r**ajul*) in order to determine the type of disorder if it is in the beginning of words only.
- The second pronunciation word (*Ma**r**ium*) to determine the type of the disorder (addition, substitution, or omission), and if it is in the middle of the word
- Third let the child pronounce the word (*kabee**r***) to determine the type disorder (addition, substitution, or omission) and if it is in the end of the words ending with the letter R.

V.6.2 THE USE OF AUTOMATIC SPEECH RECOGNITION TECHNIQUES IN DIAGNOSIS AND TREATMENT

A) Diagnosis

The previous operation which is carried out by a specialist can be done automatically using computer programs where the machine is programmed with disorders then it can classify any disorder for a new child, by the rules of this particular voice.

B) Treatment

It has been indicated in this section that such inorganic disorders , especially in children who have the ability to assimilability (section V.4.5) are treated mainly with respect to the psychology of the child and its ability to correct the voice by himself and under the guidance of the specialist, encouragement , instigation and specialist follow-up throughout the treatment period , after that the child has to continue to exercise even after the improvement which is very difficult so that , the interactive software dedicated to the child can achieve this goal easily and make the child reacts in a manner faster and easier as well as saving time , costs and make the educational and rehabilitation the process of the child available to the child in any place away from schools or rehabilitation centers . Therapeutic program will depend mainly on comparing the voice of the child to a model ready-to-pronounce the correct letters then interactive educational user interfaces will try to show him the correct pronunciation and encourage him using the awards moral or in the form of computerized games that urge him to pronounce the correct

pronunciation , the program will compare the voice of the child to the correct model and urging for a better performance to continue to play and interact with the program.

V.7 CONCLUSION

This segment of the study provides information about childhood speech disorders including the types, treatments and field studies regarding native Arabic children. The study investigates computer use in the treatment and diagnostics of speech disorders. We record the voices of children and start building a vocal-speech- database that is the basis of any computer program aimed at treatment and diagnosis. This work is presented in the following chapter of the thesis.

Chapter VI

Data Base Development and Automatic Diagnosis

VI. INTRODUCTION

The work performed in the previous chapter led us to focus on the most prevalent speech disorder among Arabic children, which is the mis pronunction of the letter /r/. This chapter presents the registration process and details the construction of the database (including the recording conditions, preliminary processing, and speech file databases). This chapter presents the experimental results and discussions for the automatic diagnosis of the /r/ letter disorder by using a different classifier in the developed database.

VI.2 (RA, /R/)-LETTER DISORDER DIAGNOSIS DATA SET

VI.2.1 DATA BASES DESCRIPTION

This section describes /r/-letter disorder diagnosis (/r/-LDD) speech data set. The database consists of three separate data sets. The first one is for /r/-disorder types on the beginning of the words (/r/-DTBW). The second data set for /r/-disorder types on the middle of the words c(/r/-DTMW). The last one is designed for /r/-discord types at the end of the word (/r/-DTEW) . The data sets will be used for the automatic disorder diagnosis on first step and the ratability on second step. The data sets hierarchy is described in FigureVI.1.

Each class consists of 300 utterances spoken by 60 speakers (30 male, 30 female). Their age varies from 4 to 12 years.

- /r/-Letter types disorder on the beginning of words (/r/-LTDLW) is composed of 1500 utterances (5 classes x 300(5 repetitions x 60 speakers) utterances) .
- /r/-Letter types disorder on the middle of words (/r/-LTDLW) is composed of 1200 utterances (4 classes x 300(5 repetitions x 60 speakers) utterances) .
- /r/-Letter types disorder at the end of words (/r/-LTDLW) is composed of 1200 utterances (4 classes x 300(5 repetitions x 60 speakers) utterances) .

Table VI.1 shows the recorded words for each class.

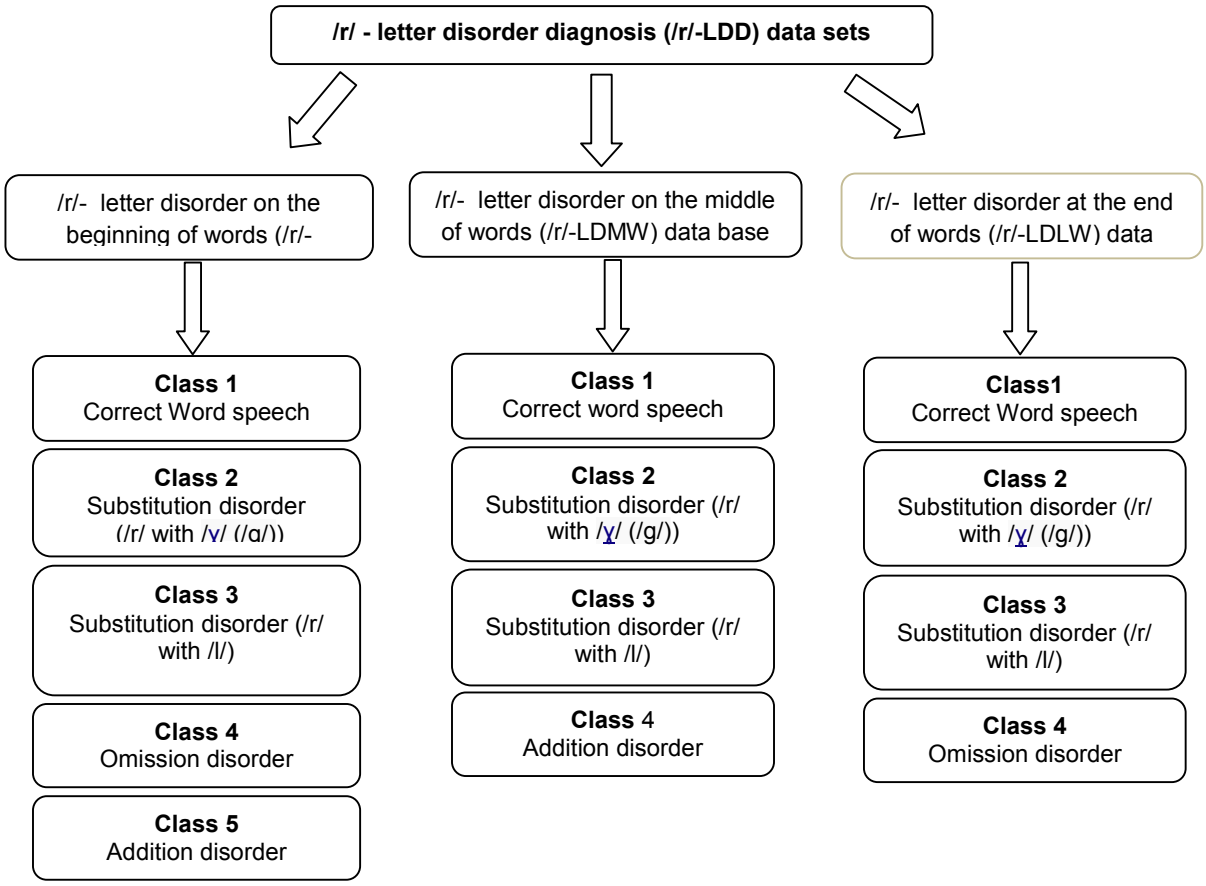


Figure VI.1. HIERARCHY OF /R/-LDD

VI.2.1 RECORDING CONDITIONS

The condition used in recording was not enough to find sufficient number of children. For this reason we have recorded the utterances in non-sever condition, and in real world noisy environment. The databases may be a challenge for researchers to develop robust ASR.

- The speech was recorded in the natural noisy environment for children (schools, home, street, yard games … etc).
- The microphone was placed on a stand about 5 cm to 20 cm, slightly off-axis from the speaker.
- The recording equipment used was different and recording gain was not constant across all recordings.
- The utterances were recorded at one sitting for each speaker.

- Speakers are Arabic native speakers covering most Arabic countries (Algeria, Saudi Arabia, Egypt, Syria , Sudan, Jordan , morocco ,Tunisia, Iraq ...etc)
- Each child was asked to read the word before the recording started by using a developed program for such task, (see Appendix B).

VI.2.2 PRELIMINARY PROCESSING

The speech data dose is not subject to any preprocessing except to mark the beginning and the ending of the word as well as so to omit silent clips.

Figure VI.2 shows an example in omitting the silent sound in the pronunciation of the word "*Merraryam*" *(مرريم)* from the 4th class of (/r/-LTDMW) data set.

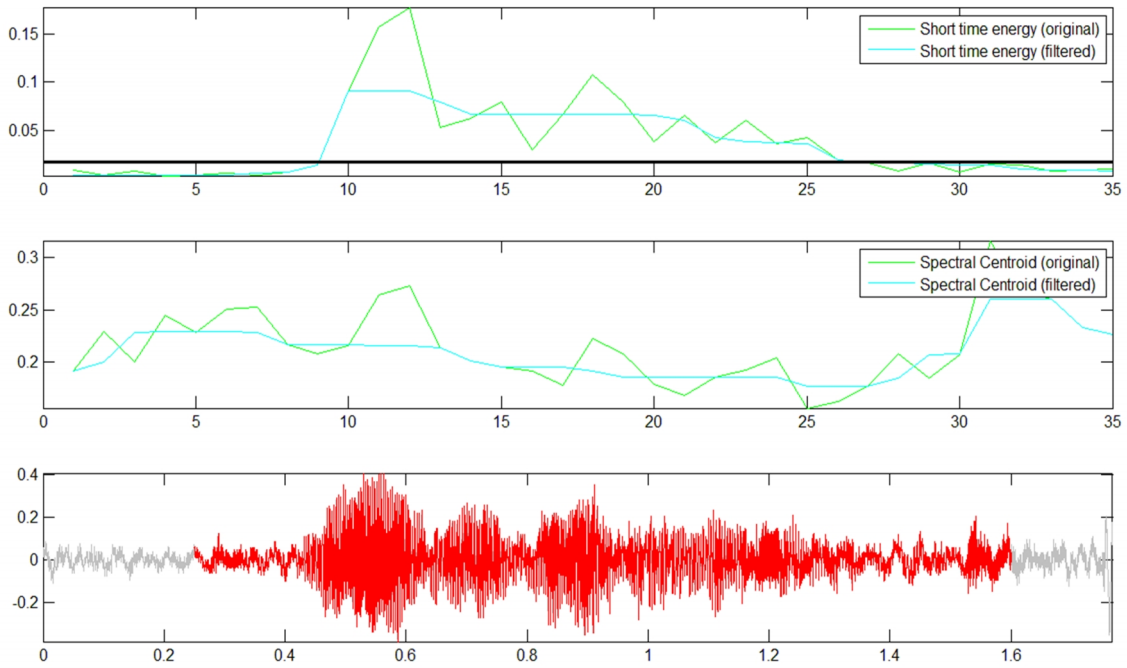


Figure VI.2. SILENT SOUND OMISSION IN THE PRONUNCIATION OF THE WORD "MERRARYAM" (مرريم)

/r/ - letter disorder diagnosis (/r/-LDD) data base														
Data Set	/r/-LTDBW					/r/-LTDMW				/r/-LTDEW				
Class	Correct	Substitution disorder (/r/ with /ɣ/ (/g/))	Substitution disorder (/r/ with /l/)	omission	Addition	Correct	Substitution disorder (/r/ with /ɣ/ (/g/))	Substitution disorder (/r/ with /l/)	Addition	Correct	Substitution disorder disorder (/r/ with /ɣ/ (/g/))	Substitution disorder (/r/ with /l/)	Omission
Word to speech	ريجيون [Reajolon]	غيجيون [Gheajolon]	لجون [Lagolon]	جون [Jolon]	ريريجيون Rearearejolon	(مريم) Mearyam	مغيم Meaghyam	مليم [Mealyam]	ميريريم Mearereryam	كيبير Keabeer	كيبغ [Keabeegh]	كيبيل [Keabeel]	كيبيا [Keabea]

TABLE.VI.1 WORDS RECORDED FOR EACH CLASS FROM DATA SETS

VI.2.4 SPEECH FILE DATABASE

VI.2.4.1 FILE NAMES

Originally, there are 30 male and 30 female speakers, 8 male and 8 female for testing (The male speakers go first (loc1_. to loc8_.), female speakers go next (loc9_. to loc16_.)). 22 females and 22 males for training (The male speakers go first (loc1 to loc22) and female speakers go next (loc23_. to loc44_.)).

Each utterance appears as a separate file. For instance the file with name "**loc7_5.wav**" is utterance **5** from the *Locuteur* (speaker) number **7**.

VI.2.4.2 FILE FORMAT AND SAMPLING RATES

The files format is Waveform Audio File (WAVE) .

- **Description**
  File format for audio. Wrapper file format that can incorporate an audio bitstream with other data chunks. One common bitstream encoding is LPCM (Linear Pulse Code Modulation). Preservation reformatting projects generally use the one of the Broadcast WAVE variants,WAVE_BWF_1 or WAVE_BWF_2, both standards of the European Broadcast Union[1].
- **Sampling Rates (FS) and the number of bits per sample (NBITS)**
  The original sampled data is available as files with a 11025 Hz sampling rate and 8 bits.

VI.3 AUTOMATIC /R/-LETTER DISORDER DIAGNOSIS FOR ARABIC LANGUAGE

VI.3.1 EXPERIMENTAL DESIGN

This section is for using the database developed in Chapter VI, section VI.2.The feature parameters are based on MFCCs, which can simulate characteristics of the channel spectrum and the human ear's hearing [16]. In the signal analysis phase, the input speech signal is transformed into feature vectors containing spectral and/or temporal information using Mel frequency spectral coefficients (MFCCs) /r/-LDD data set(sec II.2.) .

[1] http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml

VI.3.2 DIAGNOSIS ALGORITHM

The identification methodology for /r/-disorders goes through 3 stages:
The child is first asked to pronounce the following words in its correct form for each database:

- (*Rajolonn*) for database /r/-LTDBW
- *(Mariam)* for database /r/-LTDMW
- *(Kabee-ronn)* for database /r/-LTDEW

Then the automatic speech recognition methods are applied to identify to which type of disorders belongs the speech which has been input in each of the three speech databases.

In the experimental results we use the part specified for testing (on/from /r/-LDD) to determine the success rate and the correct classification.

VI.3.3 EXPERIMENTS EVALUATION FOR AUTOMATIC DIAGNOSIS FOR /R/- DISORDERS

The most common statistical methods in the field of automatic speech recognition have been used (HMM and GMM), in addition to the method proposed in this thesis which depends on Copula (chapter III).

VI.3.3.1 RESULTS IN THE BEGINNING OF THE WORD

To diagnose the /r/-disorder in the beginning of the word we use the speech database /r/-LTDBW. For the HMM method, it has been applied using a different variety of parameters, and Table VI.2 shows the results obtained, while Table VI.3 shows the results obtained using GMM & GCGMM for different values of the Gaussian mixture.

	HMMs Success Rate [%]. Classification by Number of selected states			
	3	5	7	9
2	63	58.25	70.5	62.25
3	66.5	58	66.5	70.5
5	58.75	64.5	**71.15**	71
7	66.25	70.5	64	64.5
11	63.5	68	66.25	60
13	61.25	71	70	55.75

Table VI.2
RESULTS OF /R/-LETTER DIAGNOSIS IN THE BEGINNING OF THE WORDS, CLASSIFICATION BY HMMS

Number of mixtures	GMM Success Rate [%].	GCGMM Success Rate [%].
2	66.75	55.25
3	67.25	53
4	**71.75**	59.75
5	69	64.25
6	67.75	64.5
7	67.5	67.5
8	67.25	68
9	67	67.5
10	66.75	68.5
11	/	67.25
12	/	**71.5**
13	/	71.25
14	/	69.5
15	/	68.5

TABLE VI.3
RESULTS OF /R/-LETTER DIAGNOSIS IN THE BEGINNING OF THE WORDS, CLASSIFICATION BY GMM AND GCGMM

By analyzing the tables it is clear that the Gaussian mixtures which gave the best results is 5 with (7 states) for HMM, and the Gaussian mixture which gave the best result

using the GMM model is 4, while the best result for GCGMM was obtained using 12 mixed Gaussian.

Table VI.4 shows a comparison between the 3 methods using the best parameters, where the success rate of /r/-disorders in the beginning of the word was close with a slight edge in GCGMM (71.75% vs. 71.15% in GMM and 71.15% in HMM).

/r/-LTDFW	Success Rate [%]. Classification by		
	CHMM	GMM	GCGMM
Class 1: Correct Speech	58.75	66.25	76.25
Class 2: Sub-(/r/ with /ɣ/ (/g/))	73.75	68.75	60
Class 3: Sub- (/r/ with /l/)	62.25	70	53.75
Class 4: Omission disorder	83.25	77.5	96.25
Class 5: Addition disorder	77.75	75	72.5
Average	**71.15**	**71.5**	**71.75**

Table VI.4
CLASSES RESULTS OF /R/-LETTER DIAGNOSIS ON BEGINNING OF THE WORDS, CLASSIFICATION BY HMMS, GMM AND GCGMM

VI.3.3.2 RESULTS IN THE MIDDLE OF THE WORD

To diagnose the /r/-disorder in the middle of the word we use the speech database /r/-LTDMW. The same experiments which were carried out before (VI.3.3.1) have been used, and the results are shown in Tables VI.5, VI.6, and VI.7.

	HMMs Success Rate [%]. Classification by Number of selected states			
	3	5	7	9
2	59.375	75.875	72.5	75.3125
3	74.375	63.75	70.625	65
5	**76.56**	68.75	55	70
7	66.5625	66.25	69.375	72.1875
11	71.5625	69.6875	69.375	70.625
13	69.375	72.5	75.5625	66.875

Table VI.5
RESULTS OF /R/-LETTER DIAGNOSIS IN THE MIDDLE OF THE WORDS, CLASSIFICATION BY HMMS

Number of mixtures	GMM Success Rate [%].	GCGMM Success Rate [%].
2	51.5625	71.5625
3	58.75	72.5
4	58.4375	73.75
5	60	**77.4375**
6	61.5625	70.625
7	62.8125	70
8	70.3125	70.9375
9	71.25	68.4375
10	**72.56**	66.5625
11	68.125	/
12	66.87	/

Table VI.6
RESULTS OF /R/-LETTER DIAGNOSIS IN THE MIDDLE OF THE WORDS, CLASSIFICATION BY GMM AND GCGMM

In the HMM model, the number of Gaussian which gives the best results is 3 with (5 states), for the GMM the Gaussian mixture which gave the best result is 10 while the best result obtained for GCGMM was obtained using 5 mixed Gaussian.

Table VI.7 shows a comparison between the 3 methods where the success rate in identifying the /r/-disorder in the middle of the word was weak in GMM model when compared with HMM and GCGMM respectively (72.56% vs. 76.56% in HMM and 77.18% in GCGMM)

It can be noted however that the Markov model has exceeded -in diagnosis- a success rate of 90% (correct words (no disorders in the middle of the word)) with a success rate of 90%, and a 91.25% success rate in the diagnosis of the /r/-disorder in the middle of the word).

/r/-LTDMW	Success Rate [%]. Classification by		
	HMM	GMM	GCGMM
Class 1 Correct Speech	90	70.75	73.75
Class 2 Sub-(/r/ with /ɣ/ (/g/))	63.75	68.75	86.25
Class 3 Sub- (/r/ with /l/)	61.25	78.75	71.25
Class 4 Addition disorder	91.25	72	77.5
Average	**76.56**	**72.56**	**77.43**

Table VI.7
CLASSES RESULTS OF /R/-LETTER DIAGNOSIS IN THE MIDDLE OF THE WORDS, CLASSIFICATION BY HMMS,GMM AND GCGMM

VI.3.3.3 RESULTS AT THE END OF THE WORD

To diagnose the /r/-disorder at the end of the word we use the speech database /r/-LTDLW. We repeat the same experiments in sec (VI.3.3.1) and sec (VI.3.3.2), and the results are shown in Tables VI.8, VI.9, and VI.10.

	HMMs Success Rate [%]. Classification by Number of selected states			
	3	5	7	9
2	60.625	57.1875	59.0625	57.5
3	**63.75**	57.8125	60	59.6875
5	58.125	58.125	58.75	57.1875
7	61.5625	60.9375	55.9375	47.5
11	59.0625	60.9375	59.0625	57.1875
13	55.625	58.4375	56.5625	53.4375

Table VI.8
RESULTS OF /R/-LETTER DIAGNOSIS ON THE END OF THE WORDS, CLASSIFICATION BY HMMS

Number of mixtures	GMM Success Rate [%].	GCGMM Success Rate [%].
2	45	63.43
3	51.875	69.6875
4	57.1875	**74.06**
5	53.75	70
6	52.8125	69.0625
7	50.9375	69.6875
8	51.25	67.1875
9	52.1875	67.8125
10	55.9375	66.25
11	54.0625	65
12	55.3125	64.375
13	**59.31**	66.875
14	56.875	65.625
15	57.5	64.375

TABLE VI.9
RESULTS OF /R/-LETTER DIAGNOSIS ON THE END OF THE WORDS, CLASSIFICATION BY GMM AND GCGMM

In the HMM model, the Gaussian mixtures which gave the best results is 3 with (3 states), for the GMM the number of Gaussian mixture which gave the best result is 13 while the best result obtained for GCGMM was obtained using 4 mixed Gaussian.

Table VI.10 shows a comparison between the 3 methods, where the adventage of GCGMM model was clear and better than the rest of the models, where its success rate in diagnosing the /r/-disorder at the end of the word was 74.06% vs. 63.75% in the HMM model, and a weak diagnosis success rate in the GMM model (59.31%).

We can note that the results obtained using the GCGMM model in diagnosing the correct pronunciation of the word (no disorder at the end of the word) with a success rate of 82.5%, and a success rate of 90.00% in diagnosing the /r/-disorder at the end of the word.

/r/-LTDMW	Success Rate [%]. Classification by		
	HMM	GMM	GCGMM
Class 1: Correct Speech	51.25	60	82.5
Class 2: Sub-(/r/ with /ɣ/ (/g/))	65	49.75	72.5
Class 3: Sub- (/r/ with /l/)	60	63.75	51.25
Class 4: Omission disorder	78.75	63.75	90
Average	**63.75**	**59.3125**	**74.0625**

TABLE VI.10
CLASSES RESULTS OF /R/-LETTER DIAGNOSIS ON THE END OF THE WORDS, CLASSIFICATION BY HMMS,GMM AND GCGMM

VI.3.4 DISCUSSIONS

Observing the previous results in the three models, we could generally conclude that the weakest results were obtained when classifying and diagnosing the disorder characterised by switching the letter /r/ with the letter /l/, especially when the disorder is in the middle of or at the end of the word because /r/ and /l/ share the same points of articulation or "output". The shared points of articulation condition is called "like letter" in the field of phonetics, meaning that the two letters are very similar, which ensures that interference between them is very likely to occur. The process of phonetically distinguishing them is difficult (Figure VI.3)[1]. Details concerning alphabet phonetics and their characteristics are presented in Appendix A.

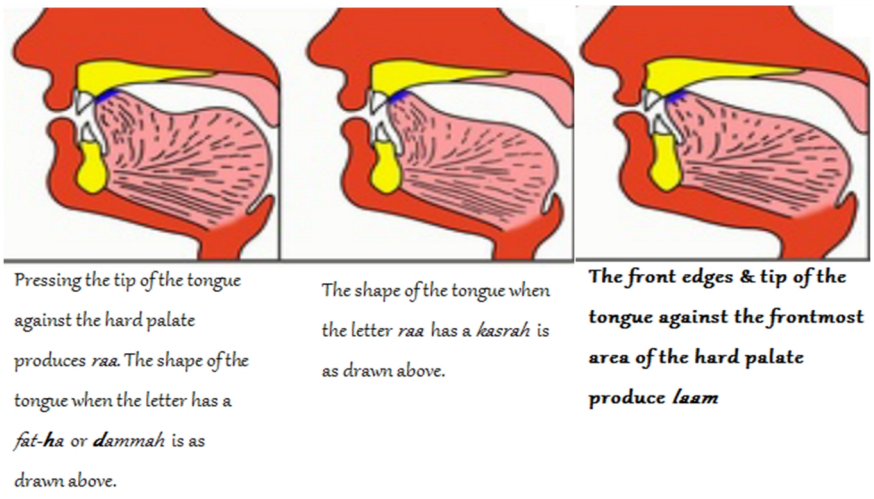


Figure VI.3 THE POINTS OF ARTICULATION OF THE LETTERS /R/ AND /L/

In addition to the assimilation between the two letters, they share five characteristics and differ in only one characteristic, which is the repetition characteristic of /r/ (Appendix A). A weakness in the classification results in the correct pronunciation for /r/ in the beginning, middle, or at the end of the word, which is most likely caused by the similarity between the correct model of pronunciation, and the pronunciation associated with the disorder (/l/ instead of /r/).

[1] http://hidayahacademy.blogspot.com

The best overall results were obtained in the omission disorders models that is most likely caused by the disappearance of the letter /r/ completely from the pronounced word, which ensures that the corresponding model for it interferes less with the other models.

Generally, the results require improvement, particularly because t/r/ is a weak letter in the context of articulatory classification (Appendix A). In addition, /r/ is the letter with the most characteristics (7 characteristics), which causes the correct pronunciation to be difficult, requiring more effort for automatic recognition including focusing on phoneme recognition instead of word recognition.

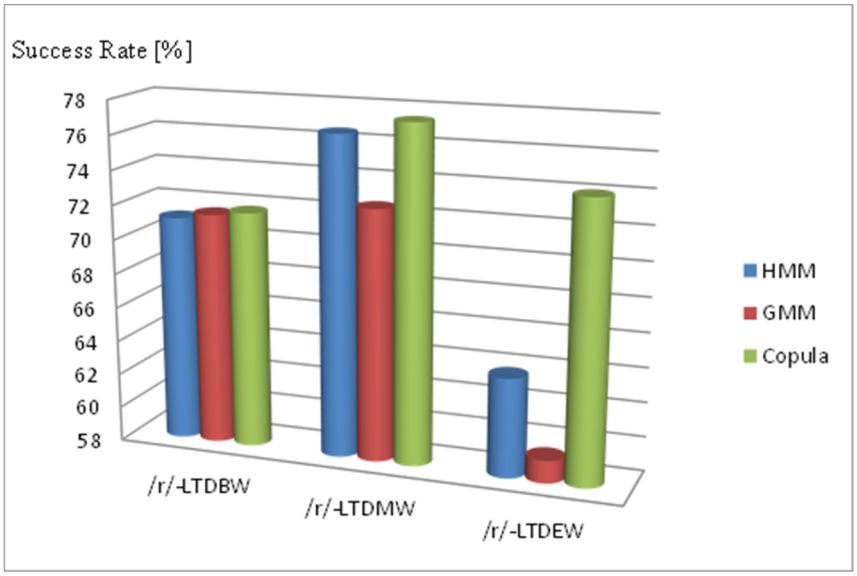


Figure VI.4
RESULTS OF /R/-LETTER DIAGNOSIS, CLASSIFICATION BY GMM AND GCGMM

Figure VI.4 shows the difference between the three models in the diagnosis of the letter /r/.

Figure VI.4 shows the excellence of the copula-dependent model over the other models in the three stages of diagnosis, and its excellence appears specifically in the diagnosis of the /r/-disorder at the end and the middle of the word; in the beginning of the word, the models have similar results, which shows the effectiveness of this model in the field of automatic speech recognition.

VI.5 CONCLUSION

Through the research results in chapter V, we demonstrated the database of the /r/-letter disorder diagnosis. The database consists of three data sets. A different classifier, using the (/r/-LDD) data base, was applied, and the results, despite their modesty, encourage serious research to improve the results through the application of other recognition methods, other methods to extract better parameters, or the application of recognition methods dependent on the segmentation of words into letter/phonemes instead of consideration of a word as one entire signal. Experiments showed the efficiency and excellence of the suggested GCGMM model with data containing noises, and it promises results and competitiveness with the other known methods.

Chapter VII

Thesis Conclusion

VII. THESIS CONCLUSION

This chapter is a summary of the work and conclusions presented in this dissertation, followed by future proposals. In this work, we have created a flexible, versatile, and compact model for automatic speech recognition. We presented a new database for an application related to an automated diagnosis for speech disorders among children, using the model we developed and comparing it to other models.

VII.1 SUMMARY

First, this thesis attempts to show the importance of ASR, providing a comprehensive survey of the research and yearly progress of speech recognition development and a discussion of the existing approach to ASR.

Second, we propose that our contribution is based on statistical tools that are the classical approach for speech recognition. These tools are the major contributor to the field and have been used in many recent applications for pattern recognition. They have not yet had an effect on automatic speech recognition.

The contributions of this thesis are manifold and are summarised below.

A) Tree Distributions Approximation (TDA) Model for Robust Discrete Speech Recognition

This new method which investigates the capability of graphical models based on tree distributions that are widely used in many optimization areas. A novel spanning tree structure that utilizes the temporal nature of speech signal is proposed. The proposed tree structure significantly reduces complexity in so far that can reflect simply a few essential relationships rather than all possible structures of trees. The application of this model is illustrated with different isolated word databases. Experimentally it has been shown that, the proposed approaches compared to the conventional discrete hidden Markov model (DHMM) yield reduced error rates of 2.54% -12% and improve recognition speed minimum 3-fold. In addition, an impressive gain in learning time is observed. The overall recognition accuracy was 93.09% - 95.34%, thereby confirming the effectiveness of the proposed methods.

B) Gaussian copula with the Gaussian mixtures marginal (GCGMM) for speech Recognition

The proposed classifier depends on one of the advanced statistical tool called (Copula), the approach combines finite Gaussian mixture modeling for marginal distribution and the Gaussian copula. First, we empirically investigate the applicability of the Gaussian copula with a nonparametric kernel density estimator (KDE) for marginal distribution. Next, we estimate the marginal distribution with Gaussian mixtures model. Our experimental results using a benchmark Arabic speech database indicate that the Gaussian copula with the Gaussian mixtures marginal (GCGMM) distribution is a sufficiently accurate model. Compared to the conventional hidden Markov model (HMM), our proposed approaches yield convergent recognition accuracy and improve recognition speed a minimum of 15-fold. We also observed an impressive gain in learning time. The overall recognition accuracy is 97.95%, thereby confirming the effectiveness of our proposed methods. The results demonstrate that our proposed methods improve on conventional methods and have excellent performance.

A case study was conducted, and a real application using automatic speech recognition was used in real life, where –with the aid of speech specialists- a speech database was built to diagnose speech disorders in children. The results for the automated diagnosis of speech disorders in children have been encouraging, given the nature of the database recorded with full variation in the following aspects including the environmental conditions, speaker variability and noise. Experiments have shown the efficiency and excellence of the suggested GCGMM model, and the correct automatic diagnoses exceeded 90% for some speech disorder types.

VII.2 FUTURE RESEARCH

Future research could focus on two main axes, as follows:

The first theory, in trending towards future developments that adopt more than one tree in the TDA classifier design, would lead to more robust classification results and exploit the generality of copula-based models using other margin distribution modelling methods adapted for various copulas. The copula classifier could be extended to include a

combined author classifier as a hidden dynamic by leveraging a state-space framework. One possible future research direction is to introduce a hybrid classifier by using a copula.

The second axis interest for applications by developing future systems and software for Arabic language as the automatic rehabilitee of speech disorder for children and invest research for the manufacture of devices that could serve children.

Vivo (Voice-In/Voice-Out) technology most likely forms the future of device control and future applications. The ability to manipulate and operate devices by voice is attracting increasing interest, especially for people with disabilities. One prospective application of the technology presented in our work concerns people with low or no vision who would appreciate listening to the Quran on smart devices [84].

Appendix A

Points of Articulation of Arabic Letters and its Characteristics

A.1 INTRODUCTION

Although Arabic language underwent significant developments, the pronunciation of the letters (Points of articulation and Characteristics) is theoretically still the same along 14 centuries thanks to the Holy Quran. Muslims were keen on establishing the basis of the language and imparting how to pronounce letters very accurately within what is called *Tajweed*[1], to impart the way of how prophet Muhammad (On Him are the blessings and the peace of Allah) was reciting the holy Quran without any misrepresentation. In this appendix we will focus on points of articulation, Characteristics of Arabic language and Illustrations as well, which is one of *Tajweed*'s sections. We look forward that this appendix is a comprehensive and important reference to whoever desires to work on Arabic Automatic Speech Recognition.

A.2 THE POINTS OF ARTICULATION OF ARABIC LETTERS (*Makhaarij*)

The points of articulation of Arabic letters is known as *Makhraj*, singular and *Makhaarij*, plural. There are five main points of articulation which coincide with the organs of speech from where the letters emanate from, (see Fig A.1). These five parts are further divided into 17 subsections of actual point of articulation for each letter.

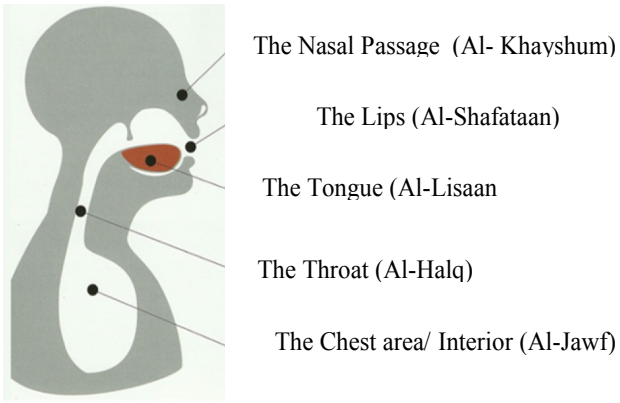


Figure A.1. THE MAIN POINTS OF ARTICULATION OF ARABIC LETTERS

[1] Literally, *Tajweed* means the recitation of the holy Quran as it was revealed to the prophet Muhammad (On Him are the blessings and the peace of Allah). Technically, it means excellence and precision in reciting the Quran giving each letter its right in pronunciation and qualities, as well as proper lengthening, assimilation. Knowledge of *Tajweed* is divided into 3 parts:

i. *Makhaarij*: The point of articulations of letters, which means the exact place the letters emanate from.
ii. *Siffat*: Characteristics of individual letters.
iii. *Ahkaam*: Rules applied to the letters and recitation in *Tajweed.*

Knowing the point of articulation of each letter in Arabic helps with correct pronunciation, even if one is struggling with pronunciation as it is common with non-Arabic speakers, even some Arabs. The more one practices the easier it becomes with a qualified teacher versed in *Tajweed* or the rehabilitation of children who suffer from disorders speech, as in the application of thesis, who can easily pick up and correct mistakes. The design of any software for the automatic correction speech disorder need to take on consideration the importance of the points of articulation of Arabic letters and its individual characteristics.

Table A.I shows the five parts of speech with number of specific *Makhaarij* for the 28 Arabic letters. Table A.II shows the 17 Specific *Makhraj* and the letters from each with the correspondent points of articulation shown on figures[1].

	Part of speech General *Makhraj*	**Meaning**	**Specific *Makhraj***	**Letters**
1	*Al-Khayshum*	The nasal passage	1	*Ghunnah*[2] sound
2	*Ash-Shafataan*	The two lips	2	4
3	*Al-Lisaan*	The tongue	10	18
4	*Al-Halq*	The throat	3	6
5	*Al-Jawf*	The chest area	1	3 (Madd[3] = Elongation)
	Total Specific *Makhraj*		17	

Table A.I . THE FIVE PARTS OF SPEECH WITH NUMBER OF SPECIFIC MAKHAARIJ

---

[1] Ghothani home for Quranic studies [http://www.gwthani.com/gwthani/index.php?]

[2] ***Ghunnah***: making sound from the nose by blocking the flow of air from the mouth with the tongue [ن (Nūn)] or lips [م (Mīm)].

[3] ***Harakat* (short vowel marks)**

The ḥarakāt, which literally means 'motions', are the short vowel marks are markers for both vowels and consonants.

Fatḥah(◌َ) The *fatḥah* is a small diagonal line placed above a letter, and represents a short /a/. The word *fatḥah* itself means opening, and refers to the opening of the mouth when producing /a ./

Kasrah(◌ِ) A similar diagonal line below a letter is called a *kasrah* ('breaking') and designates a short /i/. Example ⟨دِ⟩ : /di./

Ḍammah(◌ُ) **:** The *ḍammah* is a small curl-like diacritic placed above a letter to represent a short /u/. Example/ ⟨دُ⟩ :du./

Madd(◌ٓ) : The maddah is a tilde-like diacritic which can appear only on top of an *alif* and indicates a glottal stop /ʔ/ followed by a long /a:./.Example/ ⟨قُرْآن⟩ :qurˈʔa:n./

Sukun(◌ْ) The sukūn is a circle-shaped diacritic placed above a letter. It indicates that the consonant to which it is attached is not followed by a vowel; this is a necessary symbol for writing consonant-vowel-consonant syllables, which are very common in Arabic. Example ⟨دَدْ⟩ :dad.

Tanwin (final post nasalized or long vowels(◌ً ◌ٍ ◌ٌ)):In some Semitic languages, such as Arabic, *nunation* (Arabicتنوين : *tanwīn*) is the addition of a final nūn sound to a noun or adjective to indicate that it is fully declinable and syntactically unmarked for definiteness.There are three of these vowel diacritics, and the signs indicate, from left to right, the endings-un (nominative case), -in (genitive), and -an (accusative) .These endings are used as non-pausal grammatical indefinite case endings in literary Arabic or classical Arabic (trip totes only)**.**

Shaddah (consonant gemination mark)(◌ّ): The *shadda* or *Shaddah* (*Shaddah*), or *tashdid* (*tashdīd*), is a diacritic shaped like a small written Latin "w". It is used to indicate gemination (consonant doubling or extra length), which is phonemic in Arabic.

Part of Speech	Specific Makhraj	Letters from Specific *Makhraj:* Arabic letter, Name, Value in International Phonetic Alphabet (IPA[1]).			No
Al-Khayshum The nasal passage (Fig. A.2)	This is a single *Makhraj* for the sound of Ghunnah. This is inherent characteristic of the two letters that cannot change.	/, ن (Nūn) with Shaddah (Fig. A.2. (b))	/,[م (Mīm) with Shaddah (Fig. A.2. (a))		1
Ash-Shafataan The two lips (Fig. A.3)	Between the two lips	و (Wāw), /w/ (Fig. A.3. (c))	م (Mīm), /m/ (Fig. A.3. (b))	ب (Bāʾ), /b/ (Fig. A.3. (a))	2
	Between the inside of the lower lips and the upper incisors.	ف (Fāʾ),/f/ (Fig. A.3. (d))			3
Al-Lisaan The tongue (Fig. A.4)	The innermost part of the tongue next to the throat touching the roof of the mouth opposite it.	ق (Qāf), /q/ (Fig. A.4. (e))			4
	The innermost part of the tongue towards the mouth touching the roof of the mouth opposite it.	ك (Kāf),/k (Fig. A.4. (d))/			5
	One or both edges of the tongue, usually the left along with the upper back molars.	ض (Ḍād), /dˤ/ (Fig. A.4. (j))			6
	Between the edges of the tongue usually the right and the gums of the front molars, canine and incisors.	ل (Lām), /l/ (Fig. A.4. (i))			7
	Between the tips of the tongue and the gums of the two upper central incisors.	ن (Nūn), /n/ (Fig. A.4. (h))			8
	Between the upper part of the tip of the tongue and the gums of the two upper central incisors.	ر (Rāʾ), /r/ (Fig. A.4. (f&g))			9
	The middle of tongue with the opposite from the roof of the mouth.	ج (Gīm) ,[ʤ] ~ [ʒ] ~ [g]. (Fig. A.4. (c))	ش (Shīn),/ʃ/ (Fig. A.4. (b))	ي (Yāʾ), /j/. (Fig. A.4. (a))	10
	The tip of the tongue near the inner plates of the upper central incisors.	ز (Zayn), /z/ (Fig. A.4. (b))	س (Sīn), /s/ (Fig. A.4. (b))	ص (Ṣād),/sˤ/ (Fig. A.4. (b))	11
	The tip of the tongue along with its upper surface touching the roots of the central incisors.	ت (Tāʾ), /t/ (Fig. A.4. (o))	د (Dāl), /d/ (Fig. A.4. (o))	ط (Ṭāʾ),/tˤ/ (Fig. A.4. (p))	12
	Between the upper surface of the tongue near the end of the tips of the two upper central incisors.	ث (Ṯāʾ),/θ/ (Fig. A.4. (k))	ذ (Ḏāl),/ð/ (Fig. A.4. (k))	ظ (Ẓāʾ) ,[ðˤ] ~ [zˤ] (Fig. A.4. (l))	13
Al-Halq The throat (Fig. A.5)	The deepest part of the throat.	[ʔ](hamzah) (Fig. A.5. (c))	هـ (Hāʾ),/h/ (Fig. A.5. (c))		14
	The middle of the throat.	ع (ʿayn), /ʕ/ (Fig. A.5. (b))	ح (Ḥāʾ), /ħ/ (Fig. A.5. (b))		15
	The nearest part of the throat.	غ (Ġain),/ ɣ/ (Fig. A.5. (a))	خ (Ḫāʾ),/x/ (Fig. A.5. (a))		16
Al-Jawf The chest or interior (Fig. A.6)	This is a single Makhraj comprising the empty space of the open mouth for the letters of *Madd* or elongation preceded by the Arabic vowels.	وُ(ū or oo), /uː/ (Fig. A.6. (b))	يِ(ī or ee), /iː/ (Fig. A.6. (a))	ا(ā), /aː/ (Fig. A.6. (c))	17

Table A.II THE 17 SPECIFIC MAKHRAJ AND THE LETTERS FROM EACH.

A.3 ESTABLISHING THE EXACT POINT OF ARTICULATION OF THE ARABIC LETTERS

In order pinpoint the exact point of articulation of each of the letters as outlined in the previous section, the letters are pronounced with *Sukun (ْ)* preceded by the letter *Hamza* (ا).

Example : To establish the exact point of articulation for the letter ر (Rāʾ), /r/, we pronounce اِرْ, (īr), /i:r/.

A.4 CHARACTERISTICS OF THE ARABIC LETTERS (*SIFAAT*)

Sifaat is the plural in Arabic while *Sifah* is singular. It is the special characteristics that are inherent in each of the Arabic letters; the *Sifaat* enable the differentiation between letters that comes from the same *Makhraj*, as already shown in section 2. Perfection in pronunciation cannot be obtained unless both the ***S****ifah* (characteristic/quality) and *Makhraj* are correct.

There are 18 *Sifaat* for the Arabic letters, 10 with opposites and 8 without opposites. The 10 with opposites are 5 in pairs (Table A.III). Each of the Arabic letters has at least 5 *Sifaat* with some having additional *sifah*(Table A.IV). Table A.V shows the classification of each letter depending on characteristics.

NOTICE

- If the letters being merged come from the same *Makhraj* (point of articulation), and have the same *Sifah* (characteristic) the letters called : ***"Mutamaathil"*** **(like letters).**
- when the letters being merged come from two *Makhaarij* – close in proximity, and have different (but similar) *Sifaat the letters called* ***"Mutaqaarib"*** **(close letters).**
- If the letters being merged come from the same *Makhraj*, but have different *Sifaat, the letters called "****Mutajaanis****"* **(Homogeneous).**

	Sifaat		Opposite
1	***The Strength (As-Shiddah)*** Trapping the flow of sound, strengthening the complete reliance on the *Makhraj* (point of articulation), associated with these letters: ت ك ط ب د ق أ ج	***Moderation(At-Tawasut)*** In between the strength and the weakness is the moderation, where the sound emerges but does not flow from the point of articulation associated with these letters : ر م ن ع ل	***The Weakness (Ar-Rikhaawah)*** A flow of sound during pronunciation, weaking the reliance on *Makhraj*(point of articulation) associated with all the letters not included in "The Strength (*As-Shiddah*)" and "Moderation(*At-Tawasut*)".
2	***The Whispers (Al-Hams)*** A flow of breath (air) during pronunciation due to weakness in the reliance on the *Makhraj* (point of articulation)associated with these letters: ت ك ص س خ ه ش ث ح ف		***The Audible (Al-Jahr)*** The trapping of the flow of breath (air) due to heavy dependence on the *Makhraj* (point of articulation) associated with all the letters not in "The Whispers (Al-Hams)".
3	***The Elevation (Al-Istia'la)*** Raising the tongue to the roof of the mouth during articulation, associated with these letters: ظ ط ق غ ص ض خ		***The Lowering (Al-Istifaal)*** Lowering the tongue to the floor of the mouth during articulation associated with all the letters not in "The Elevation (*Al-Istia'la*)".
4	***The Closing (Al-Itibaaq)*** The meeting of the tongue and what is opposite it from the roof of the mouth during articulation, associated with the letters: ظ ط ص ض		***The Opening (Al-Infitaah)*** The separation of the tongue from the roof of the mouth during articulation, associated with all the letters not included in "The Closing (*Al-Itibaaq*)".
5	***The Fluency (Al-Idhlaaq)*** The easy flowing of the letters: ب ن ل ر م ف from the tip of the tongue and lips. However this *Sifah* and its opposite is not included in the study of *Tajweed* but included for completeness.		***The Restraint (Al- Ismaat)*** The emergence of the remaining letters not included in "The Fluency (*Al-Idhlaaq*" from inside of the mouth and throat

Table A.III *SIFAAT* WITH OPPOSITE

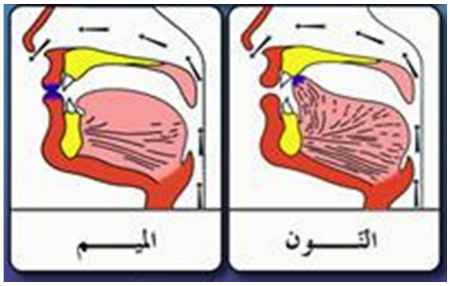
1	***The Whistling (As-Safeer)*** A sound emerging between the tip of the tongue and the upper central incisors which resembles the sound of a bird, associated with ز س ص. It is usually like a buzzing sound with ز.
2	***The Vibration (Al-Qalqalah)*** The vibration of the *Makhraj* (point of articulation) with the emergence of the letter when it has *Sukun* (ْ) associated with letters in this phrase: د ب ج ط ق
3	***The Ease (Al- Leen)*** This is pronunciation without exertion or difficulty. It is associated with letters و and ى with *Saakin* (ْ) preceded by *Fatḥah*(َ).
4	***The Drifting (Al-Inhiraaf)*** The inclination of the letter after its articulation from the *Makhraj* towards another *Makhraj* associated with ل and ر. ل inclines towards the tip of the tongue and ر inclines towards the *Makhraj* of ل.
5	***The Repetition (At-Takreer)*** This is the natural tendency to vibrate or roll the tongue during articulation of the letter ر. However this is to be avoided for correct pronunciation by controlling the tongue and not relaxing it.
6	***The Diffusion (At-Tafashshii)*** The spreading of air throughout the mouth during articulation of the letter ش
7	***The Elongation (Al- Istitaalah)*** This is the extension of the sound over the entire edge of the tongue from front to back during articulation, associated with letter ض .
8	***The Nasalisation(Al-Ghunnah)*** This is the sound emitted from the nose, an inherent characteristics of letters نand م when accompanied by *Sukun* or *Shaddah*. *Ghunnah* emerges from the nose when the flow of sound is blocked in the mouth, by the tongue with ن and by the lips with م.

Table A.IV SIFAAT WITH NO OPPOSITE

Characteristics	ء	هـ	ي	و	ن	م	ل	ك	ق	ف	غ	ع	ظ
The Audible *(Al-Jahr)*	ء		ي	و	ن	م	ل		ق		غ	ع	ظ
The Strength *(As-Shiddah)*								ك	ق				
The Elevation *(Al-Istia'la)*									ق		غ		ظ
The Closing *(Al-Itibaaq)*													ظ
The Restraint *(Al- Ismaat)*	ء	هـ	س	و				ك	ق		غ	ع	ظ
The Whistling (As-Safeer)													
The Vibration *(Al-Qalqalah)*									ق				
The Drifting *(Al-Inhiraaf)*							ل						
The Repetition *(At-Takreer)*													
The Diffusion *(At-Tafashshii)*													
The Elevation *(Al-Istia'la)*													
The Nasalisation*(Al-Ghunnah)*					ن	م							
The Whispers *(Al-Hams)*		هـ						ك		ف			
The Weakness *(Ar-Rikhaawah)*		هـ	ي	و						ف	غ		ظ
The Lowering *(Al-Istifaal)*	ء	هـ	ي	و	ن	م	ل	ك		ف	غ	ع	
The Opening *(Al-Infitaah)*	ء	هـ	ي	و	ن	م	ل	ك	ق	ف	غ	ع	
The Fluency *(Al-Idhlaaq)*					ن	م	ل			ف			
The Ease *(Al- Leen)*			ي	و									
Moderation*(At-Tawasut)*					ن	م	ل					ع	
Strong *Sifaat*	**3**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**5**		**3**	**2**	**4**
Weak *Sifaat*	**2**	**4**	**4**	**4**	**4**	**4**	**4**	**3**	**1**	**5**	**3**	**3**	**1**
Total	**5**	**6**	**6**	**6**	**6**	**6**	**6**	**5**	**6**	**5**	**6**	**5**	**5**

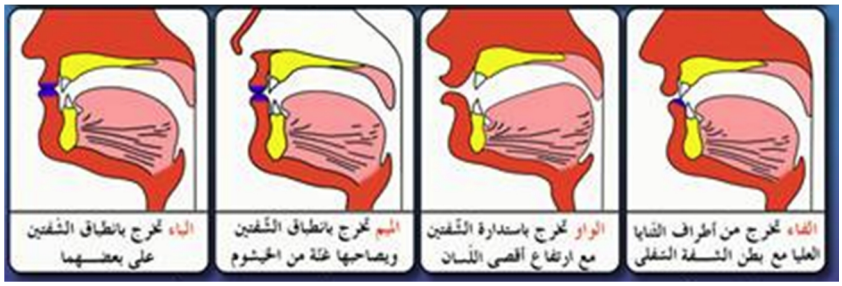
Characteristics	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	ا
The Audible *(Al-Jahr)*	ط	ض				ز	ر	ذ	د			ج			ب	ا
The Strength *(As-Shiddah)*	ط								د					ت	ب	
The Elevation *(Al-Istia'la)*	ط	ض	ص							خ						
The Closing *(Al-Itibaaq)*	ط	ض	ص													
The Restraint *(Al- Ismaat)*	ط	ض	ص	ش	س	ز		ذ	د	خ	ح	ج	ث	ت		ا
The Whistling (As-Safeer)			ص		س	ز										
The Vibration *(Al-Qalqalah)*	ط								د			ج			ب	
The Drifting *(Al-Inhiraaf)*							ر									
The Repetition *(At-Takreer)*							ر									
The Diffusion *(At-Tafashshii)*				ش												
The Elevation *(Al-Istia'la)*		ض														
The Nasalisation*(Al-Ghunnah)*																
The Whispers *(Al-Hams)*			ص	ش	س					خ	ح		ث	ت		
The Weakness *(Ar-Rikhaawah)*		ض	ص	ش	س	ز		ذ		خ	ح		ث			ا
The Lowering *(Al-Istifaal)*				ش	س	ز	ر	ذ	د		ح	ج	ث	ت	ب	ا
The Opening *(Al-Infitaah)*				ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	ا
The Fluency *(Al-Idhlaaq)*							ر								ب	
The Ease *(Al- Leen)*																ا
Moderation*(At-Tawasut)*							ر									
Strong *Sifaat*	**6**	**5**	**4**	**2**	**2**	**3**	**3**	**2**	**4**	**2**	**1**	**4**	**1**	**2**	**3**	**2**
Weak *Sifaat*	**0**	**1**	**2**	**4**	**4**	**3**	**4**	**3**	**2**	**3**	**4**	**2**	**4**	**3**	**3**	**4**
Total	**6**	**6**	**6**	**6**	**6**	**6**	**7**	**5**	**6**	**5**	**5**	**6**	**6**	**5**	**6**	**6**

Table A.V CLASSIFICATION OF EACH LETTER DEPENDING ON CHARACTERISTICS



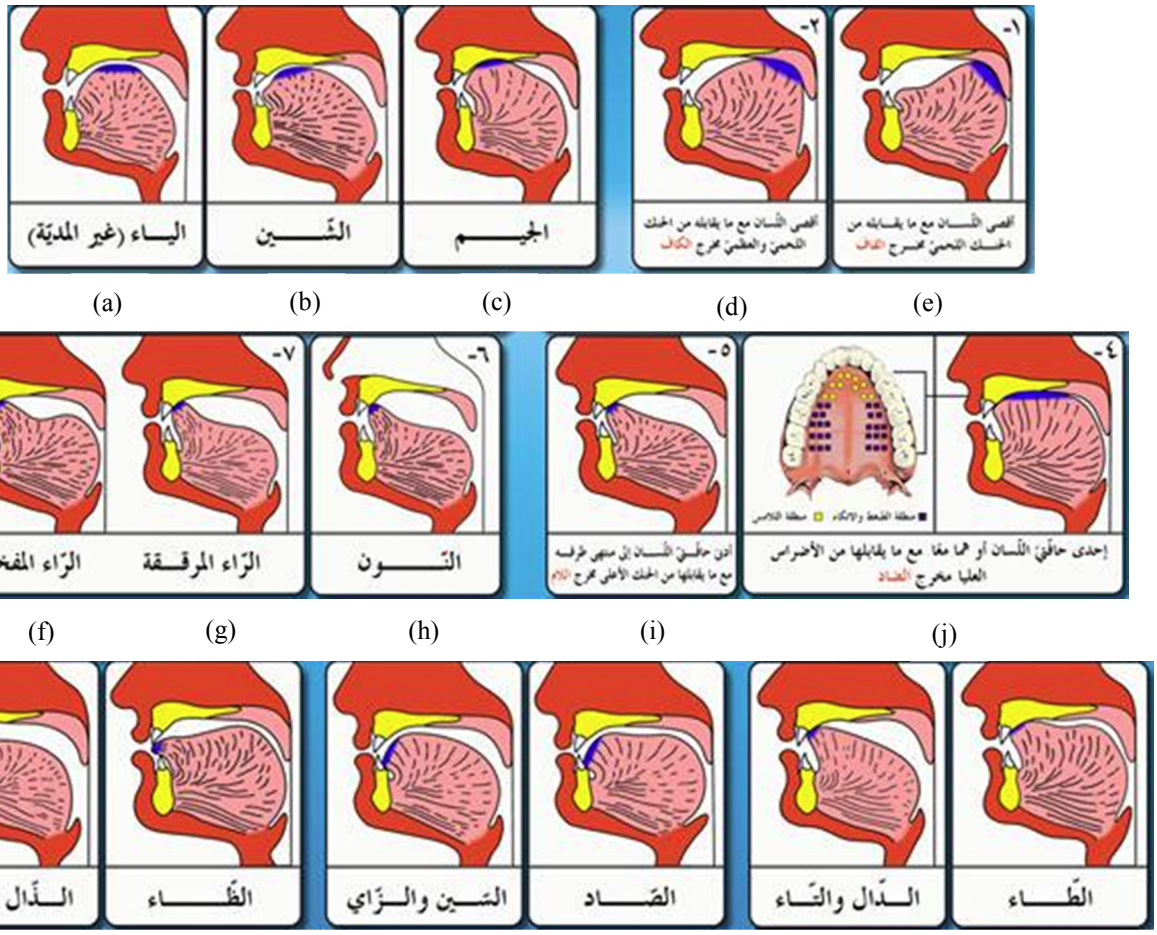
(a) (b)

Figure A.2. *"AL-KHAYSHUM", THE NASAL PASSAGE* (*MAKHRAJ 1*)



(a) (b) (c) (d)

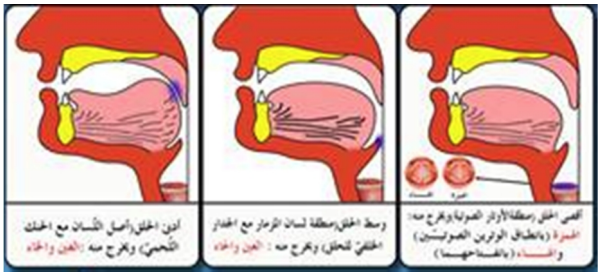
Figure A.3. *"ASH-SHAFATAAN", THE TWO LIPS* (*MAKHRAJ 2,3*)



(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o) (p)

Figure A.4. *"AL-LISAAN", THE TONGUE* (*MAKHRAJ 4-13*).



(a) (b) (c)

Figure A.5. *"AL-HALQ", THE THROAT (MAKHRAJ 14-16*)

(a) (b) (c)

Figure A.6 THE CHEST OR INTERIOR (*MAKHRAJ 17*)

Appendix B

The Recording Software

B.1 INTRODUCTION

In order to facilitate the recording process, a program has been specifically developed for this, using the visual programming language.

The program can be used by anyone for the recording process as it is designed the way it the error possibility very low.

B.2 SOFTWARE WINDOWS

The Program has three main windows:

The first is for choosing the letter that which the child finds pronunciation difficulty (Fig B.1).

Figure B.1. THE MAIN INTERFACE TO CHOOSE THE LETTER THAT THE CHILD FINDS DIFFICULTY TO PRONOUNCE

The second is meant for recording. It has 3 parts; one has to finish each part to move to the next (Fig B.2).

- ✓ The first section for child's data registration
- ✓ The second for choosing the disorder type
- ✓ The third section is for confirmation process

Third window for confirmation and submission (Fig B.3).

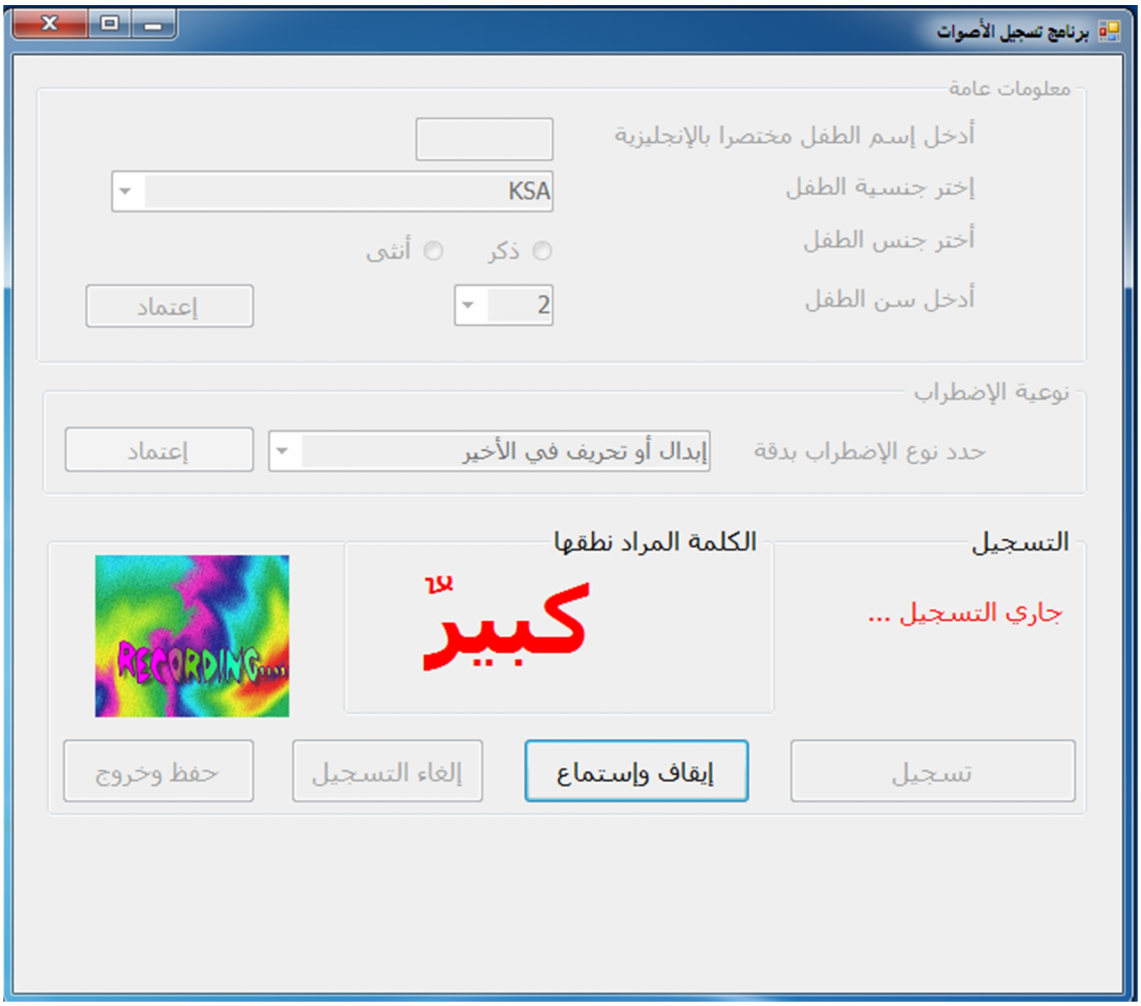


Figure B.2. CHILD'S DATA REGISTRATION, DISORDER TYPE AND CONFIRMATION OF REGISTRATION

Upon completion of the recording process, voices are recorded automatically in the appropriate folders holding the child's name, nationality, age and sex, without the user's intervention.

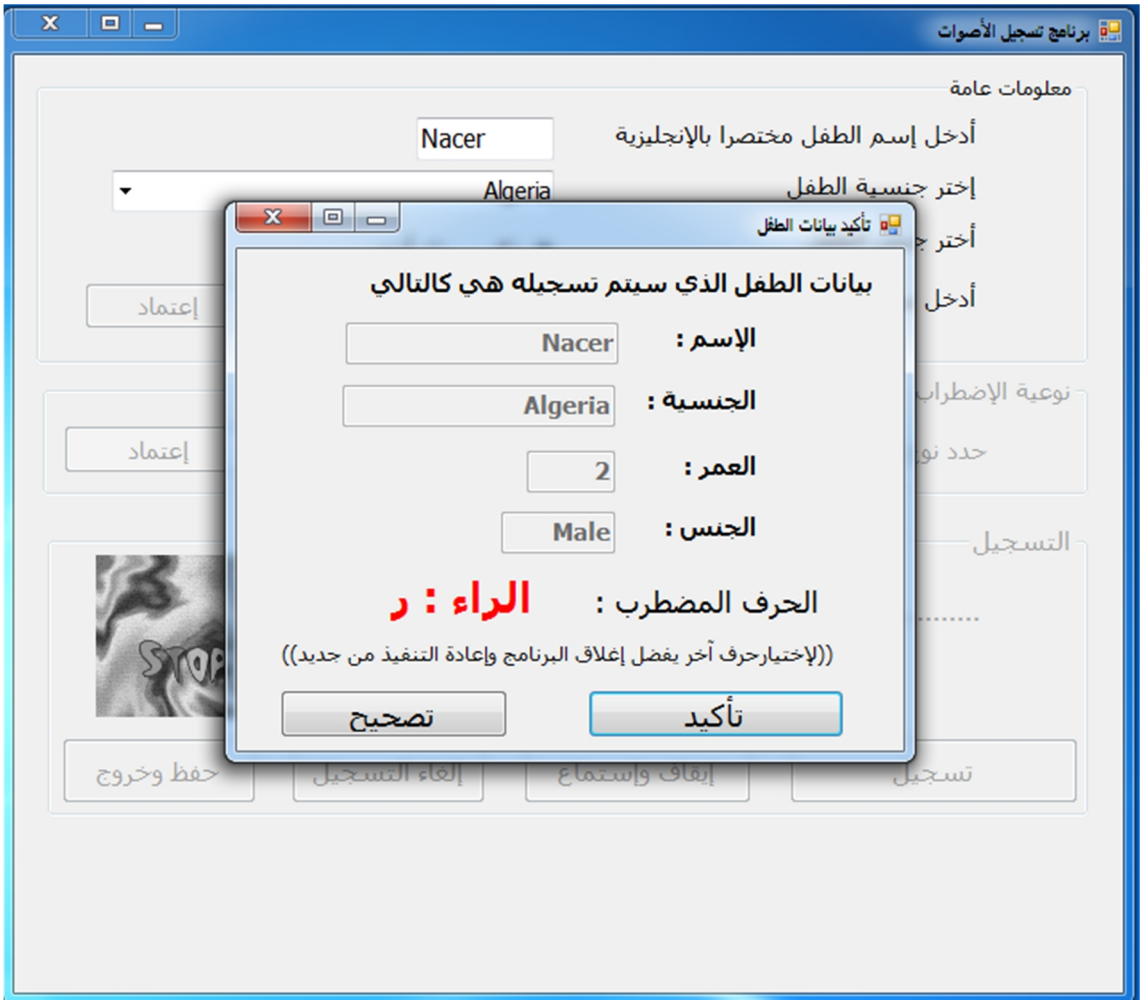


Figure B.3. CONFIRMATION AND SUBMISSION

Bibliography

[1] Stuckless, R. (1994). Developments in real-time speech-to-text communication for people with impaired hearing. In M.Ross(Ed.), Communication access for people with hearing loss (pp.197-226). Baltimore, MD: York Press.

[2] A. G. Adami, "Automatic speech recognition: from the beginning to Portuguese language", in Proc. International Conference on Computational Processing of Portuguese Language (PRO- POR), 2010.

[3] Flanagan, J.L, "Speech analysis, synthesis, and perception". Springer, Berlin (1965).

[4] B.H. Juang and L.R. Rabiner, "Automatic speech recognition - a brief history of the technology development", Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

[5] S.Furui, "50 years of progress in speech and speaker recognition", Proceedings of the International Conference on Speech and Computer, pp 1–9 (2005).

[6] J.L.Flanagan, "Speech Analysis, Synthesis and Perception", Second Edition, Springer-Verlag,1972.

[7] H. Dudley, R. R. Riesz, and S. A. Watkins, "A Synthetic Speaker", J. Franklin Institute, Vol.227, pp. 739-764, 1939.

[8] V. Dellwo, M. Huckvale, and M. Ashby, "How Is Individuality Expressed in Voice?. An Introduction to Speech Production and Description for Speaker Classification". In Christian Müller, editor, Speaker Classification I, 4343 of Lecture Notes in Computer Science, pages 1–20. Springer Berlin / Heidelberg, 2007.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. The HTK Book (for HTK Version 3.4). 2009.

[10] P. T Nguyen "Automatic Speaker Classification Based on Voice Characteristics", Master thesis, University of Canberra, 2010.

[11] L. Bahl, F. Jelinek, and R. Mercer, "BA maximum likelihood approach to continuous speech recognition", IEEE Trans.Pattern Anal. Mach. Intell., vol. PAMI-5, no. 2, pp. 179–190, Mar. 1983.

[12] Picheny, M.; Nahamoo, D.; Goel, V.; Kingsbury, B.; Ramabhadran, B.; Rennie, S.J.; Saon, G., "Trends and advances in speech recognition," IBM Journal of Research and Development , vol.55, no.5, pp.2:1,2:18, Sept.-Oct. 2011.

[13] Urmila Shrawankar, Dr. V M Thakare, “Techniques For Feature Extraction In Speech Recognition System : A Comparative Study” ,IJ-CA-ETS : International Journal Of Computer Applications In Engineering, Technology And Sciences, ISSN: 0974-3596, Volume 2 : Issue 2, Pg: 412-418, 2010.

[14] Shanthi Therese, S., and Chelpa Lingam. "Review of Feature Extraction Techniques in Automatic Speech Recognition." International Journal of Scientific Engineering and Technology , Volume No.2, Issue No.6, pp : 479-484, 2013.

[15] R. Mammone, X. Zhang, and R. Ramachandran, “Robust speaker recognition: A feature-based approach”, IEEE Signal Processing Mag., vol. 13, no. 5, 1996, pp. 58–71.

[16] M. A. Anusuya and S. K. Katti, "Speech Recognition by Machine, A Review", (LICSIS) International Journal of Computer Science and Information Security, vol. 6, no.23, pp. 181-205,2009.

[17] R. Togneri and D. Pullella, “An overview of speaker identification: Accuracy and robustness issues”, Circuits and Systems Magazine, IEEE, vol. 11, no. 2, pp.23–61, 2011.

[18] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “Hidden Markov model toolkit (htk) version 3.4 user’s guide”, 2002.

[19] J. Deller, J. Hansen, and J. Proakis, “Discete-Time Processing of Speech Signals”. Wiley,N.Y, 2000.

[20] S. Schȯtz, “Acoustic analysis of adult speaker age”. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intel ligence and Lecture Notes in Bioinformatics),4343 LNAI:88–107, 2007.

[21] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Frame vs. Turn-Level: “Emotion Recognition from Speech Considering Static and Dynamic Processing”. In Ana Paiva, Rui Prada, and Rosalind Picard, editors, A ective Computing and Intel ligent Interaction, 4738 of Lecture Notes in Computer Science, pages 139–147. Springer Berlin / Heidelberg, 2007.

[22] F. Eyben, M. Wöllmer, and B. Schuller," OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit". Proceedings - 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009.

[23] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals". 2, pages 881–884, 2007. - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007.

[24] Jain, A.K.; Duin, R. P W; Jianchang Mao, "Statistical pattern recognition: a review," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.22, no.1, pp.4,37, Jan 2000.

[25] L. Xu, A. Krzyzak, and C.Y. Suen, "Methods for Combining Multiple Classifiers and Their Applications in Handwritten Character Recognition",: IEEE Trans. Systems, Man, and Cybernetics, vol. 22, pp. 418-435, 1992.

[26] M.Al-Zabibi, "An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition", the British Library in Association with UMI, 1990.

[27] R., Gordon (Ed.). "Ethnologue: Languages of the World". SIL International, Dallas, TX, 2005.

[28] K. Kirchho, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta, "Novel Approaches to Arabic Speech Recognition," Technical Report, Johns-Hopkins University, 2002.

[29] Mrayati M., Makhoul J., "Man-machine communication and the Arabic language", Lecture notes, Applied Arabic linguistics and signal and information processing, pp : 133-145, 1984.

[30] [http://www.sakhr.com/]

[31] Al-Ani S., "Abstract and concrete interaction in the arabic sound system", Islamic and middle eastern societies, ed. Robert oslan and Salman Al-Ani, Amana Books 1987.

[32] Djoudi M., "Contribution à l'etude et à la reconnaissance de la parole en arabe standard », Thèse de l'Université de Nancy I, 1991.

[33] D. Vergyri, K. Kirchhoff, K. Duh, A. Stolcke, "Morphology-based language modeling for Arabic speech recognition", In INTERSPEECH, pp. 2245-2248, 2004.

[34] D. Vergyri, K. Kirchhoff. “Automatic diacritization of Arabic for acoustic modeling in speech recognition”, In Ali Farghaly and Karine Megerdoomian, editors, COLING 2004, Computational Approaches to Arabic Scriptbased Languages, pp. 66–73, Geneva, Switzerland, 2004.

[35] H. Satori, M. Harti, N. Chenfour, “Introduction to Arabic Speech Recognition Using CMUSphinx System”, Proceedings of Information and Communication Technologies International Symposium (ICTIS'07), Fes, Morocco, pp. 139-115, July 2007.

[36] K. Daqrouq, K.Y. Al Azzawi, "Arabic vowels recognition based on wavelet average framing linear prediction coding and neural network", Speech Communication, Volume 55, Issue 5, June 2013, Pages 641-652.

[37] N.Hammami, M.Bedda, N.Farah, "Spoken Arabic Digits recognition using MFCC based on GMM," Sustainable Utilization and Development in Engineering and Technology (STUDENT), 2012 IEEE Conference on , vol., no., pp.160,163, 6-9 Oct. 2012.

[38] N.Hammami, M.Bedda, N.Farah,"The second-order derivatives of MFCC for improving spoken Arabic digits recognition using Tree distributions approximation model and HMMs," Communications and Information Technology (ICCIT), 2012 IEEE Conference on , vol., no., pp.1,5, 26-28 June 2012.

[39] K. Saeed and M. Nammous, “A speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image,” IEEE Trans. Ind. Electron., vol. 54, no. 2, pp. 887–897, Apr. 2007.

[40] N. Hammami, M. Bedda, and N. Farah, " Tree distributions approximation model for robust discrete speech recognition" ,International Journal of Speech Technology, vol. 15, no .4,pp 455-462, Springer, 2012.

[41] Sarikaya, R.; Afify, M.; Yonggang Deng; Erdogan, H.; Yuqing Gao, "Joint Morphological-Lexical Language Modeling for Processing Morphologically Rich Languages With Application to Dialectal Arabic," Audio, Speech, and Language Processing, IEEE Transactions on , vol.16, no.7, pp.1330,1339, Sept. 2008.

[42] Yun Lei; Hansen, J.H.L., "Dialect Classification via Text-Independent Training and Testing for Arabic, Spanish, and Chinese," Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.1, pp.85,96, Jan. 2011.

[43] Hai-Son Le; Oparin, I.; Allauzen, A.; Gauvain, J.; Yvon, F., "Structured Output Layer Neural Network Language Models for Speech Recognition," Audio, Speech, and Language Processing, IEEE Transactions on , vol.21, no.1, pp.197,206, Jan. 2013.

[44] Lehr, M.; Shafran, I., "Learning a Discriminative Weighted Finite-State Transducer for Speech Recognition," Audio, Speech, and Language Processing, IEEE Transactions on , vol.19, no.5, pp.1360,1367, July 2011.

[45] Soltau, Hagen; Saon, G.; Kingsbury, B.; Kuo, Hong-Kwang Jeff; Mangu, L.; Povey, D.; Emami, A., "Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program," Audio, Speech, and Language Processing, IEEE Transactions on , vol.17, no.5, pp.884,894, July 2009.

[46] Halima ABIDET-BAHI, “NESSR : Un système neuro-expert pour la reconnaissance de la parole”, Phd thesis, University of badji mokhtar , Annaba, Algeria, 2005.

[47] Alsulaiman, M.; Muhammad, G.; Ali, Z.,"Comparison of voice features for Arabic speech recognition," Sixth International Conference on Digital Information Management (ICDIM), pp.90-95, 26-28,Sept. 2011.

[48] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," Signal Processing Magazine, IEEE, vol. 22, pp. 89-100, 2005.

[49] H. Songfang and S. Renals, "Hierarchical Bayesian Language Models for Conversational Speech Recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 18, pp. 1941-1954, 2010.

[50] A. Wiesel, Y. C. Eldar, and A. O. Hero, "Covariance Estimation in Decomposable Gaussian Graphical Models," Signal Processing, IEEE Transactions on, vol. 58, pp. 1482-1492, 2010.

[51] A. Miguel, A. Ortega, L. Buera, and E. Lleida, "Bayesian Networks for Discrete Observation Distributions in Speech Recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, pp. 1476-1489, 2011.

[52] V. Y. F. Tan, A. Anandkumar, T. Lang, and A. S. Willsky, "A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures," Information Theory, IEEE Transactions on, vol. 57, pp. 1714-1735, 2011.

[53] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures," Signal Processing, IEEE Transactions on, vol. 58, pp. 2701-2714, 2010.

[54] S. El Fkihi, M. Daoudi, and D. Aboutajdine, "The mixture of K-Optimal-Spanning-Trees based probability approximation: Application to skin detection," Image and Vision Computing, vol. 26, pp. 1574-1590, 2008.

[55] A. Torsello and E. R. Hancock, "Learning shape-classes using a mixture of tree-unions," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, pp. 954-967, 2006.

[56] S. Ioffe and D. Forsyth, "Mixtures of trees for object recognition," in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, 2001, pp. II-180-II-185 vol.2.

[57] N. Hammami and M. Sellami, "Tree distribution classifier for automatic spoken Arabic digit recognition," in Internet Technology and Secured Transactions, 2009.IEEE, ICITST 2009. International Conference for, 2009, pp. 1-4.

[58] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," Information Theory, IEEE Transactions on, vol. 14, pp. 462-467, 1968.

[59] M. Meila, "An accelerated Chow and Liu algorithm: fitting tree distributions to high dimensional sparse data," 1999.

[60] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference: Morgan Kaufmann, 1988.

[61] T. H. Gormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to algorithms," ed: MIT press Cambridge, MA, 1990.

[62] U. o. B.-M. Laboratory of Automatic and Signals, "Spoken Arabic Digits " UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit], 2008.

[63] M. Kudo, J. Toyama, and M. Shimbo, "Japanese Vowels," UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Japanese+Vowels], 1999.

[64] R. Gray, "Vector quantization," ASSP Magazine, IEEE, vol. 1, pp. 4-29, 1984.

[65] N. Hammami, M.Beda, N.Farah , " HMM parameters estimation based on cross-validation for Spoken arabic digits recognition", international conference on communications, computing and control applications (CCCA), IEEE, 2011, pp. 1-4.

[66] Q. Ji, J. Luo, D. Metaxas, A. Torralba, T. S. Huang and E. B. Sudderth, "Guest Editors' Introduction to the Special Section on Probabilistic Graphical Models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 10, pp. 1729-1732, Oct. 2009.

[67] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.

[68] Sklar, A. (1959), “Fonctions de repartition ‘a n dimensions et leurs marges,” Publ.Inst. Statist. Univ. Paris 8, 229-231.

[69] G. Mercier, G. Moser, and S. B. Serpico, “Conditional copulas for change detection in heterogeneous remote sensing images,” IEEE Trans.Geosci. Remote Sens., vol. 46, no. 5, pp. 1428–1441, May 2008.

[70] Y. Stitou, N. Lasmar, and Y. Berthoumieu, “Copula based multivariate gamma modeling for texture classification,” in Proc. ICASSP, 2009, pp. 1045–1048.

[71] A. Sundaresan, P. Varshney, and N. Rao, “Copula-based fusion of correlated decisions,” IEEE Trans. Aerosp. Electron. Syst., vol. 47, no. 1,pp. 454–471, Jan. 2011.

[72] Cui, Y.; Yang, J.; Yamaguchi, Y.; Singh, G.; Park, S.-E.; Kobayashi, H., "On Semiparametric Clutter Estimation for Ship Detection in Synthetic Aperture Radar Images," Geoscience and Remote Sensing, IEEE Transactions on , vol.51, no.5, pp.3170,3180, May 2013.

[73] Voisin, A.; Krylov, V. A.; Moser, G.; Serpico, S. B.; Zerubia, J., "Classification of Very High Resolution SAR Images of Urban Areas Using Copulas and Texture in a Hierarchical Markov Random Field Model," Geoscience and Remote Sensing Letters, IEEE , vol.10, no.1, pp.96,100, Jan. 2013.

[74] Kwitt, R.; Meerwald, P.; Uhl, A.; , "Efficient Texture Image Retrieval Using Copulas in a Bayesian Framework," Image Processing, IEEE Transactions on , vol.20, no.7, pp.2063-2077, July 2011.

[75] Krylov, V.A.; Moser, G.; Serpico, S.B.; Zerubia, J.; , "Supervised High-Resolution Dual-Polarization SAR Image Classification by Finite Mixtures and Copulas," Selected Topics in Signal Processing, IEEE Journal of , vol.5, no.3, pp.554-566, June 2011.

[76] Chaorong Li; Jianping Li; Bo Fu, "Magnitude-Phase of Quaternion Wavelet Transform for Texture Representation Using Multilevel Copula," Signal Processing Letters, IEEE , vol.20, no.8, pp.799,802, Aug. 2013.

[77] Hammami, N.; Bedda, M.; Nadir, F., "Probabilistic classification based on copula for speech recognitation: an overview," Computer Applications Technology (ICCAT), 2013 International Conference on , vol., no., pp.1,3, 20-22 Jan. 2013.

[78] Nelsen, Roger B.: An Introduction to Copulas. Springer-Verlag New York. Inc,

[79] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture models,” IEEE Trans. Speech Audio Process., vol. 3, no. 1, pp. 72–83, Jan. 1995.

[80] Hammami, N., Bedda, M., Farah, N., 2011. HMM parameters estimation based on cross-validation for Spoken Arabic Digits recognition. IEEE Proc. 2011 International Conference on Communications, Computing and Control Applications (CCCA), 3-5.

[81] NG, Eddie KH. Kernel-based copula processes.. PhD Thesis. University of Toronto.2010.

[82] Dr. Mustafa Fahmi, “Logopedics” , 5th edition, 1975, Dar Messr, Egypt. (in Arabic).

[83] Dr. Faisal Al-Afif “Speech and Language Disorders”, arabbook,[www.arabbook.com] (In Arabic).

[84] N.Hammami, M.Bedda, N.Farah, R.Lakehal-Ayat3 “Spoken Arabic Digits recognition based on (GMM) for e-Quran voice browsing: Application for blind category”, Proc.IEEE (ICAITHQ’13), Medina, KSA, Sep 2013.

CURRICULUM VITAE

Name and surname: HAMMAMI NacerEddine

Date of Birth: July 12th, 1982 **in** Souk Ahras (Algeria)

Marital Situation: Married with Three children

Postal Address: Laboratoire LabGed,
University of Badji Mokhtar ,
Annaba, 23000, Algeria

Mobile: +213 5 53 31 09 58 ; +213 37 33 47 51

E-mail nacereddine.hammami@gmail.com

TITLES & DIPLOMAS

- **Since 2013** — ***PhD Student*** **in Computer Science**
  Option***: Artificial Intelligence and pattern recognition***
  Badji Mokhtar University- Annaba.

- **2008-2010** — *Magistère* **in Computer Science**
  Option***: Artificial Intelligence and pattern recognition***
  Badji Mokhtar University – Annaba.

- **2004-2005** — **Master in Computer Science**
  Option: ***Artificial Intelligence and pattern recognition***
  Ecole Polytechnique of Tours university of François Rabelais - France

- **1999-2004** — **Engineer in Computer Engineering**
  Option: ***Artificial Intelligence and pattern recognition***
  Computer Science Department, University of Jijel - Algeria

- **1998-1999** — **The General Secondary Education Certificate**
  Path: Sciences of Nature and Life
  Secondary School El-Aouana – Algeria

EXPERIENCES

- ***2007-2008*** **Lecturer**, Department Of Human sciences, University of Grenoble, France
- **2005-2007** **Manager of The company PRIMACOM Services** Ltd (Company for computer science services and telecommunication) Grenoble – France

TRAININGS (Collaboration and Assistance)

- **03/2005-07/2005** Training course in Master 2 Research Computer Science **"**Probabilistic models for imagery – Detection of human skin" , the main subject is to bloc a pornographic sites laboratory of computer science of the Ecole Polytechnique of Tours – France.

  CNRS(*National Center of Scientific Research) *project
  MathStic project (2004-2005) supported by CNRS
  (with Bruno Jedynak and Laurent Younes)
  http://www.enic.fr/people/daoudi/MathStic/projet_math_stic.html

- **01/2004-06/2004** Project of end of studies **"Study** and realization of an algorithm of data compression on a parallel architecture"
  Laboratory of computer science of Awlade Issa, Jijel-Algeria
- **09/2000-06/2003** Collaboration and assistance in a Unit of the National Constabulary El Khroub - Algeria
  - Computer Science and Statistics Management
  - Courses in Computer Sciences for the gendarmes of the Unit.

PROGRAM LANGUAGES & OPERATING SYSTEMS

- A very good knowledge of Java, C/C++, SQL, Turbo Pascal, Assembleur (Z80, 680X0, 80X86,…) and Delphi
  Especially:
  - The programming oriented subject,
  - Programming of Image Processing,
  - Parallel or Shared Programming (Rmi in Java),
  - Nets Programming, Data Mining and data analysis,
  - Software and Statistic Programming (R, S-pulse,Matlabe)
  - Excellent manipulation of peripherics, Office automation (Word, Excel, etc.)
- Operating Systems: DOS, Windows 95/98/2000/XP/NT, Unix/Linux, Mac.
- Miscelellaneous : LaTeX, HTML, Graphics,etc ..

PUBLICATIONS & SCIENTIFIC COMMUNICATIONS

[1] **N. Hammami**, M. Bedda, and N. Farah, " Tree distributions approximation model for robust discrete speech recognition" ,International Journal of Speech Technology, Springer, 2012.
http://link.springer.com/article/10.1007%2Fs10772-012-9141-9

[2] **N.Hammami**, M.Bedda, N.Farah, R.Lakehal-Ayat "Spoken Arabic Digits recognition based on (GMM) for e-Quran voice browsing: Application for blind category", Proc.IEEE (ICAITHQ'13), Medina, KSA, Sep 2013
http://nooritc.org/sites/default/files/ListofPublicationsEnglishNOORIC1435.pdf

[3] **N.Hammami**, M.Bedda, N.Farah, "Probabilistic Classification Based on Gaussian Copula for Speech Recognition: Application to Spoken Arabic Digits", Proc.IEEE (SPA'13), Signal Processing, Poland, Sep 2013.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[4] **Hammami Nacereddine**, Mohamed Goudjil and Alruily Meshrif, "*Probabilistic Classification Based on Copula for Text categorization: An Overview*", Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6618781&queryText%3DProbabilistic+Classification+Based+on+Copula+for+Text+categorization%3A+An+Overview%E2%80%9D

[5] Alruily Meshrif , **Hammami Nacereddine**, Goudjil Mohamed, "*Using Transitivity for developing Arabic Text Summarization System*", Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6618786&queryText%3DUsing+Transitivity+for+developing+Arabic+Text+Summarization+System

[6] Mohamed Goudjil , Mouloud Koudil , **Nacereddine Hammami** ,Mouldi Bedda and Meshrif Alruily; "*Arabic Text Categorization Using SVM Active Learning Technique : An Overview*"; Proc. IEEE (ICIA'13) The International Conference on Artificial Intelligence,Jun. 2013.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[7] Khelifa, M.A.; Boukabou, A.; **Hammami, N**., "Data transmission based on chaotic synchronization system," Computer Applications Technology (ICCAT), 2013 International Conference on , vol., no., pp.1,2, Tunisia, 20-22 Jan. 2013.
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6521983&queryText%3DData+transmission+based+on+chaotic+synchronization+system%2C%22+Computer+Applications+Technology

[8] **N.Hammami**, M.Bedda, N.Farah, "Spoken Arabic Digits recognition using MFCC based on GMM," Proc. IEEE, (STUDENT'12) Sustainable Utilization and Development in Engineering and Technology, pp.160,163,Malysia, 6-9 Oct. 2012.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[9] **N. Hammami**, M. Bedda, N.Farah ," Spoken Arabic Digits recognition Using MFCC based on GMM ", IEEE STUDENT'12 Conference, Malysia, 2012.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[10] **N. Hammami**, M. Bedda, N.Farah "The second-order derivatives of MFCC for improving spoken Arabic digits recognition using Tree distributions approximation model and HMMs," Communications and Information Technology (ICCIT), 2012 International Conference on , vol., no., pp.1,5, Tunisia, 26-28 June 2012.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[11] **N. Hammami**, M. Bedda, N.Farah "HMM parameters estimation based on cross-validation for Spoken Arabic Digits recognition," Communications, Computing and Control Applications (CCCA), 2011 International Conference on , vol., no., pp.1,4,Tunisia, 3-5 March 2011.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[12] **N. Hammami**, M. Bedda "Improved tree model for arabic speech recognition," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.5, no., pp.521,526,China, 9-11 July 2010.
http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=n+hammami

[13] **N. Hammami**, M. Bedda, Adaptation and preprocessing of the Repository dataset:"spoken Arabic digit",
http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit

[14] **N. Hammami**, M. Sellami "Tree distribution classifier for automatic spoken Arabic digit recognition," Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for , vol., no., pp.1,4,UK, 9-12 Nov. 2009
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5402575&queryText%3DTree+distribution+classifier+for+automatic+spoken+Arabic+digit+recognition

[15] A.Douzal, **N. Hammami**,C.gabry, "Local Analysis of multivariate time series". ISI 2007 (International Statistical Institute), Portugal 2007
http://hal.archives-ouvertes.fr/docs/00/36/04/89/PDF/ISI1-07.pdf

RECOMMENDATION & APPRECIATION

- **Pr. Nadir FARAH,**
  Laboratoire LabGed,
  Universite Badji Mokhtar Annaba
  Annaba, 23000, Algeria
  http://www.labged.net/
  Farah@labged.net
- **Pr. Mouldi BEDDA,**
  Faculty of Engineering
  University of Al Jouf
  Sakaka, Kingdom of Saudi Arabia
  Mouldi_ bedda@yahoo.fr
- **Pr. Jacques DEMONGEOT**,
  Director of the laboratory TIMC-IMAG (Techniques of Medical Engineering and Complexity – Computer Science, Mathematics and Applications of Grenoble- France). http://www-timc.imag.fr**.**
- **Pr. Daoudi Mohamed,**
  Professor in Computer Science and supervising of my Master's training, University of Lille - France.
  http://www.telecom-lille1.eu/people/daoudi/index.htm**.**
- **Pr. Jean-Yves Ramel**,
  Lecturer, Teacher Searcher at Poly Tech Tours, Department of Computer Science and Responsible(2) of the staff: Forms recognition and Images Analysis, http://www.li.univ-tours.fr**./**

LANGUAGE

- Mother Tongue: Arabic.
- Very Good knowledge of English (reading, writing and speaking).
- Very Good knowledge of French (reading, writing and speaking).

CULTURAL & SPORTS ACTIVITIES

Cultural Activities: All interesting cultural activities.

Sports: Football, Swimming, table tennis, PC chess.

Abstract

Automatic speech recognition (ASR) techniques are evolving in manufacturing and public use, in which the techniques of voice recognition have been implemented in many electronic multimedia devices for daily use or in other fields including manufacturing, the military, and medical science. These techniques and their applications are rapidly advancing; automatic speech recognition is becoming one of the most important means of communication between humans and machines. In spite of tremendous progress in theoretical and applied technology for ASR), it is limited to English and some other languages, whereas ASR research and application in the Arabic language is limited. Many devices that support speech recognition techniques do not support Arabic, although it is an international language used by more than 400 million people in more than 22 countries and an official language of the United Nations; it is an important language for over one billion people because of its relationship to the Islamic religion. This thesis is an attempt to enrich the field of Arabic speech recognition in which two new methods in the field of speech recognition have been suggested and adopted. The first method is a statistical method depending on the tree probability model, and the second method depends on copula, an advanced statistical tool. These methods were suggested and used for the first time and are an addition to ASR in Arabic and other languages. The methods have produced excellent results that compete with other known models in the field of speech recognition. After this thesis, a case study was conducted and an application using automatic speech recognition has been used- with the aid of speech specialists - to develop a speech database to diagnose and treat speech disorders in children. The database is unique, and we hope that it could be a reference for research and in the use of Arabic automatic speech recognition. The results of automated diagnosis for accent disorders among children using this method have been encouraging, given the nature of the database, which contains noise; the database has encouraged investment in this research for the manufacture of devices that could serve children, and it is dependent on the techniques developed in this thesis.