OPTIMAL FEATURE SELECTION USING MUTUAL INFORMATION FOR SPEECH RECOGNITION AT LOW SNR

BY

ESAM ABID AL-MASHABI

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

TELECOMMUNICATION ENGINEERING

JUNE 2011

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by ESAM ABID SAEED ALMASHABI under the direction of his thesis advisor and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE IN TELECOMMUNICATION ENGINEERING.

Thesis Committee

Dr. Mohammed A. Deriche (Chairman)

Dr. Samir H. Abdul-Jauwad (Member)

azzedine zerguine

Dr. Azzedine Zerguine (Member)

Department Chairman

Dean: College of Graduate Studies

12/7/2011

Date

Dedicated

To

My family for their love, patient, and support.

Special dedication to my parents for their continuous praying, support and encouragement towards the success.

ACKNOWLEDGMENTS

In the name of ALLAH, The most Gracious, The most merciful. Best prayers and blessings to our prophet, peace be upon to him and to his family.

All admire be to ALLAH, the Great, for giving me the courage, patience and determination to accomplish this work. I would like to thank King Fahd University of Petroleum and Minerals for supporting this research. My special gratitude is deeply expressed to my thesis advisor, Dr. Mohamed A. Deriche, for his patient guidance, great insights and motivated support thoughts on my work. I also thank my thesis committee members Dr. Samir H. Abdul-Jauwad and Dr. Azzedine Zerguine who have been very kind and supportive. I also thank Mr. Umar M. Johar for his cooperation and coordination to facilitate my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THESIS ABSTRACT

Name:ESAM ABID AL-MASHABI

Thesis Title: OPTIMAL FEATURE SELECTION USING MUTUAL INFORMATION FOR SPEECH RECOGNITION AT LOW SNR

Major Field:TELECOMMUNICATION ENGINEERING

Date of Degree:JUNE 2011

This thesis proposes a new approach for feature selection in speech recognition at low Signal to Noise Ratio (SNR). This concept is based on a two dimensional optimization of information content of different features by using the concept of mutual information. These two dimensions are: maximizing the information content towards the target classes and minimizing it between features.

This feature extraction and selection stage is critical stage in speech recognition systems, as it is the first step in both training and recognition, and upon which all other steps are highly dependent. There have been numerous feature sets proposed in speech recognition, many of which are correlated (LPC, Cepstrum, LAR, etc.). The selection of the optimal features from the various features is expected to lead to improved overall recognition accuracy.

The main aim of this research is to formulate the problem of feature selection as an optimization problem using information theory concepts. The proposed algorithm was first tested with English isolated word using standard databases. The algorithm was also tested with Arabic language. Our experiments were based on Arabic words obtained from a database that was collected locally. The experimental results showed an improved performance over existing techniques especially at low SNR.

MASTER OF SCIENCE DEGREE

King Fahd University of Petroleum & Minerals, Dhahran

ملخص الرسالة

الإسم:عصام عابد سعيد المشعبي

عنوان الرسالة: اختيار الصفات الصوتية المثلى لأنظمة التعرف الصو تي في بيئة ضوضائية باستخدام مفهوم المعلومات المتبادلة.

التخصص:هندسة الإتصالات

تاريخ التخرج:يونيو 2011

تقترح هذه الرسالة طريقة جديدة لاختيار السمات الصوتية المثلى لاستخدامها في أنظمة التعرف الصوتي في بيئة ضوضائية (نسبة الإشارة إلى الضوضاء منخفضة). القاعدة المقترحة تقوم على أساس الاستفادة المثلى من محتوى المعلومات للصفات المختلفة باستخدام مفهوم المعلومات المتبادلة. تعتبر مرحلة استخراج الصفات الصوتية المرحلة الحرجة في أنظمة التعرف الصوتي، لأنها هي الخطوة الأولى في عمليتي التدريب والتعرف ومن ثم تعتمد عليها كل المراحل الأخرى بدرجة كبيرة. هناك مجموعة عديدة من الصفات الصوتية تستخدم في أنظمة التعرف الصوتي إلا أن الكثير منها يكون متماثلاً ومترابطاً ومن أمثلتها (*LPC, Cepstrum, LAR*). من المتوقع أن مجموعة الصفات الصوتية المثلى ستؤدي إلى تحسن نتائج أنظمة التعرف الصوتية بشكل عام. إن الهدف الرئيسي لهذه الرسالة هو صياغة قاعدة علمية للاختيار الأمثل للصفات الصوتية. تم اختبار هذه القاعدة للتعرف على الأرقام الإنجليزية المجردة في نظام التعرف الصوتي المبني على قاعدة بيانات قياسية. ولتأكيد نجاح التجربة تم إجراء الإختبار على الأرقام العربية المجردة في نظام التعرف الصوتي العربي والذي بنيت قاعدة بياناته من أصوات جمعت محلياً ومن قواعد بيانات معتمدة.

رسالة الماجستير في العلوم

جامعة الملك فهد للبترول والمعادن، الظهران

CHAPTER 1

INTRODUCTION

1.1 Background

Speech is the most effective mean of communication among humans. The speech communication process begins with an idea, a thought, or information (message) that the speaker wants to deliver to the listener. This message is generated as a sequence of basic speech sounds called phonetics. The phonetics are encoded by language rules in a form of a word. Words are modulated and transmitted to the recipient who will demodulate and decode the words to get back the original message. Different perspectives can be taken when analyzing speech; these include acoustics, linguistics, and psychology, to mention a few. The specific origins of speech are unknown with a number of questions remaining to be answered or tracked, even in the twenty-first century [96]. Traditionally, speech has been seen as a combination of sentences which are composed of words, which in turns are composed of phonemes. Phonemes are the smallest units of speech and are the sounds that distinguish one word from another. For a complete speech communication system, two processes must be considered: speech production and speech perception (these are called a speech chain). Speech chains can be modeled as in Figure 1.1 [95] and are illustrated in more details in the following sections.

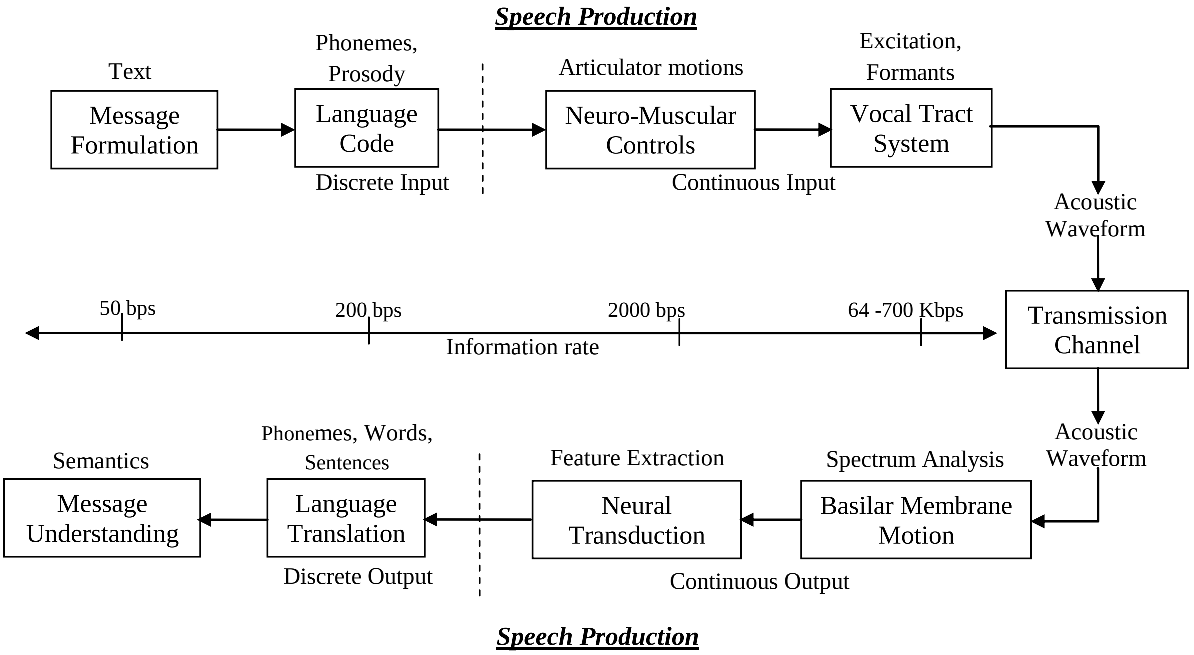

Figure 1.1: The Speech Chain: from Message Formulation to Message Understanding [95]

Speech Production

Speech production is the process by which the vocal system produces the sound which is then encoded in a form of a language word. The vocal system consists of: vocal tract, larynx, and glottis, (Figure 1.2) [83]. The Vocal tract includes: laryngeal pharynx, oral pharynx, oral cavity, nasal pharynx and nasal cavity. The larynx includes the vocal cords (vocal folds), the top of cricoid cartilage, the arytenoid cartilages, and the thyroid cartilage (known as "Adam's apple"). The vocal cords are stretched between the thyroid cartilage and the arytenoid cartilages. The glottis is the area between the vocal cords and the larynx. The pharynx connects the larynx to the oral cavity. The soft palate connects and isolates the route from the nasal cavity to the pharynx. At the bottom of the pharynx, we find the epiglottis and the false vocal cords to prevent food reaching the larynx, and to isolate the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing.

The oral cavity is one of the most important components of the vocal tract. The lips control the size and shape of the mouth opening through which speech sound is radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. The air stream to the nasal cavity is controlled by the soft palate. The oral cavity can change considerably in size enabling sound pronunciation.

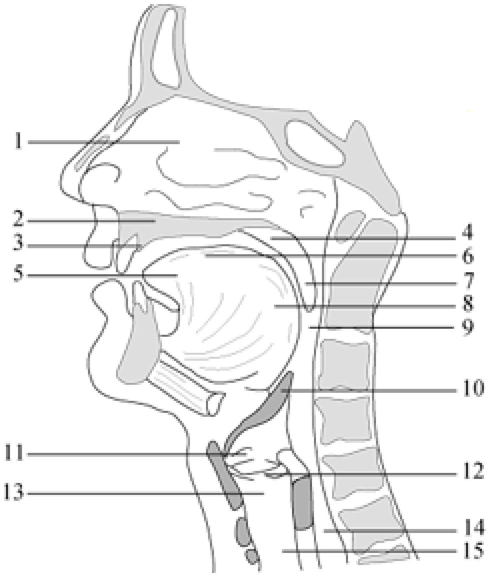

Figure 1.2: Human Vocal System [83]

(1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea.

Speech is originated by the lungs (with assistance from the diaphragm) where air flow is pressed through the glottis to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities, the air flow exits through the nose and mouth, respectively. The vocal cords module the air flow by smooth systolic and flatter while glottis rapid opening and closing generating buzzing sound which are the phonemes. Phonemes are usually considered as voiced, however they can also be unvoiced, and in some cases they are between the two states. Voiced phonemes consist of a fundamental frequency (F_0) and its harmonic components produced by the vocal cords. The vocal tract modifies this excitation signal causing the production of formants (poles) and sometimes anti-formants (zeros) frequencies.

On the other hand, with purely unvoiced sounds, there is no fundamental frequency in the excitation signal, and therefore no harmonic structure exists; hence the excitation can be seen as a white noise process. Unvoiced sounds are usually more silent and less steady than voiced ones. Phonemes are also classified as vowels and consonants. Consonants can also be classified according to the air flow blockage by vocal tract: total (voiced consonant), partial (stop consonant) or none (unvoiced consonant). Vowels (voiced) are characterized by harmonic patterns and relatively free passage of air flow through vocal tract. The fundamental frequency of vocal cords vibration depends on their mass and tension.

The most common model used for the vocal system is the single non-uniform acoustic tube (between the glottis and mouth). This model simulates vocal tract as a tube that varies as a function of time and displacement along the axis of sound propagation [1]. The non-uniform model is known as the Multitube Lossless Model because it consists of series of *N* concatenated lossless acoustic tubes; Figure 1.3 shows an example of a 7-tube lossless model. Each single tube, k , is characterized by its cross-sectional area, A_k , and its length l_k. Resonance frequency is produced when the sound wave propagates through the conjunction area between two adjacent tubes. Appropriate estimation for the number of tubes in Multitube model leads to an accurate approximation of human vocal tract [1].

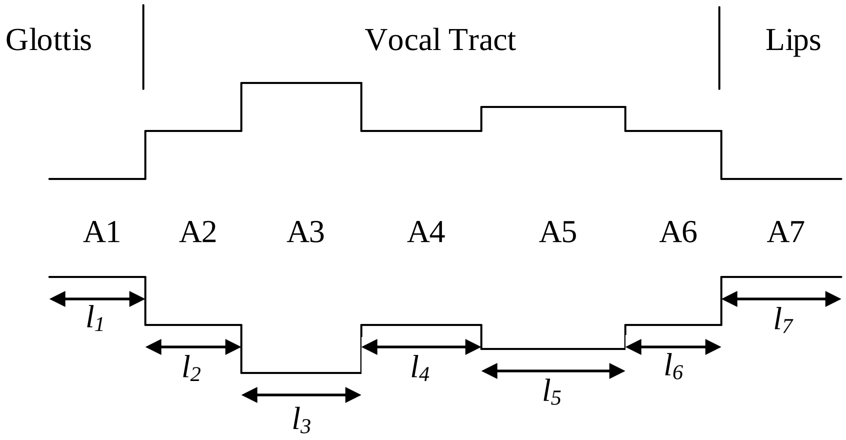

Figure 1.3: Acoustic Tube Model for the Human Vocal Tract

Acoustic Speech Variability:

Human speech sound carries not only the basic acoustic message, but it also provides some indication of the speaker's main characteristics. Several speech studies showed that speaker characteristics can significantly affect the sound produced. Examples of speaker's characteristics include:

A. Emotional state:

Emotional state of the speaker may affect the acoustic content of the utterances. Speaker's emotion such as anger, fear, contentment, happiness and love could all be communicated within the acoustic speech. Generally, negative emotions are perceived and communicated more accurately from the speech than the positive ones.

B. Physical state:

Poor health, tiredness and physical exertion can all cause variability in the speech signal. Speakers in good physical condition “produced vowel phonation for maximum duration with significantly less jitter and shimmer and had larger phonation range than did subjects of similar chronological ages in poor physical condition” [45].

C. Gender:

Previous research work has shown difference in speech signals resulting from male and female speakers especially in the fundamental frequency [73]. The fundamental frequency of men, women, and children are estimated to be around 110 Hz, 200 Hz, and 300 Hz, respectively. A scaling factor of 1.6 based on membranous vocal cord length was found to account for the difference in the fundamental frequency between male and female.

D. Age:

Age related changes in the physiology of vocal tract can have significant impact on the speech produced. Studies found that utterances from older speakers contain significantly more shimmer in the vowel duration than younger speakers. Also the vowel articulation becomes more central in elderly speakers while the formants frequency is higher for children than for adults. Models trained with speech from older people become more and more unstable for the recognition of new utterances from speakers as time passes [46]. In most speech recognition systems, the only practical way to maintain recognition reliability over time is to retrain the recognition system with new training data or to progressively retrain the system with new samples at appropriate time intervals.

E. Dialect:

Significant differences exist between speech of native and non-native speakers of a language; and between speakers of different dialects. For example, Fokes and his partners [47] found that non-native speakers of English have difficulty producing unstressed syllables appropriately, although they tend to use appropriate acoustic parameters for these.

In a comparison study of three different Arabic dialects, Yeou, and his partners [48] found variation in acoustic correlates, like F_0 peak alignment, F_0 contour shape and vowel duration [48].

F. Other Factors:

There is also considerable speaker specific information present in speech signals. This information can be useful for speaker recognition, but can also cause difficulty for speech recognition systems.

Along with these speaker specific characteristics, noise can be a major factor of acoustic variability. Noise can by due to surrounding environment or an imperfect channel for transmission such as low microphone quality. To reduce noise effect, it is common to attempt to control the environment where the recognition system is used. Features extracted from the speech signal may also be selected based on their robustness against noise.

Speech perception

The auditory system is the sensory system that allows humans to hear and understand sounds. The auditory system enables human also to recognize dangers that can't be seen. The ear (Figure 1.4) represents the main auditory system for humans. The ear can hear sounds ranging from 20 Hz to 20 kHz. It is most sensitive to frequencies between 500 Hz and 4 kHz, which corresponds almost to the speech frequency band using the telephone based system. There are three main sections in the ear: outer ear, middle ear, and inner ear. The outer ear is the part that we can see. The outer ear protects the inner and middle ears and it composed of the Pinna, auditory canal and ear drum. The Pinna is positioned on each side of the head to make help with the directional identification of sounds by differences of intensity, and time of arrival. The auditory canal behaves as a quarter wavelength resonator at approximately 3 kHz. This is the reason why the human ear has its maximum sensitivity around this frequency. The ear drum is a stiff conical diaphragm. The ear drum is seen as the boundary between the auditory canal and the middle ear. The middle ear is located between the outer and the inner ear and it composed of three bones called ossicles. These are: the malleus (also called the hammer), the incus (also called the anvil), and the stapes (also called the stirrup). The malleus is attached to the inner layer of the ear drum. The incus connects the malleus to the stapes. The stapes has a footplate and a superstructure. Its footplate is seated in the oval window, which is the separation between the middle ear and the inner ear. The eustachian tube connects the middle ear to the back of the throat. It is normally closed but opens when swallowing or coughing to equalize pressure between the middle ear and the ambient pressure in the throat. The inner ear is composed of cochlea, auditory nerve and the three semi-circulars canals. The cochlea is filled with liquid and contains the hairy sensory cells. The three semi-circulars are arranged in planes orthogonal to each another. They have no auditory functions but help in maintaining balance.

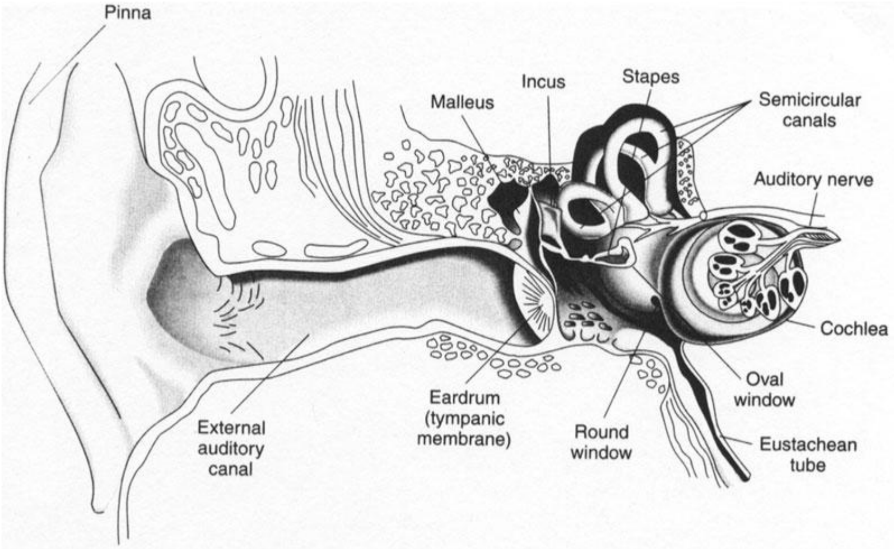

Figure 1.4: The Human Auditory System [www.skidmore.edu].

The funnel-shaped pinna collects sounds. These are transmitted along the auditory cane; and set the ear drum into vibrations. At the ear drum, sound energy, which is in fact air pressure fluctuation, is transformed into mechanical energy of ear drum movement. As the ear drum vibrates, so does the malleus. The vibrations are transmitted inwards from the malleus to the incus, and to the stapes. The vibrations of the stapes make the fluids inside the inner ear to vibrate. The vibration of the stapes footplate sets up a travelling wave pattern within the cochlea. This wavelike pattern causes a shearing of the hairy sensory cells of the cochlea, resulting in the generation of neural impulses which are then sent through the auditory nerve to higher brain centers. The brain then interprets neural impulses as sounds.

In summary, the overall auditory system can be schematically modeled as shown in Figure 1.5.

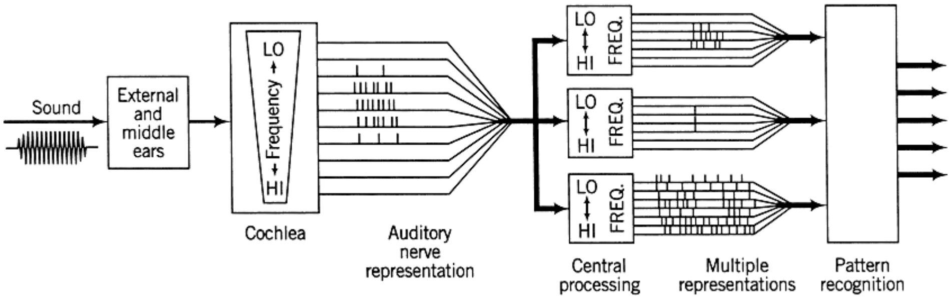

Figure 1.5: Schematic Model of Human Auditory System [95].

1.2 Overview of the Arabic Language

“Arabic is a Semitic language, and it is one of the oldest languages in the world” [23]. Arabic has 34 basic phonemes, six are vowels, and 28 are consonants [23]. There are three short vowels and the other three are long. The short vowels are not part of Arabic alphabet; they are represented as marks over or below the consonant or the long vowel. The Arabic short vowels are called “Harakats”, and they are shown in Table 1.1. The long vowels are part of Arabic alphabet. The long vowel is about twice the length of a short one and it is pronounced as a stress on the short one. The Arabic long vowels are called “Al-Mad Letters”, these are also shown in Table 1.1. The 28 consonants used in Arabic are the letters: (see Table 1.1). A special case of the consonant when it becomes a vowel is called “Shaddah (ّ)” which is a stress sound to pronounce a repeated consonant, Table 1.1.

Table 1.1: Arabic Vowels and Consonants

Arabic Short Vowels							
Vowel	Name	Vowel	Name	Vowel	Name	Vowel	Name
◌َ	Fathah	◌ُ	Dammah	◌ِ	Kasrah	◌ّ	Shaddah

Arabic Long Vowels					
Vowel	Pronunciation	Vowel	Pronunciation	Vowel	Pronunciation
ا	Aa	و	oo OR uu	ي	ee OR uu

Arabic Consonants					
Consonants	Name	Pronunciation	Consonants	Name	Pronunciation
ا	ʾalif	ʾ(a)	ض	Ḍād	ḍ
ب	Bāʾ	B	ط	Ṭā'	ṭ
ت	Tāʾ	T	ظ	Ẓā'	ẓ
ث	Thāʾ	Th	ع	ʿayn	ʿ
ج	Jīm	J	غ	Ghayn	Gh
ح	Ḥā'	ḥ	ف	Fāʾ	f
خ	Khāʾ	Kh	ق	Qāf	q
د	Dāl	D	ك	Kāf	k
ذ	Dhāl	Dh	ل	Lām	l
ر	Rāʾ	R	م	Mīm	m
ز	Zāy	Z	ن	Nūn	n
س	Sīn	S	هـ	Hāʾ	h
ش	Shīn	Sh	و	Wāw	w
ص	Ṣād	ṣ	ي	Yāʾ	y

“In Arabic, vowels cannot be initials as they can only occur between two consonants or at the end of word; [while] consonants can occur as initials, intervocalics, or syllable closings. Intervocalic and initial consonants have durations which are about half those of syllable closing or syllable suffix consonants” [14]. “The allowed syllables in Arabic are: CV, CVC, and CVCC where V indicates a vowel and C indicates a consonant. … All Arabic syllables must contain at least one vowel. Arabic syllables can be classified as short or long. … The CV type is a short one while all others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant" [23]. Table 1.2 shows syllabus representation of the Arabic digits as a typical example.

Arabic has a special characteristic for its words by using the "harakates". With "Al-harakates", the Arabic spoken word is cited differently than the written one (the same combination of letters pronounced differently with different harakates) [37]. "Arabic has a rich [and productive] morphology characterized by high degree of affixation and interspersed vowel patterns in the word root" [27]. Arabic is a rich language of dialects, of which there are four main types [27]:

1. Gulf, which includes Gulf (Gulf coast from Kuwait to Oman) and Iraqi dialects.
2. Levantine, which includes Shami (Syrian, Lebanese, Palestinian, and Jordanian) dialects.
3. North African, which includes Maghreb (Tunisian, Algerian, Moroccan, Libyan) and Hassaniiya (Mauritania) dialects.
4. Standalone, which include Hijazi, Najdi, Yemeni, Egyptian and Sudanese dialects.

Researchers working in speech recognition agree that the Arabic language is difficult because of these different dialects. It is important for Arabic ASR system to be capable of accurately recognizing the word/words irrespective of its dialect.

Table 1.2: Syllables of Arabic Digits

Digit	Arabic Writing	Pronunciation	Syllables	No. of Syllables
1	واحد	WA HID	CV-CVC	2
2	اثنين	!ITH NIN	CVC-CVCC	2
3	ثلاثة	THA LA THAH	CV-CV-CVC	3
4	أربعة	AR BA !AH	CVC-CV-CVC	3
5	خمسة	KHAM SAH	CVC-CVC	3
6	ستة	SIT TAH	CVC-CVC	3
7	سبعة	SAB !AH	CVC-CVC	3
8	ثمانية	THA MA NE YAH	CV-CV-CV-CVC	4
9	تسعة	TES !AH	CVC-CVC	2
0	صفر	SEFR	CVCC	1

1.3 Research Motivation and Objectives of The Thesis

Speech recognition systems attempt to simulate the human auditory system to deduce spoken words. A typical speech recognition system detects the sound signal and decodes it into the original word/words. Speech recognition technology also known as Automatic Speech Recognition (ASR) has gradually evolved from limited vocabulary speaker dependent recognition systems, to large vocabulary speaker independent recognition systems. This progression becomes clearer in the past few years with the evolution from isolated word recognition systems to continuous speech recognition system. The fast processors, the significant growth of Interactive Voice Response (IVR) technology, and the development of advanced algorithms; all support current research in speech recognition. As a result speech recognition has reached a significant high level of performance in terms of accuracy and speed in diverse applications.

ASR systems consist mainly of 4 major blocks: feature extraction, acoustic modeling, pronunciation and language modeling, and decision (Figure 1.6).

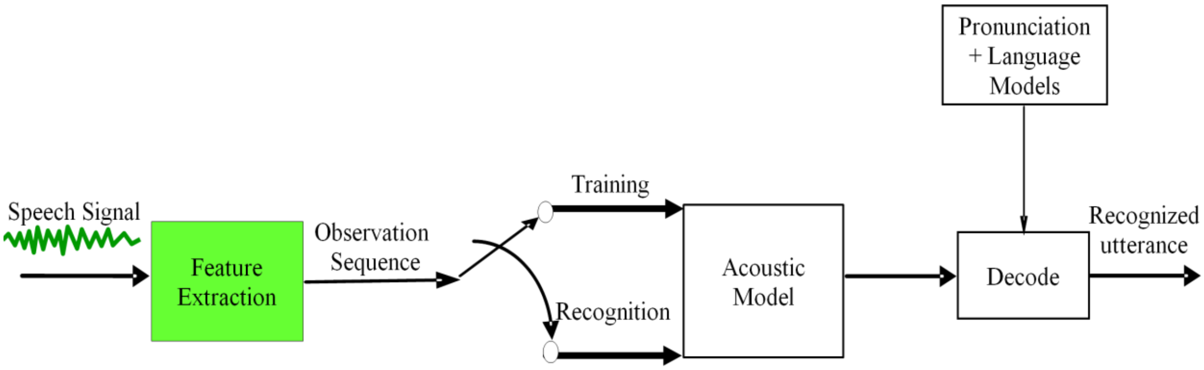

Figure 1.6: Major Components of Automatic Speech Recognition System.

Feature extraction is the first stage of the speech recognition process in both training and classification modules. The feature extraction stage converts a given input sound signal into a sequence of acoustical vectors that characterize the sound. This makes the feature extraction stage extremely important. Some considerable efforts have been carried out attempting to optimize the extraction of the most effective features to improve recognition accuracy.

Although Arabic is the sixth most widely spoken language in the world [27], “there has been relatively minimal research [carried] on Arabic compared to other languages” [23]. Arabic; my native language; is a native language of 206 million speakers and 24 million people as their second language [110]. Arabic is also the formal language of Islam all over the world. Continuous efforts and growing demand require the development of reliable Arabic speech recognition systems. Reliable systems are these that perform very well in different environments and produce accurate recognition. A number of attempts have been made to build such reliable systems for Quranic Verse recitation and delimitation [37, 38, 39, 40 and 41].

Our approach in selecting the optimal features in speech recognition is to formulate the problem as an optimization framework over the most relevant features for speech recognition applications. Such features are obtained from a large pool of traditional features.

In particular, the objectives of this research are:

1. To introduce a new approach for the optimal selection of features in speech recognition applications using the concept of mutual information.
2. To compare the performance of the optimal feature set using (1) with different standard feature sets traditionally used in speech recognition applications.
3. To apply the above model for Arabic speech recognition and compare the results to those obtained for the English language.

1.4 Thesis Organization

This thesis is organized as follows. Chapter 2 gives an overview of speech processing focusing on speech recognition. Chapter 3 describes the mutual information concept and introduces the mRMR algorithm for optimization. Chapter 4 illustrates the experiments carried in this work and the different results obtained. Chapter 5 provides a conclusion followed by some future research directions.

CHAPTER 2

OVERVIEW OF SPEECH RECOGNITION TECHNIQUES

2.1 Introduction

Speech Processing is usually used to mean analysis of human speech signals for a given application of interest.

Speech processing covers a wide range of applications. Speech processing can be subdivided into three main areas of applications (Figure 2.1); namely:

- Recognition; which covers applications that aim at recognizing or even understanding speech. Speech recognition applications include:
  - ➢ Systems for identifying input speech to extract message content.
  - ➢ Systems for identifying the speaker using voice input.
  - ➢ Systems for verifying the speaker claimed identity.
  - ➢ Automatic Language Translation used for online translation of words/sentences between languages.
- Analysis and Synthesis; this area covers techniques used to process speech waveforms to extract original or important information from the data. Examples include:

  - ➢ Systems for enhancing signal quality by reducing/removing the effects of noise.
  - ➢ Word Spotting applications such as word searching.
  - ➢ Systems for producing natural-sounding synthetic speech also known as Text To Speech (TTS) systems. TTS has some modern applications like: reading emails in unified messaging services and reading GPS information and providing instruction in automobiles.

- Coding; this area covers the techniques for representing speech at lower dimensions while preserving information content. Speech coding is used in storage and transmission. Examples of speech coding applications include:
  - ➢ Voice Communication which enables communication between people including: narrowband/broadband wired communication e.g. ADSL and VoIP and narrowband/broadband wireless communication e.g. GSM and CDMA.
  - ➢ Speech Storage which is used when speech is required to be efficiently stored for services or security purposes. Interactive Voice Response (IVR) systems are a common application of such systems.

2.2 Speech Recognition

2.2.1 <u>Historical Background</u>

For humans, speech is the most natural and effective way of communication. As can be seen from Figure 2.1, speech recognition is an important branch of speech processing.

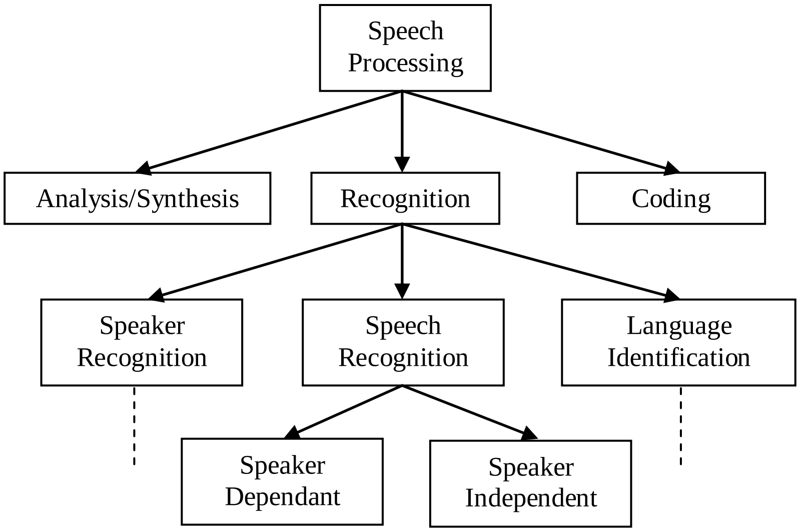

Figure 2.1: Speech Processing Hierarchy

Speech Recognition (SR) technology enables the identification of spoken words using PC-based systems. The goal of research in SR technology is to develop systems that can receive and analyze spoken information accurately and efficiently and act appropriately based on the extracted information. Accurately and efficiently means here that the SR process should be independent of the device used (i.e., the transducer, telephone or microphone), the speaker's accent, or the acoustic environment where the speaker is located (e.g. quiet, noisy, indoors, outdoors, standing or moving) [95]. Acting appropriately depends primarily on the accuracy of the recognized message. So the ultimate goal of SR, which has not yet been achieved yet, is to perform as well as a human licenser (hear, understand and act "and sometimes speak with help of Text To Speech technology") [95]. Nowadays, the commonly used speech recognition systems fall into one of the following three categories [1]:

1. Small vocabulary (~10 – 100 words).
2. Isolated words (> 10.000 words).

3. Continuous speech with constrained “task domain” e.g. business words and directory names (~1000 – 5000 words).

SR can be classified based on the speaking methods: Isolated Word Recognition (IWR), Connected Word Recognition (CWR) and Continuous Speech Recognition (CSR). SR can also be defined as Speaker Dependant (SD) or Speaker Independent (SI) based on the constraints of recognizing the speakers involved in the training phase. SR can also be classified from an application point of view as an artificial syntax system which is domain-specific or as a natural language processing system which is language-specific.

The first success story in speech recognition was a sound-activated toy dog named “Radio Rex”, this effort was not a real engineering system, but was an excellent initiative. The first speech recognizer appeared in 1952 and was a device for isolated digit recognition for a single speaker [98]. Another early device was the IBM Shoebox, exhibited at the 1964 New York World's Fair [1]. These early systems were based on spectral resonance extracted by an analog filter bank and a logic circuit. By the mid 1970’s, the basic ideas of applying fundamental pattern recognition technology to speech recognition were developed [98]. The first pattern recognition technology was based on Linear Predictive Coding and Dynamic Time Warping (DTW) matching methods. In 1970s, commercial systems became available. These speaker dependent systems were designed to recognize small vocabulary discrete utterances (words) in relatively noise-free environment [1]. The 1980s witnessed a major evolution in SR life cycle by moving from the more intuitive template-based approach towards a more rigorous statistical modeling framework [98]. The introduction of the Hidden Markov Model (HMM) in 1985 as a speech modeling technique gave a major boost to SR. By end of 1980s, IBM; using HMMs; developed an experimental system capable of recognizing 20,000 isolated words or naturally spoken utterances drawn from a 5,000- word vocabulary. In the late 1980's, Artificial Neural Networks (ANN) technology was also introduced as another statistical modeling technique.

In the 1990's, the methods for stochastic language understanding and modeling, as well as statistical learning of acoustics were the key technologies enabling the building of large vocabulary systems with unconstrained language models, and constrained task syntax models for continuous speech recognition. In late 1990's, real speech-enabled applications were developed. AT&T's developed the "Voice Recognition Call Processing (VRCP)" which was an automated handling of operator-assisted calls. AT&Ts also developed the "How May I Help You? (HMIHY)" system to which was an automatic flight information system for all customers help line calls [98]. In 2000s, SR has progressed towards multi-modal systems i.e. audio-visual speech recognition systems and even multi-languages systems.

Main Applications of Speech Recognition:

Speech Recognition is widely used in many different areas with diverse applications (services), some of these include:

- Telecom Services, for example:
    - Voice Activated Dialer where the caller can request his mobile phone with an activated voice recognition service to dial a number using voice commands.

    - Voice Enabled Unified Messaging where the caller can access and handle his email, voice mail and fax messages over the phone using voice commands.
- Computing Services when a person can interface with his computer’s functions and applications using voice commands instead of mouse and keyboard, e.g. Windows Speech Recognition.
- Call Center Services, for example:
    - Airline Services where the customer can inquire about flight information, reservation information, booking information or check-in through speech recognition services and without help desk agents.
    - Automatic Directory Assistance where one can speak the name of the entity or the corporate to the system which will respond with the contact number without any human interaction.
- Financial Services that allow customers to inquire about their banking accounts and making different transactions using voice commands.
- Medical Services, for example:
    - Medical Transcriptionist (MT) systems where one can talk to the deaf through a speech recognition system which transcripts the words into text.
    - Electronic Medical Records (EMR) where the doctor dictates the medical diagnosis, illness description, and medicine prescription, which is updated in the patient record.

  - Disabilities Services that enable disabled people to use their voice to perform activities that they cannot do them with their hands.
- Education Services where the student is taught or trained through speech recognition systems e.g.: recitation the Holy Quran.
- Military and Civil Aviation Services, for example:
  - Airplane Control Systems where the pilot can interact with the airplane control system through speech recognition systems for different instructions.
  - Aviation Training where trainees can talk to speech recognition systems that simulates different functions of the land controller.
- Traffic Facility where car drivers can speak to the Global Positioning Systems (GPS) to retrieve information on trip, map, etc...
- General Services, for example:
  - Transcription and Automatic Translation Services that recognize speech and convert it to text or to another language.
  - Robotics with speech recognition enabled systems that allow a person to interact with robots and give instructions.

In general, we can say that speech recognition applications are only limited by imagination. We will see more and more of such applications at home and at work. Open source speech recognition systems are also available, including HTK, ISIP, AVCSR and CMU Sphinx-4 systems [3].

2.2.2 Speech Recognition Models

Recognition of speech involves identifying the unknown input utterance by utilizing a reference model obtained from the training phase. There are two classes of speech recognition models: Template matching models which use a pattern matching algorithm, and Stochastic models such as the Maximum Likelihood (ML) algorithm [1]

A. Template/Pattern Matching Models (deterministic models):

The basic idea here is to use certain distance measures to compute pattern similarities [3]. This requires aligning temporally the features of the test utterances with those of the reference utterances before computing the matching score. The major drawback of this approach is that pattern similarity requires time alignment. Time alignment and distance computation are often performed simultaneously. Dynamic Time Warping (DTW) is the popular technique used for time alignment. DTW temporally wraps (stretch or compress in time) the frames of a test pattern to fit the corresponding frames of the reference pattern and use dynamic programming to accumulated distance between these. DWT has been successfully employed in simple applications [1].

Let the test pattern be denoted as *T,* and the reference pattern denoted as *R.* The warping distance *D(T,R)* between the test pattern and the reference pattern is the sum of the local distances between each correspondence frame assuming both patterns have the same number of frames [2] and it is given by:

$$D(T,R)=\left[\sum_{n=1}^{N} d\big(T(n),R(n)\big)\right], \qquad □(2.1)$$

Where *d(T(n),R(n))* is the local distance between frame *n* of *T* and frame *n* of *R*. The Euclidean distance and LPC log likelihood distance have been shown to be reasonable distance measures for this problem [36]. DWT can also be used with the feature extracted from the frames such as LPC. It is worth noting that there are several approaches that can be used for matching patterns including the nearest neighbor rule or the K-nearest neighbor rule, among others.

B. Stochastic Models:

Template Matching Models exhibit some limitations that restrict their use to relatively small to medium vocabulary databases and speaker-independent applications. For large and complicated applications where the cost of variability, storage and computation is high, Stochastic Models are more appropriate. The term stochastic indicates that some variable characteristics of the speech are used. There are two commonly used models: the Hidden Markov Model (HMM), and Artificial Neural Networks (ANN). HMM has been the basis for successful large-scale laboratory and commercial speech recognition systems, while ANNs have been a part of much more general research efforts to explore alternative computing architectures mimicking biological neural systems [1]. ANNs give computing an amazing capability in applications that usually require human observation and thought process. ANNs are built from a large number of neural cells, each representing a simple processor that deals with part of the problem. The concept of ANNs is to distribute the computation load among the cells [37]. An input is propagated through the ANNs network, and then the output is compared with a desired reference to produce an error. The load weights of the cells are adjusted to minimize the error until an optimal solution is reached based on some performance criterion. The main disadvantage of ANNs is that the training is a very long process. On the other hand, HMM converges faster and it is a more preferred stochastic modeling technique used in speech recognition. The power of HMM resides in the feature that it is able to model time-varying patterns and track the changes, hence, it is appropriate in modeling the tracking of speech features with time. HMM is discussed in details in section 4.3. The Stochastic approach generally uses an ML criterion to decide on the reference utterance that most likely matches the test utterance.

Let O represents the observation sequence (feature vector) from the test utterance, and let M_i represent the different model of the possible utterances. The ML formulation is written: select utterance u such that:

$$u = \max_{i} \left[P\left(O | M_{i} \right) \right] \qquad \square(2.2)$$

where $P(O|\ M_i\ )$ is the probability of the observation O given a model M_i, (i=1, 2, …N) with N being the number of possible utterances.

2.2.3 Automatic Speech Recognition (ASR) Systems

Automatic speech recognition systems consist generally of 2 stages: training and recognition. The training phase is the process of learning from known speech units with the aim of building reference models. This simulates the teaching of a kid a given word until he memorizes it and stores it in his brain. The goal of training is to develop a dictionary of speech units. ASR systems operate by matching the index of input speech signals with the ones of known speech units from a dictionary. The training phase, in general, consists of three stages, these are:

1. Building the database and codebook.
2. Building the classification model.
3. Validation.

The recognition phase is the process of identifying the spoken unit based on the reference models. This involves a classification step followed by a decision. Pattern classification basically measures the similarity between the input utterance pattern and each reference pattern with either of the classification models discussed earlier. The decision step identifies the unknown utterance based on algorithm’s result which depends on similarity measurements of classification. The recognition phase simulates a word recognized by the human brain based on a stored referenced model for that word built earlier.

The recognition phase consists of two stages namely:

1. Classifying the sequence of symbols.
2. Declaring the utterances with the highest probability.

Figure 2.2 shows a block diagram for a typical ASR system.

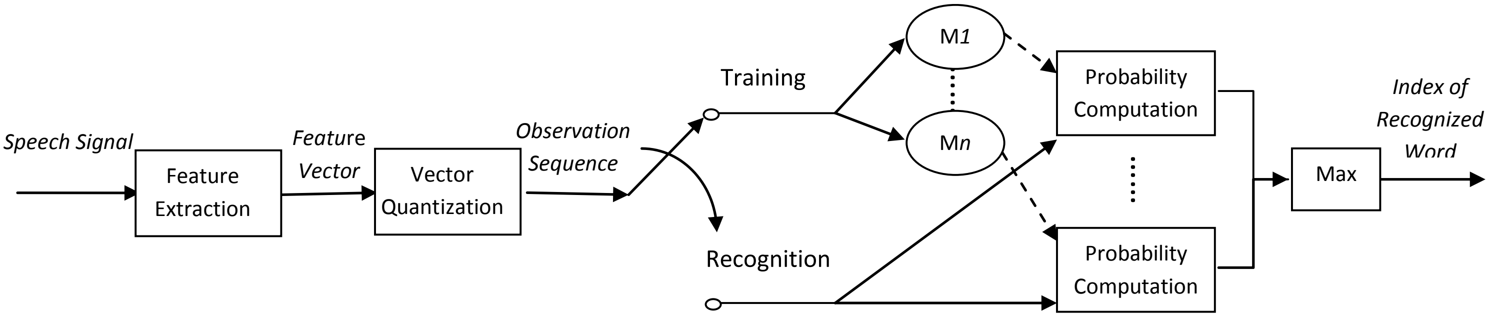

Figure 2.2: Block Diagram for Automatic Speech Recognition.

Before the training and the recognition phases, the utterances are preprocessed and analyzed to extract the different sets of feature vectors. The utterances (acoustic unit) can be a word or sub-word, such as phonemes, diphones (a phoneme with its neighbor), triphones (a phoneme with its left and right neighbors) or syllables [28]. The feature vectors are seen as the raw data for both training and recognition. The process of determining the feature vectors is known as the feature extraction stage. This stage is very important; hence it will be discussed in more details in the next section.

2.3 Feature Extraction for Speech Recognition Applications

Feature extraction is a key step in developing robust speech recognition systems. This stage is used to convert speech utterances into sets of small size characterizing features. The estimation of the time varying spectrum is usually the first step in most feature extraction methods [8]. This is carried out by pre-emphasizing speech data and then segmenting it into frames. Each frame (usually 25 msec) is windowed to smooth out the frame ends. The different features are then extracted. In the training stage, segments of feature vectors are stored as pattern for different utterances or words. These vectors may be used to develop models for the different utterances. In the recognition phase (or testing), the estimated feature vectors from the unknown utterances are mapped to the best matching pattern using probability (or distance). Figure 2.3 presents the basic block diagram for the feature extraction stage.

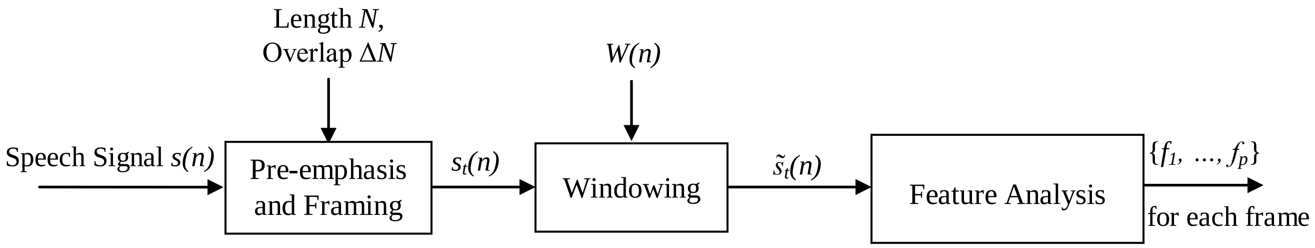

Figure 2.3: Block Diagram for Feature Extraction

Note that the feature extraction stage is preceded with two important stages: segmentation and windowing. These are briefly explained before discussing in more details the different features commonly used in speech analysis.

Speech Preemphasis:

Spectral tilt is one characteristic of human speech spectrum. Spectral tilt represents the spectrum of the voiced signal which has more energy at lower frequencies than higher. To overcome this unbalanced energy distribution, a first order preemphasis filter $(1 - \alpha z^{-1})$ is used to boost high frequency energy. Typical pre-emphasis filter coefficient is between 0.95 - 0.97.

Speech Segmentation:

Human speech frequency characteristics change with time, so speech is said to be nonstationarity. To reflect this nonstationarity, we need to recalculate the feature coefficients margin at regular time intervals. To perform short time analysis of segments over the whole speech, the concept of sliding windows is used. Previous research has shown that speech can be stationary over short time intervals of 10-30 ms. Based on this assumption, the window sliding approach is used with short overlap between consecutive frames. Typical frame duration is 20-30 ms and overlap duration of 10-20 ms.

Windowing (Frames):

Large prediction errors may result at frame edges because of prediction length (small number of non-zero samples). To reduce this error, speech frames are first multiplied by soft window functions to smooth out the frame edges. A Soft window has slow truncations at both edges, so multiplying it by the speech frame will smoothly taper the frame ends. The Hamming window is the most commonly used window with 99% of its energy residing in the main lobe and the highest side lobe is at -43 dB [4]. The Hamming window is defined as [1]:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \qquad \square(2.3)$$

where N = the length of the window.

Other reasons for multiplying frames with windows include spectral leakage [4]. Simple segmentation can be obtained using a rectangular window which means in frequency domain a convolution of the signal spectrum with the frequency response of rectangle window. The rectangle window response has a very narrow main lobe and relatively large side lobes. This causes a corruption of frequency components in the region of the side lobes. On the other hand, Hamming window frequency response has a wider main lobe and smaller side lobes (Figure 2.4). The spectral leakage in Hamming window case is reduced because of its small side lobes [1].

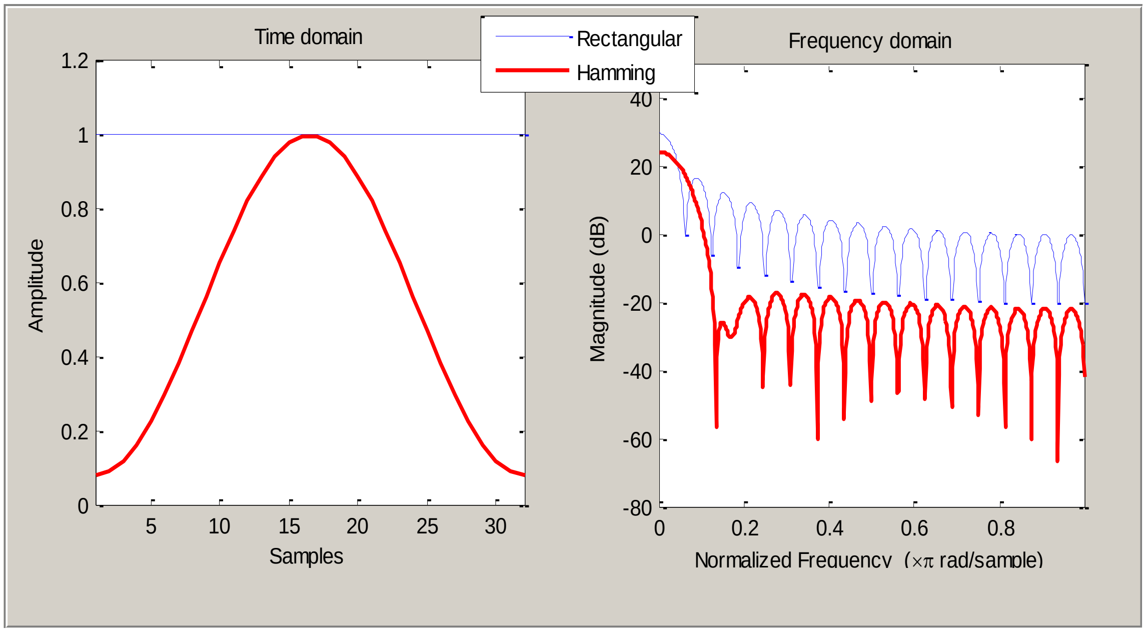

Figure: 2.4: Rectangular and Hamming Filters of Length 32.

Feature Analysis:

After segmenting speech into frames, we are now ready to analyze the frames to extract the different characteristic features. Traditionally, there have been 2 approaches used for speech parameterization. In particular, these two approaches are: parametric and nonparametric. Parametric methods are usually used for time domain analysis, and represent the resonant structure of the human vocal tract (human production system). Whereas, nonparametric methods are implemented in frequency domain and it is used in representing the human auditory perception system [22]. The most popular features are the Linear Prediction coefficients (LPC) and the Cepstral Coefficients. Broadly speaking, features can be classified into Auditory based or Articulatory based features:

2.3.1 Auditory Based Features

Auditory features are mainly based on models of the human auditory system. The Human auditory system exhibits a number of limitations including a nonlinear frequency scale, spectral amplitude compression, and decreasing hearing sensitivity at lower frequencies. These constrains are taken into account in the extraction of the auditory based features. We will briefly describe in the next paragraphs the most popular auditory based features used in speech recognition.

Linear Prediction Coefficients (LPC):

Linear Prediction Coefficients (LPCs) were proposed by Atal and Hanauer as dominant speech features [39]. LPCs are based on the modeling of the acoustic parameters in representing a given sample of speech from previous samples. The LPC model is an all-pole system and represented by:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} \alpha_k z^{-k}} = \frac{G}{A(z)} \qquad (2.4)$$

where:

α_k 's are the LPC coefficients (features),

G is the Gain, and

p is the LPC order.

By evaluating the parameters $\alpha_k\,'s$; a given speech sample can be approximated as:

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (2.5)$$

The error $e(n)$ between $s(n)$ and $\tilde{s}(n)$ is defined as:

$$\begin{aligned} e(n) &= s(n) - \tilde{s}(n) \\ &= s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \end{aligned} \tag{2.6}$$

The residual signal energy is given by:

$$\begin{aligned} E &= \sum_{k=-\infty}^{\infty} e^2(n) \\ &= \sum_{k=-\infty}^{\infty} [s(n) - \tilde{s}(n)]^2 \end{aligned} \tag{2.7}$$

The LPC coefficients are estimated by minimizing the residual energy E over the parameters. Such optimization results in solving the famous Yale-Walker equations·

LPC is a powerful model in estimating the acoustic parameters and representing speech using a smooth spectrum. Enhanced forms of LPC can be obtained from the original LPC; these include: the Line Spectrum Pairs (LSP), the Reflection Coefficients (RC), and the Log Area Ratio Parameters (LAR).

Line Spectrum Pairs (LSP):

The LSP model is another representation of LPC. The LSP model represents the predictors in the Z-domain in which the zeros of *A(z)* are mapped into the unit circle through a pair of polynomials. The LSP polynomials are represented by:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned}$$

So that

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{2.8}$$

Since the zeros of the polynomials *P(z)* and *Q(z)* are complex conjugate pairs, there are only p real parameters (frequency or phase) which are required to define the LSP parameters (since the magnitude is unity). LSP can be interpreted in term of format frequencies of the model. Each zero of *A(z)* maps into one zero in each of the polynomials *P(z)* and *Q(z.)* If these 2 zeros are close in frequency, their parent zero in *A(z)* represents a formant (narrow bandwidth) in the model. Otherwise, the original zero represents a wide bandwidth spectral component [1].

Reflection Coefficients (RC):

The RC model is based on wave theory and the acoustic tube model of the vocal tract where at each boundary of the model a portion of the sound wave is transmitted while the remainder is reflected. RC coefficients are obtained from LPC coefficients by using the following backward recursion:

$$\begin{aligned}
&\alpha_j^{(p)} = \alpha_j, && 1 \le j \le p \\
&k_i = \alpha_i^{(i)} && \\
&\alpha_j^{(i-1)} = \frac{\alpha_j^{(i)} - \alpha_i^{(i)}\alpha_{i-j}^{(i)}}{1-k_i^2}, && \begin{cases} 1 \le j \le i-1 \\ i = p, p-1, ..., 1 \\ k_i \ne \pm 1 \end{cases}
\end{aligned} \qquad (2.9)$$

where:

a_{j}'s are the LPC coefficients, and

k_{i}'s are the reflection coefficients.

Log Area Ratios (LAR):

The LAR coefficients exhibit similar advantages to the RC coefficients. The LAR model is also based on the acoustic tube model of the vocal tract but uses the ratios of cross-sectional areas of adjacent sections. The LAR coefficients are derived from the RC coefficients by the following non-linear transformation:

$$LAR_i = \frac{1}{2}\log\left(\frac{1+k_i}{1-k_i}\right) = \tanh^{-1}(k_i), \qquad \begin{cases} i = 1, 2, \ldots.., p \\ k_i < 1 \end{cases} \qquad (2.10)$$

where:

LAR

LAR_i's are the LAR coefficients, and

k_i's are the reflection coefficients.

Filter Bank Features (FB):

The Filter bank used for speech is designed to match the human ear spectral characteristics. Each of the filters used is called an auditory filter representing a resonance frequency of the human speech signal. Each auditory filter is centered at the tone frequency, and the bandwidth (critical band) is seen as the minimum bandwidth of the filter that blocks adjacent resonance frequency from interfering with it.

To calculate the power of each tone in an auditory band, we will assume that the input noise is a wide-sense stationary (WSS) process and the auditory filter is a symmetric, linear, time-invariant system. Based on that, the power spectrum of tone for each filter is the power spectrum of the input noise multiplied by the transfer function squared of the filter. Then the total power is calculated as:

$$P_i = K_i \int_0^\infty N_i(f)\left|H_i(f)\right|^2 df, \qquad i = 1, 2, \ldots\ldots \qquad (2.11)$$

where:

K_i is a constant,

$N_i\,(f)$ is the power spectral of the noise in a given frequency band, and

H_i (f) is the transfer function of the auditory filter at the same frequency band.

Cepstrum Coefficients (CC):

The Cepstrum Coefficients (CCs) were introduced as speech features by Oppenheim [39]. The CC model is based on the linear separation of the speech generator spectrum (Excitation *E(z)*) from (the vocal tract filter *H(z)*) [32]. Since the speech spectrum (*X(z)*) is represented by:

$$X(z) = E(z)H(z) \tag{2.12}$$

By applying the log operation, the speech spectrum can be represented using a linear combination of the excitation and the filter as follows:

$$\log X(z) = \log E(z) + \log H(z) \tag{2.13}$$

The CC is defined as the inverse z-transform of the log spectrum of *X(z)*. The lower order coefficients of CC represent the vocal tract characteristics, whereas the higher order coefficients represent the excitation characteristics [32].

Mel Frequency Cepstrum Coefficient (MFCC):

The MFCC parameters of Davis and Mermelstein were introduced as one of the most dominant speech features in the literature [39]. The MFCC model is a modified form of the CC. the MFCC model provides a representation of a smooth short-term spectrum that has been compressed and equalized hence better simulates the human hearing system characteristics. The MFCC coefficients are the most popular features used in speech analysis and shown to be more robust to noise and speech distortion than any other features [36]. The MFCC coefficients are generated from the filter bank parameters using the Mel-frequency scale. The Mel-scale filter banks represent the non-linear pitch perception pattern of the human ear. To calculate the MFCCs, the Discrete Fourier Transform (DFT) is first applied to the input speech frame to obtain the magnitude spectrum. The magnitude spectrum is then frequency-warped in order to transform the spectrum into Mel-frequency. Then, each magnitude spectrum is multiplied by the corresponding filter gain to compute the energy for each filter. Finally, the Discrete Cosine Transform (DCT) of the each log energy is computed resulting into the MFCC coefficients [10].

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(j \frac{\pi}{N} \left(i - \frac{1}{2} \right) \right) \tag{2.14}$$

where:

c_i's are the MFCC coefficients,

N is the number of points used, and

m_j's are the log energies.

Delta Cepstrum (DC):

The DC is a good measurement of dynamic behavior of speech. It is based on an estimate of the local time derivative of the short time cepstrum. This is typically implemented as a least-square approximation to the local slope and so it is a smoother estimate of the local derivate [2]. The DC coefficients can be expressed as [32]:

$$\Delta C_t = \frac{\sum_{i=-n}^{n} k_i \left(C_{t+i} - C_{t-i} \right)}{\sum_{k=-n}^{n} k_i^2}, \tag{2.15}$$

Where n is the window width and k_i's are the regression coefficients. Thus, each stream of delta cepstral coefficients is computed by correlating the corresponding stream of cepstral values with a straight line that has a slope of one [2].

The second derivative, (known as **Delta-Delta Cepstrum**) is also useful as it corresponds to a similar correlation, but with a parabolic function [2]. The delta–delta parameters, $\Delta^2 C_t$'s, are derived in the same manner, but using differences of the delta parameters. These features tend to emphasize the dynamic aspect of speech spectrum over time. However, these feature vectors miss some of the coarse characteristics that are important in static spectral representation [2]. DC and D-DC are not sufficient by themselves for good recognition performance, but they are used as add-on features to static measures such as MFCC or PLP coefficients [2].

Perceptual Linear Predictive coefficients (PLP):

The PLP parameters were introduced as speech features by Hermansky [39]. The PLP model is similar to the MFCC in relation to good smooth estimation of local spectrum and for optimal modeling of the human hearing system characteristics. The main difference between PLPs and MFCCs is the nature of the cepstral smoothing [2]. PLP is viewed as a Mel-cepstral analysis with LPC-like spectral smoothing, or as an LPC analysis for an implicit version of the speech that has been warped according to auditory properties [2]. PLP is as efficient as standard LPC, and has practical advantages over some other perceptually based processing techniques.

Cepstral Mean Subtraction (CMS) coefficients:

The CMS coefficients are a special case of cepstral coefficient where the estimates of local time derivatives of cepstral parameters are more robust by suppressing the constant spectral components in the data. Consider a linear time-invariant system, let *X(ω,t)* is the short-term spectrum of the observed signal,

$$X(\omega,t) = S(\omega,t)H(\omega,t) \tag{2.16}$$

Then the corresponding short-term log power spectrum becomes:

$$\log|\,X(\omega,t)|^2 = \log|\,S(\omega,t)|^2 + \log|\,H(\omega,t)|^2 \tag{2.17}$$

For instance, if *H* is constant over time, and if the constant component of *S* is not useful, then we can estimate the constant component of the sum by computing the mean of the log spectrum. Alternatively, we can take the Fourier transform of equation 2.18 to get the cepstrum, then remove the mean. The CMS is more robust in a system that has sudden changes (disturbances) resulting from unexpected environment events such as a disturbance affecting the speech, including; a change in telephone channel, a switch in microphones, or just a turn off the speaker's headset so that the overall spectral characteristic is modified. Disturbances that are of convolution nature in the time domain become additive in the log spectral domain. If such additive components have different spectral characteristics, we can use linear filters to separate these.

Relative Spectra (RASTA):

The RASTA model is comparable to CMS, but is obtained from PLP. In RASTA, conventional critical-band short-term spectrum in PLP is replaced with a spectral estimate in which each frequency channel is filtered with a band-pass filter of sharp spectral zero at the zero frequency. With this design, any constant or slowly varying component in each channel is suppressed.

Speech-Signal-Based Frequency Cepstral Coefficients (SFCC):

The Speech-Signal-Based Frequency Cepstral Coefficients are a new features introduced by Paliwal, Shannon, Lyons and Wojcicki in April, 2009 [99]. Unlike the auditory based frequency wrapping scale (MFCC), the new feature's wrapping function is based purely on the properties of the acoustic speech signal [99]. The concept of Speech-Signal-Based frequency wrapping is based on the hypothesis that the majority of the linguistic information is carried in the high energy (formants) regions of the speech spectrum. SFCC is computed by ensemble averaging the short-time power spectrum, $\overline{P}(f)$ over the entire speech signal. Then, the frequency axis is divided into M non-overlapping intervals, such that the area under $\log \overline{P}(f)$ is equal for all intervals. Finally, the middle frequency of each interval represents the warped frequency. Mathematically, consider the short-time Fourier transform of a discrete signal *s(n)* as:

$$S(n, f) = \sum_{m=n-N+1}^{n} s(m)w(n-m)e^{-j2\pi fm/F_s} \qquad (2.18)$$

where:

w(n) is a window of length *N*, and

F_s is the sampling frequency.

With assumption that speech is a wide-sense stationary (WSS) process, the short-time power spectrum for the speech signal *s(n)* is computed as:

$$P(n,f)=\frac{1}{N}\left| S(n,f)\right|^2 \qquad (2.19)$$

An ensemble average is computed by averaging *P(n, f)* over the entire signal as:

$$\overline{P}(f)=\langle\, P(n,f)\rangle \qquad (2.20)$$

Then, the logarithm of the ensemble spectrum is divided into equal interval as [99]:

$$A_i = \int_{f_i}^{f_{i+1}} \log \overline{P}(f)\, df, \qquad i=1,2,....M \qquad (2.21)$$

$$A_i = A_{i+1}, \qquad i=1,2,....M-1 \qquad (2.22)$$

where:

A_i is the area of the i^{th} interval, and

f_i and f_{i+1} are the cutoff frequencies of the interval.

The Speech-Signal-Based frequency warping function is simply the middle frequency of the interval and it is calculated by:

$$W\left(\frac{f_i + f_{i+1}}{2}\right)=\left(\frac{i}{M}\right), \qquad i=1,2,......M \qquad (2.23)$$

The SFCCs are computed using the MFCCs (mentioned above); however, the speech-signal-based frequency warping is used for the triangular filterbank design instead of the Mel-scale [99].

Power Normalized Cepstral Coefficients (PNCC):

The Power Normalized Cepstral Coefficients are new features introduced by Kim, and Stern in 2009 and 2010 [100]. PNCC is an auditory based processing technique. The PNCC feature extraction process is comparable to MFCC except for two steps. PNCC uses power-law function instead of the log function used in MFCC. In addition, PNCC uses gammatone filter while MFCC uses the triangular filter. Also, PNCC has an additional step over MFCC that is using Power-Bias Subtraction (PBS) to suppress background noise. The computational cost of PNCC is slightly larger than that of conventional MFCC processing [100].

The power-law nonlinearity function used in PNCC is described by the equation [111]:

$$y = x^{0.1} \tag{2.24}$$

where the value of 0.1 is the best observed value which approximates the physiological rate-intensity function [100]. An attractive feature of the power-law nonlinearity over the log nonlinearity is that the dynamic behavior of the output does not depend critically on the input amplitude [100]. PBS is used in PNCC to bias the power in each of the frequency channels to maximize the sharpness of the power distribution. This procedure is motivated by the fact that the human auditory system is more sensitive to changes in power over frequency and time than to relatively constant background excitation [105].

The normalized power based on power-bias subtraction $\tilde{P}(m,l)$ is given by [105]:

$$\tilde{P}(m,l) = \left(\frac{1}{l_2 - l_1 + 1} \sum_{l'=l_1}^{l_2} \frac{\tilde{Q}(m,l')}{Q(m,l')} \right) P(m,l) \tag{2.25}$$

where:

l$_1$ = *min(l−N, L)*, *l*$_2$ = *max(l+N, 1)*, *L* is the total number of channels and *N* the weight smoothing factor, and

$\tilde{Q}(m,l)$ is the normalized power given by:

$$\tilde{Q}(m,l) = \max\left\{Q(m,l) - q_0,\ q_f\right\} \tag{2.26}$$

where p_0 is the peak power value after normalization, q_f provides power flooring.

Q(m,l) is the medium-duration power, which is the running average of the short-time power *P(m,l)* and it is given below:

$$Q(m,l) = \frac{1}{2M+1}\sum_{l'=l-M}^{l+M} P(m,l') \tag{2.27}$$

where *m* is the frame index, the *l* is the gammatone channel index and *M* = 2 based on speech recognition results obtained with different values of *M*.

P(m, l) is the normalized short-time power by the peak power and it is given by:

$$P(m,l) = P_0\,\frac{P_{org}(m,l)}{P_{peak}} \tag{2.28}$$

where $P_{org}(m, l)$ is the short-time spectral power in the m^{th} frame and the l^{th} gammatone channel, P_{peak} is the peak power that is the 95th percentile of the short-time power and P_0 is a constant value.

2.3.2 <u>Articulatory Based Features</u>

In human biology, articulators are the finer anatomical features critical to speech production and include the vocal cords, soft plate, tongue, teeth and lips [1]. Articulation is the science of how the movement of articulators produces different phonemes. Based on the biological human speech production system, a number of articulatory based features were developed. In particular, the three articulatory feature dimensions that describe consonants:

- Place of articulation which describes the position of the main constriction of the vocal tract. This dimension includes bilabial, labiodentals, dental, alveolar, postalveolar, retroflex, palatal, velar, uvular and glottal features.
- Manner of articulation which describes the degree of the constriction of the vocal tract, the position of the velum, and some other characteristic of articulators' behavior. This dimension includes aspirated, plosive, nasal, trill, flap, fricative, affricate and approximant features.
- Vibration of vocal cords which classifies consonants into either voiced or unvoiced.

For vowels, it is described by the position of the highest point of the tongue called dorsum. There are two dimensions in which dorsum's position describes vowels:

- Horizontal which includes close, close-mid, open and open-mid features.
- Vertical which includes front, central and back features.

We list in what follow the most commonly used articulatory based features:

Speech Intensity:

Intensity is the loudness of sound [14]. The intensity of a continuous sound can be calculated using the following equation:

$$I_c = 10\log\left[\frac{1}{TP_0^2}\right]\int X^2(t)dt \tag{2.29}$$

where;

P_o is the auditory threshold pressure and equal $(2.10)^2$ Pascals,

$X(t)$ is the sound pressure measured in Pascal,

I_c is the sound intensity in continuous time domain, and

T is the duration of the sound.

For a given discrete sound signal the intensity, I_d, is calculated by the same manner with the equation:

$$I_d = 10\log\left[\frac{1}{nP_0^2}\right]\sum_i x_i^2, \qquad i = 1, 2, ...n \qquad (2.30)$$

where: x_i is the instantaneous sound pressure in Pascal, and

n is the number of samples.

Formant Frequencies:

Formant frequencies are the resonant frequencies of the vocal tract and are crucial in speech production modes. “One of the most common methods used for formant analysis is Linear Predictive Coding in which formant frequencies are estimated from the spectral peaks of the speech signal” [14]. The formant transition tracks provide us with the important “hidden” information of the formant frequency trends for different phonemes.

Voicedness Feature:

Voiced sounds have a periodic structure when viewed over short time intervals and perceive of the fundamental frequency of this periodic signal is known as pitch [95].

Pitch is commonly used interchangeably with fundamental frequency [1]. Pitch is constrained by the individual’s larunx. For men, the possible pitch range is usually between 50 – 250 Hz, while for women the range is between 120 – 500 Hz [1].

Voicedness is a state measurement of the vocal cords and it describes how periodic the speech signal is in a given time period *t* (frame index) [56]. To measure the periodicity, the autocorrelation function is used. Let us define $R^t(\tau)$ as the autocorrelation function of $x^t(v)$:

$$R^t(\tau) = \frac{1}{T-\tau}\sum_{v=0}^{T-\tau-1} x^t(v)\,x^t(v+\tau) \qquad (2.31)$$

where T is the length of a time frame, τ is the time delay or lag and t is the frame index.

It is worth noting that the autocorrelation function of cyclostationary signals is periodic with maxima $R^t(0)$ reached at $\tau = kT,\ k = 0 \pm 1, \pm 2, \ldots\ldots$ Therefore, a peak in the range of possible pitches with a value close to $R^t(0)$ is a strong indication of periodicity.

To produce a bounded measure of voicedness in the interval [-1, 1], the autocorrelation function is normalized. The voicedness measure v^t; which is the maximum value of the normalized autocorrelation in the interval of natural pitch periods [2.5ms...12.5ms]; is defined as:

$$v^t = \frac{\max_{2.5ms.\,fs \le \tau \le 12.5ms.\,fs} R^t(\tau)}{R^t(0)} \qquad (2.32)$$

where f_s denotes the sample rate. When v^t close to "1", the time frame described as voicedness. And when it is close to "0", it is described as voiceless [56].

Based on the analysis of different features discussed above, we summarized the advantages and disadvantages of different features in Table 2.1.

Table 2.1: Comparison of the Most Popular Acoustical Features

Technique	**Summary**	**Advantage**	**Disadvantage**
LPC	Represents the spectral envelope of the short time spectrum of speech.	1. Good approximation for voiced speech.	Very sensitive to quantization noise.

		2. Low computation.	
LSP	Represents the resonant frequencies in speech spectrum.	1. Less sensitive to quantization noise. 2. Stable.	Resonant frequencies may provide redundant information.
RC	Represents the reflection at the boundaries of acoustical tube model of the vocal tract.	1. Good representation for the vocal tract. 2. Stable.	Highly sensitive to quantization noise if the magnitude is near unity.
LAR	Represents the area ration of acoustical tubes of the vocal tract model.	Robust to quantization noise.	Relatively high computational load.
CC	Provides a linear separation of the excitation from the vocal tract.	1. Orthogonal. 2. Improved recognition performance.	High computational load.
Delta-CC	Represents the dynamic aspect of the speech spectrum over time.	Relatively insensitive to constant spectral characteristics.	Not sufficient for good recognition.
MFCC	Represents a smoothed short-time cepstrum.	1. Robust features. 2. Best representation of all classes of speech sounds. 3. Best recognition performance.	High computational load.
Intensity	Represents sound loudness.	Low computational load.	Speaker dependant.
Formant Frequencies	Represent the peaks in the spectral envelope of the sound.	1. Good tracking of the phonemes in frequency. 2. Unaffected by changes in the voice source.	Sensitive to classification error.

2.3.3 Optimal Feature Selection for Speech Recognition

We have seen that all of the features mentioned above have their own advantages and drawbacks. For this reason, it would be desirable to find an optimized set of features based on a certain optimality criterion. Such a move from the traditional use of existing features is expected to lead to an improved recognition accuracy under different environments. There are many feature selection algorithms proposed in the literature and most of them operate based on a certain definition of relevance. The main drawback of many of these algorithms is that the non-negligible redundancy between different selected features which usually leads to low classification accuracy. The optimization algorithm that we propose here is the use of the minimum Redundancy Maximum Relevance (mRMR) criterion. This algorithm has demonstrated excellent classification results by solving the problem of redundant features. The successful implementation of mRMR in biomedical field prompts us to introduce this algorithm in feature selection for speech recognition technology aiming to improve the accuracy by reducing the irrelevant or redundant features. It is worth mentioning that this is the first attempt to use the mRMR in speech recognition. In our work, we will discuss an approach for choosing an optimal set of features from the traditional following features: LPC, LSP, RC, CC and MFCC. We will show that the resulting optimized feature set significantly enhances the recognition performance of ASR systems even at low Signal-to-Noise-Ratio (SNR). Speech recognition at low SNR is a challenging problem given difficulty in extracting good characterizing features. In our research, we focus only on low SNR region since it is obvious that improving the performance in this region will be reflected at high SNR where speech has more power than noise making the problem of speech recognition simpler.

Before discussing the mRMR algorithm, we will briefly discuss the Hidden Markov Model that we will use in training and classification.

2.4 The Hidden Markov Model in Speech Recognition

The Hidden Markov Model (HMM) is one of the most powerful modeling techniques used in speech recognition [11]. The Dragon Speech Recognition System was the first ASR application using HMM in the early 1970s [26]. HMM became a popular statistical modeling technique in ASR since the mid-1980s. The introduction of HMM prior to speech processing work was related to the work of Levinson and Poritz for their approach in modeling the problem of characterizing a random process with incomplete observations [1]. Their algorithm was known as the Estimate-Maximize (EM) algorithm. In the early 1970s, Baum and colleagues worked on the special case of HMM and they developed the EM algorithm for HMM parameter estimation and decoding. This algorithm was known as the forward-backward (F-B) algorithm (it is also called the Baum-Welch reestimation algorithm). The F-B algorithm simplified computationally intractable problem into a manageable one [1]. Basically HMM is a statistical model consisting of a Markov chain process with a set of unknown parameters, and the challenge is to estimate these (hidden) parameters from a sequence of observable parameters [23]. The Hidden Markov model can be explained clearly with a simple coin tossing experiment. Consider that we have a set of numbered coins that are tossed one at the time and at each step we record the output. In this experiment, the states sequence (coins number and the probability of changing between them) is hidden and the recorded output sequence (called observations symbols) is the only observed information. The challenge is to use the HMM to statistically estimate the state sequence from the observed output sequence. For example, if we have an output sequence: T T H, the question is: what is the probability of generating this sequence from a certain state sequence, say; state#2, then state#3 then state# 2? The HMM model helps answering this question.

The term “hidden” indicates that the state sequence of interest is not directly observed, however, it can be estimated from the outputs (observed) since the state sequence affects the generated output sequence.

Another example of HMM is speaker classification. Imagine that there is group of people behind a wall and they are a random mix of personnel (doctors, students, engineers, researchers and farmers). We cannot see the people behind, but we can hear them. We ask each speaker to utter a deep sigh and since the speakers are different so their acoustics sigh are also different. The goal of this experiment is to assign the speakers from their acoustic signature to different career groups. We refer to the career group of the speakers as the states (hidden) and the heard acoustics as the observations. The career sequence (hidden states) which is the probability of being in a state and the probability of transition to another state is generated according to some distribution. And for each career state, the output acoustic signals (observations) are generated according to another distribution.

HMM uses the observed acoustic distribution to describe the career sequence distribution. To formally describe the HMM model in this experiment, let S_i's correspond to different career groups (states). Each state is associated with a probability density function for the emitted acoustics at a time t; these are represented by O_i's. We assume that the state and the observation sequences form a first order Markov chain i.e. their density functions are independent of the previous state or observations. As a Markov chain, we can describe the career sequence clearly if we can estimate the transition probabilities between different career states; say a_{ij}, and the initial probability of each career state; say π_i. Well, the HMM model supplies us with these quantities.

Formally, the HMM model with states $\{\boldsymbol{S_i}\}$ is represented by the states probabilities and transition probabilities $\{\boldsymbol{\pi_i}, \boldsymbol{a_{ij}}\}$ and the observation probabilities $\{\boldsymbol{B_i}\}$. It is written as: $M = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$.

HMM in Speech Recognition

HMM provides a simple and effective framework for modeling time-varying spectrum signal such as in speech [95]. HMM has an advantage of preserving the temporal information content of speech. There are two major phases of HMM modeling in speech recognition: the training phase and the recognition phase.

- In the training phase, a set of testing observations is used to derive the reference models corresponding to the number of classes (or words in our case).

- In the recognition phase, the probability of generating the unknown observations is computed against each of the reference models and the model leading to the highest probability is declared as the selected class (or word).

The left-to-right type of HMM is the typical choice for modeling speech recognition application. The HMM model is expressed by the set M = ($\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$) and is characterized by the following elements:

1. N is the number of states in the model. “In the speech recognition context, there [are] two methods to estimate the value of N” [16]. These are:
   a. States representing different phonetic units (phonemes, phones, syllables...). Usually, N is around 6 in speech application.
   b. States have no-phonetic meaning, they represent basically temporal frames.
2. $\boldsymbol{S} = \{s_1, \ldots, s_N\}$ is a set of all states in a model.
3. T is the number of observations in a sequence.
4. $\boldsymbol{I} = \{s_1, \ldots, s_T\}$ is a set of all possible state sequence, where s_t, represents the state at time t, $s_t \in \boldsymbol{S}$.
5. R is the total number of distinct observation symbols per state (assume to be the same for all states).
6. $\boldsymbol{V} = \{v_1, \ldots, v_R\}$ is a set of all possible observation symbols.
7. $\boldsymbol{O} = \{O_1(v), \ldots, O_T(v)\}$ is a set of all possible observation sequence, $O_t(v)$ is the observation symbol v at time t.

8. $\boldsymbol{B} = \{b_i(v)\}$, $b_i(v) = P(O_t(v) \mid x_t{=}s_i)$, the probability of getting observation $O_t(v)$ from state s_i at time t.

9. $\boldsymbol{\pi} = \{\pi_i\}$, $\pi_i = P(s_{1=}s_i)$, the initial probability of state s_i, (i =1, 2, … N, $\sum_i \pi_i = 1$).

10. $\boldsymbol{A} = \{a_{ij}\}$, $a_{ij} = P(x_{t+1}{=}s_j \mid x_t{=}s_i)$, the transition probability from state s_i at time t to state s_j at time $t+1$.

Figure 2.5 shows an example of HMM model of 3 states and 4 observations.

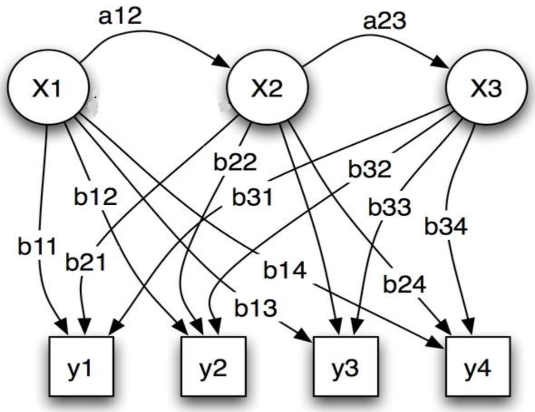

Figure 2.5: HMM Model of 3-States and 4-Observations

To fully describe the HMM model, 3 main problems should be addressed:

1. *The Evaluation Problem*: problem of computing $P(\boldsymbol{O} \mid M)$, the probability of an observation sequence given the model.

2. *The Decoding Problem*: problem of maximizing $P(\boldsymbol{O}, \boldsymbol{I} \mid M)$, the probability of an observation sequence and the state sequence given the model.

3. *The Training Problem*: problem of adjusting the model M = ($\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$) parameters to maximize $P(\boldsymbol{O}| M)$ or $P(\boldsymbol{O}, \boldsymbol{I} \mid M)$.

Problems 2 and 3 represent the training phase of HMM, while problem 1 represents the recognition phase. The mathematical solutions for those problems are briefly described below:

Solution to problem 1 (Forward-Backward Procedure)

Using Baye's rule, one can calculate $P(\mathbf{O} \mid M)$ by starting with $P(\mathbf{O}, \mathbf{I} \mid M)$ for all possible state sequence and multiplying it by $P(\mathbf{I} \mid M)$, then sum up over all possible $\mathbf{I}$'s,

$P(\mathbf{O}, \mathbf{I} \mid M) = b_{i1}(O_1)\, b_{i2}(O_2) \ldots.\, b_{iT}(O_T)$ (2.33)

$P(\mathbf{I} \mid M) = \pi_{i1} a_{i1\, i2} a_{i2\, i3} \ldots.. a_{iT\text{-}1\, iT}$ (2.34)

We can then write:

$P(\mathbf{O}\mid M) = \sum P(\mathbf{O}, \mathbf{I} \mid M)\, P(\mathbf{I} \mid M)$ (2.35)

$= \sum \pi_{i1}\, b_{i1}(O_1)\, a_{i1\, i2}\, b_{i2}(O_2) \ldots\, a_{iT\text{-}1\, iT}\, b_{iT}(O_T).$ (2.36)

Since there is N^T distinct possible state sequence of I and (2.36) involves $2T\text{-}1$ multiplication so the total number of computation is $2TN^T$ multiplications and $N^T\text{-}1$ additions, which is extensive computation. Hence, we need more efficient procedure for solving this problem. We use here the popular Forward-Backward Procedure [31]:

Forward Procedure:

Consider the forward variable $\alpha_t(i)$ defined as:

$\alpha_t(i) = P(O_1, O_2, O_3, \ldots. , O_t, i_t{=}i \mid M),$ (2.37)

Which is the probability of the partial observation sequence up to time t and the state i at time t, given the model M. Then, we can use the following iterations:

1. initialization:

$$\alpha_t(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N \qquad (2.38)$$

2. Recursion:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad t = 1....T-1, \quad 1 \le j \le N \qquad (2.39)$$

3. Termination:

$$P(O \mid M) = \sum_{i=1}^{N} \alpha_T(i), \qquad (2.40)$$

Here, we will only need *N+N(N+1)(T-1)* multiplications and *N(N+1)(T-1)* additions (i.e. order of N^2T multiplication) [31].

Backward Procedure:

Consider:

$$\beta_t(i) = \mathrm{P}(\mathrm{O}_{t+1}, \mathrm{O}_{t+2}, \ldots., \mathrm{O}_\mathrm{T} \mid \mathrm{i}_\mathrm{t} = \mathrm{i}, M), \qquad (2.41)$$

Which is the probability of the partial observation sequence from *t+1* to *T* given the state *i* at time *t*, and the model *M*. Then we perform the following:

1. initialization:

$$\beta_T(i) = 1 \qquad 1 \le i \le N \qquad (2.42)$$

2. Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2...1, \qquad 1 \le i \le N \qquad (2.43)$$

3. Termination:

$$P(O \mid M) = \sum_{i=1}^{N} \pi_i b_i(O_1) \beta_1(i), \qquad (2.44)$$

Here, we will also have an order of N^2T multiplications [31].

Both the Forward and Backward Procedures are equivalent and efficient for the computation of *P(**O**| M)*.

Solution to problem 2 (Viterbi Algorithm)

Here, we have to find **I** that maximize *P(**O**, **I** | M)*. The famous algorithm to solve that is the Viterbi Algorithm, which is an inductive algorithm in which at each instant; the best possible state sequence is kept as intermediate sequence towards the desired observation sequence.

Let $w = -\ln\left(a_{ij} b_j\left(O_t\right)\right)$ is the weight on the path from state *i* to state *j* , $\delta_t(i)$ denote the accumulative weight when we are in state *i* at time *t* and $\psi_t(j)$ denote the state at time *t-1* which has the lowest cost corresponding to the state transition to state *j* at time *t*. the algorithm proceed as follows:

1. initialization:

$$\left.\begin{aligned} \delta_1(i) &= -\ln\left(\pi_i\right) - \ln\left(b_i\left(O_1\right)\right) \\ \psi_1(i) &= 0 \end{aligned}\right\}, \qquad 1 \le i \le N \qquad \begin{aligned} &(2.45) \\ &(2.46) \end{aligned}$$

2. Recursion:

$$\left.\begin{aligned} \delta_t(j) &= \min_{1 \le i \le N} \left[\delta_{t-1}(i) \ln\left(a_{ij}\right)\right] - \ln\left(b_j\left(O_t\right)\right) \\ \psi_t(j) &= \arg\min_{1 \le i \le N} \left[\delta_{t-1}(i) \ln\left(a_{ij}\right)\right] \end{aligned}\right\}, \begin{cases} 2 \le t \le T \\ 1 \le j \le N \end{cases} \qquad \begin{aligned} &(2.47) \\ &(2.48) \end{aligned}$$

3. Termination:

$$P^* = \min_{1 \le i \le N} \left[\delta_T(i)\right] \qquad (2.49)$$

$$q_T^* = \arg\min_{1 \le i \le N} \left[\delta_T(i)\right] \qquad (2.50)$$

4. Tracing back the optimal state sequence:

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right) \qquad , \begin{cases} t = T - 1 \\ T - 2, \ldots 1 \end{cases} \qquad (2.51)$$

Here P^* gives the required state-optimized probability and $\boldsymbol{Q}^*$ = { q_1^*, q_2^*,… q_T^* }is the optimal state sequence. The complexity of this recursion is of order N^2T [31].

Solution to problem 3

As we know there are two identical probability functions used for identification; *P(**O**, **I**| M)* and *P(**O**| M)*. Based on that, there are two popular techniques used for solving the training problem.

1) **The Segmental K-means algorithm**: where the parameters of *M* = *(**π**, **A**, **B**)* are adjusted to maximize *P(**O**, **I**| M)*. This criterion is called maximum state optimized likelihood criterion. This algorithm is carried out as follows:

   a) Randomly chose N observation symbols. The initial choice of N doesn't represent the final HMM however, it is critical to estimate the number of iterations required for HMM training. So Segmental K-means algorithm is biased upon the right choice of N.

   b) Calculate the initial probability and the transition probability as:

$$\hat{\pi}_i = \frac{Number\ of\ coccurence\ of\ \{O_1 \in i\}}{Total\ Number\ of\ coccurence\ of\ O_1}, \qquad 1 \le i \le N\square \qquad (2.52)$$

$$\hat{a}_{ij} = \frac{Number\ of\ coccurence\ of\ \{O_t \in i\, and\, O_{t+1} \in j\}\ for\ all\ t}{Total\ Number\ of\ coccurence\ of\ \{O_t \in i\}\ for\ all\ t}, \ 1 \le i, j \le N\square \quad (2.53)$$

c) Calculate the mean vector and the covariance matrix for each state:

$$for\, 1 \le i \le N$$

$$\hat{\mu} = \frac{1}{N} \sum_{O_t \in i} O_t \qquad \square \ (2.54)$$

$$\hat{V}_i = \frac{1}{N} \sum_{O_t \in i} (O_t - \hat{\mu}_i)^T (O_t - \hat{\mu}_i) \qquad \square \ (2.55)$$

d) Calculate the symbol probability distribution for each training vector for each state:

$$for\, 1 \le i \le N$$

$$\hat{b}_i(O_t) = \frac{1}{(2\pi)^{D/2} \left|\hat{V}_i\right|^{1/2}} \exp\left[-\frac{1}{2}(O_t - \hat{\mu}_i)\hat{V}_i^{-1}(O_t - \hat{\mu}_i)^T \right] \qquad \square \ (2.56)$$

e) Find the optimal state sequence (solution of problem 2) for each training sequence using $\hat{M}_i = (\hat{\pi}_i, \hat{A}_i, \hat{B}_i)$. A vector is reassigned a state if its original assignment is different from the estimated optimum state.

f) If any vector is reassigned, use the new assignment and repeat the process.

2) **The Baum-Welch Re-estimation Formulas**: where the parameters of M = ($\boldsymbol{\pi}$, $\boldsymbol{A}$, $\boldsymbol{B}$) are adjusted to maximize $P(\boldsymbol{O}|\ M)$. This criterion is called maximum likelihood criterion.

Define $\gamma_t(i) = P(i_t = s_i \mid O, M)$, the probability of being in state i at time t given the observation sequence $\boldsymbol{O}$, and the model M,

$$\begin{aligned} \gamma_t(i) &= \frac{P(i_t = s_i, O \mid M)}{P(O \mid M)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{P(O \mid M)} \end{aligned} \qquad (2.57)$$

Define $\zeta(i, j)=P(i_t=s_i,\ i_{t+1}=s_j \mid \boldsymbol{O}, M)$, probability of being in state s_i at time t and making a transition to state s_j at time $t+1$ given the observation sequence $\boldsymbol{O}$, and the model M.

$$\xi(i, j)=\frac{P(i_t=s_i, i_{t+1}=s_j, O \mid M)}{P(O \mid M)}$$

$$=\frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O \mid M)} \qquad \square(2.58)$$

Hence, it is clear that:

$$\sum_{t=1}^{T-1}\gamma_t(i)=Expected\ \ Number\,of\ \ transition\ \ from\,state\,i \qquad \square \qquad (\square 2.59)$$

$$\sum_{t=1}^{T-1}\xi_t(i, j)=Expected\ \ Number\,of\ \ transition\ \ from\,state\,i\,to\,state\ \ j \qquad (\square 2.60)$$

Now the Baum-Welch reestimation formulas become:

$$\hat{\pi}_i=\gamma_t(i), \qquad\qquad 1 \le i \le N \qquad\qquad \square\ (2.61)$$

$$\hat{a}_{ij}=\frac{\sum_{t=1}^{T-1}\xi_t(i, j)}{\sum_{t=1}^{T-1}\gamma_t(i)}, \qquad\qquad \square\ (2.62)$$

$$\hat{b}_j(k)=\frac{\sum_{\substack{t=1\\ O_t=k}}^{T}\gamma_t(j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \qquad\qquad \square\ (2.63)$$

In general, the Segmental K-means algorithm is more preferable than the Baum-Welch algorithm as it involves manipulating of small values (since no summation is involved). Also Segmental K-means requires much less computation than Baum-Welch. However, many of the existing systems still use the traditional Baum-Welch re-estimation approach. In summary, we have discussed the features used in representing speech and we have provided different examples of these features. We also provided a summary table comparing the most popular features. Our focus in this research is to propose a new approach for selecting the optimal features for speech recognition applications. Before discussion our approach in detail, we will first introduce the concept of Mutual Information and describe its different applications in speech analysis.

CHAPTER 3

INFORMATION THEORY AND SPEECH ANALYSIS

3.1 Background

Suppose there is a closed envelop and you are asked to guess the message inside. You can only ask yes/no questions about the contents of the message. Assuming this exercise is repeated many times, and you get as clever as possible while choosing your questions. The question now what's the smallest number of questions needed, on average, to guess correctly the contents of the message? The answer to this question is as follows; suppose there are only a finite number of words inside the message or just there is a limit on the length of the messages. Then you can number the words/characters from 1 to N. Call the word you get on each trial *X*. Since the game is repeated many times we can define $P(X=i)=p_i$ as the probability of getting word/character number *i* for *N* words/characters on any given trial. The maximum number of yes/no questions needed to pick out any given message is calculated as *log N*. But you can do better than this: if message *i* is more frequent than message *j* (i.e. $p_i > p_j$), it will save time if you ask whether the message is *i* before considering the possibility that it is *j*. We will define *H* as the smallest average number of questions needed. In probability, *H* is called the information content or the entropy of the message.

Now suppose there are two envelopes and you are asked to know the messages inside. The question here: how many questions will that take? Call the two variables as *X* and *Y*. To find out the value of *X* takes *H[X]* questions and for it *Y* takes *H[Y]*. So we need together a maximum of *H[X]* + *H[Y]* questions. But some combinations of messages may be more likely than others which is due to information dependency of messages also called *joint entropy, H[X, Y]*.

Now suppose that you found the content of message *X*, the question is how many questions you will need to find out the contents of *Y*? This is what we call the *conditional entropy,* the entropy of *Y* conditioned on *X*, written as *H[Y|X]*.

The concept of entropy (in term of probability theory) was introduced, simultaneously, by some mathematicians and engineers during World War II (among the Americans Claude Shannon and Norbert Wiener), working on a major practical problems of coding, code-breaking, communication and automatic control. Norbert Wiener worked out the continuous case of the standard entropy/coding/ communication channel part of information theory at the same time as Shannon was working on the discrete case. Information theory was first introduced in 1948 by Claude Shannon who is known as the father of information theory. He discussed some fundamental limits on the representation and transmission of information. Since that time, the results have been extended to cover multiple areas. In a sense, information theory has provided the theoretical motivation for many great advances in digital communications and digital storage. For example, what is the optimum size of information that can be sent over the phone system of certain properties?

3.2 The Concept of Mutual Information

Mutual information, written as $I[X; Y]$, is defined as the amount of information we can learn about Y from our knowledge of X. Consider our previous example of closed envelop, the mutual information represents the number of questions we can save from the original number of questions required to identify the message content. The mutual information quantifies how much one message can tell us about another.

Mathematical Description of Mutual Information Concepts:

We will first introduce the concept of entropy which is a measure of uncertainty of a random variable. Let X be a discrete random variable with sample values $\{x_1, x_2, \ldots, x_N\}$ with probability mass function $p(x_i) = P[X=x_i]$

The entropy is then defined as the average information over all instances (N) of the random variable X:

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log\left(P(x_i)\right) \qquad (3.1)$$

The joint entropy measures the dependence of a random variable X on another random variable Y and it is defined as:

$$H(X,Y) = -\sum_{i,j} p_{x_i,y_j} \log\left(p_{x_i,y_j}\right) \qquad (3.2)$$

where p_{x_i,y_j} represents the joint probability mass function of the random variable X and Y (assuming X and Y are discrete random variables).

The conditional entropy is another measurement of randomness of *Y* given the knowledge of *X* and it is defined as:

$$H(X|Y) = H(X,Y) - H(Y) \qquad (3.3)$$

For continuous random variables, replace $\sum$ with $\int$ in all previous expressions.

The conditional entropy can provide information on 2 random variables when they are completely independent $H(X|Y) = H(X)$, however it is insufficient to inform about their dependency. A small value of *H(X|Y)* may imply that *Y* provides us a great information about *X*, the value of *H(X)* is small or the value of *H(Y)* is large. For that the mutual information is defined. The mutual information *I(X; Y)* between 2 random variables *X* and *Y* is defined as the reduction of randomness of a random variable *X* given a prior knowledge of another random variable *Y*:

$$I(X;Y) = H(X) - H(X|Y) \qquad (3.4)$$

For discrete random variables, the mutual information is calculated using:

$$I(X;Y) = \sum_i \sum_j p(x_i, y_j) \log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) \qquad (3.5)$$

where:

p(x_i, y_i) is the joint probability density function of *X* and *Y*, and

p(x_i) and *p*(y_i) are the marginal probability density functions of *X* and *Y* respectively.

For continuous random variables, the mutual information from Eq. 3.6 is rewritten as:

$$I(X;Y) = \int_y \int_x \, p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dx\, dy \qquad (3.6)$$

where:

p(x, y) is the joint probability density function of *X* and *Y*, and

p(x) and *p(y)* are the marginal probability density functions of *X* and *Y* respectively.

Conventionally, the mutual information is measured in *bit* since log base-2 is used. The mutual information measures the reduction of randomness (uncertainty) of a random variable *X* given the knowledge of another random variable *Y*. Or in other words, the mutual information measures how much our knowledge about one random variable (*Y)* reduces our uncertainty about another random variable (*X)*. For example, if *X* and *Y* are independent, then knowing *X* does not give any information about *Y* and vice versa, so their mutual information is zero. On the other hand, if *X* and *Y* are identical, then knowing *X* completely determines *Y* and vice versa, so their mutual information is the same as the uncertainty contained in one of the variables (their entropy).

The conditional mutual information is also used to describe the mutual information of two random variables conditioned on a third one. For discrete random variables *X*, *Y* and *Z*, the conditional mutual information is defined as:

$$I(X;Y|\,Z) = \sum_{z\in Z}\sum_{y\in Y}\sum_{x\in X} p_{X,Y,Z}(x,y,z)\log\left(\frac{p_Z(z)p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)}\right) \qquad (3.7)$$

Where p is represents marginal, joint, and/or conditional probability density functions.

Mutual Information conditioning on a third random variable may either increase or decrease the original mutual information, however *I(X,Y | Z)* is always positive.

The Mutual information *I(X, Y)* has a number of properties:

1. Symmetry: $I(X;Y) = I(Y;X)$, but (*H(X) ≠ H(Y)*) and (*H(X/Y) ≠ H(Y/X)*).

2. It is "0" iff X and Y are independent: $I(X;Y)= H(X)-H(X|Y)$ and $H(X|Y)= H(X)$ when X and Y are independent.
3. Non-negativity: $I(X;Y) \geq 0$.
4. Additivity: $I(X; Y, Z) = I(X;Y) + I(X; Z| Y)$.

Typical applications of Mutual information:

Mutual information has been largely used in the context of communication system. However the concept of mutual information has been expanded to solve difficult estimation and data analysis problems in biomedical applications [61], image processing and signal processing. The key in using of mutual information in these applications is to measure the independence between two random variables or distributions. In image processing [62] and speech recognition [63], the use of the maximum mutual information (MMI) between the observed data and available models has resulted in powerful algorithms for training models for classifications.

Some examples of areas where Mutual Information is applied include:

1. Coding for channel capacity.
2. Discriminative feature selection and transformation.
3. Training of hidden Markov models.
4. Error coding detection.
5. Medical study of genes.
6. Phase synchronization detection.
7. Clustering comparing measure.
8. Corpus linguistics.

In summary, the objective of this chapter was to provide an overview of the concept of Mutual Information as it plays an important role in our research. We use Mutual Information as a measure of correlation in our feature selection algorithm. The detailed description of the proposed feature selection algorithm, the experimental approach, design and results will now be explained in chapter 4.

CHAPTER 4

THE PROPOSED FEATURE SELECTION ALGORITHM USING THE MRMR

4.1 The Minimum Redundancy–Maximum Relevance Algorithm for Feature Selection

In many pattern recognition applications, identifying the most characterizing features from observed data (feature selection) is critical to minimizing the classification error [53]. From classification point of view, the extracted features are classified as [112]:

1. **Relevant:** Features with high impact on classification accuracy.
2. **Irrelevant:** Features with negligible impact on classification accuracy.
3. **Redundant:** Features that can be replaced by other features with no change on classification accuracy.

Dealing with a large number of redundant features is inefficient in terms of time and processing and may lead to inaccurate conclusion. In addition, irrelevant features may “confuse” the classification algorithms leading to wrong decisions. Hence, it is important to use the right features before a given classification task. The main task of feature selection is to determine the minimum set of relevant features that highly represent the original features. This selection is achieved by eliminating features with no or little impact on classification.

Feature selection reduces the computational cost by reducing the space dimension as well as improving classification accuracy. Other advantages of feature selection include reducing effect of noise, facilitating data visualization and data understanding, reducing measurement and storage requirements and improving data quality.

Given the input observations and M feature vectors $\boldsymbol{Q} = \{x_i; i =1, . . ., M\}$, and the target classification vector $\boldsymbol{C}$, the feature selection problem is that of finding from the *M-dimensional* feature space, S^M, a subspace of m features, S^m, that "optimally" characterizes $\boldsymbol{C}$ [54].

4.1.1 Feature Selection Algorithms (FSA):

"A feature selection algorithm (FSA) is a computational solution that is motivated by a certain definition of relevance [112]". Characterization of FSA can be seen mainly from 3 different dimensions: search strategy, successor generation and evaluation. Search strategy is the algorithm that drives the feature selection process according to a specific strategy. There are 3 types of search: exponential, sequential and random. Successor generation is the mechanism used to generate a successor from all possible variants. The most operators used as successor generators are: Forward, Backward, Compound, Weighting, and Random. Evaluation measure is the function that evaluates the generated successor to guide the search algorithm. Many evaluation measurements are used in feature selection and the famous ones are: Divergence, Dependence, Information Theory and Consistency. By combining these characteristics, numerous FSAs have been developed [90]. Table 4.1 presents the different combinations of search strategy and evaluation function that are available in the literature.

Table 4.1: Available Feature Selection Methods by Search Strategy and Evaluation Function [90]

Search Strategy	Evaluation Function	Consistency				Information Theory			Distance	Performance
		Basic consistency	Inconsistent examples (Liu)	Rough Set consistency	Inconsistent example pairs (IEP)	Mutual Information	Symmetrical uncertainty	Minimum description length criterion	Relief based feature set measure (RFSM)	Wrapper (using 10 fold cross validation)
Complete	Exhaust.	X	✓	✓	✓	✓	✓	✓	✓	✓
	FOCUS 2	✓	X	X	X	X	X	X	X	X
	B & B	X	✓	✓	✓	✓	X	X	✓	X
	ABB	X	✓	✓	✓	✓	X	X	✓	X
Sequential	SFS	X	✓	✓	✓	✓	✓	✓	✓	✓
	SBS	X	✓	✓	✓	✓	✓	✓	✓	✓
Probabilistic	LVF	X	✓	✓	✓	✓	X	X	✓	X
	LVW	X	X	X	X	X	✓	✓	X	✓
Meta-heuristic	SA	X	✓	✓	✓	✓	✓	✓	✓	✓
	GA	X	✓	✓	✓	✓	✓	✓	✓	✓
	TS	X	✓	✓	✓	✓	✓	✓	✓	✓

Feature selection techniques can mainly be classified into two categories according to how the search technique combines with the classification model. These techniques are *the wrappers* and *the filters* [100].

The Wrappers feature selection approach wraps features around a specific prediction or classification method. This means that with very small number of features, we can get high accuracy since the features' characteristics match well the learning algorithm used. The advantage of the wrapper method is that the selected feature subset is often very small and gives high performance, but it has an extensive computation process. Also it has a problem of stability and sensitivity of the selected features where the selected subset changes significantly with changing the classification methods, and/or when adding more data points. According to the search characteristic, wrapper methods can be classified into: deterministic and randomized wrappers. [55]

The Filters feature selection approach selects features based on their relevance with respect to the targeted classes. The name of this method is related to the concept of data filtering. This method is popularly used in practice. The advantage of this method is that the selected features from training data can be generalized to new data. The common practical process of the filter method is simply to rank the features and select the top-ranked ones. A problem of this approach is that these features could result in high mutual information between these i.e. the features are correlated or redundant, and therefore they are not provide good representations of target classes. There are two aspects to this problem:

1) Efficiency: if a feature set of say 50 features contains a good number of mutually highly correlated features, the true representative "independent" features are therefore much fewer, say 20. We can delete the 30 highly correlated features without affecting the performance, which means there are 30 wasted or redundant features.

2) Broadness: because the features are selected according to their discrimination, they are not highly representing the original space covered by the entire dataset. The feature set may represent several dominant characteristics of the target classes; however these could not be enough to generalize to a larger dataset. Filter methods can be univariant or multivariant based on how many features are considered at a time. The table below shows the classification of feature selection methods with their advantages and disadvantages and some popular examples for each type. The choice for the best method is based on the goal and the available resources [112].

Table 4.2: Classification of Feature Selection Methods [112]

Model		Advantage	Disadvantage	Example
Filter	Univariant	Fast, Scalable, Independent of the classifier	Ignores feature dependencies, Ignores interaction with the classifier	X^2, Euclidian distance, t-test, Information gain
	Mulivariant	Models feature dependencies, Independent of the classifier, Better computational complexity than wrapper methods	Slower than univariate techniques, Less sclable than univariate techniques, Ignores interaction with the classifier	mRMR Correlation-based feature selection(CFS), Markov blanket filter (MBF), Fast correlation-based feature selection (FCBF)
Wrapper	Deterministic	Simple, Interacts with the classifier, Models feature dependencies,	Risk of over fitting, More prone than randomized algorithms to getting stuck in a local optimum (greedy	Sequential forward selection (SFS), Sequential backward elimination (SBE),

		Less computationally intensive than randomized methods	search), Classifier dependent selection	Plus L Minus R, Beam search
	Randomized	Less prone to local optima, Interacts with the classifier, Models feature dependencies	Computationally intensive, Classifier dependent selection, Higher risk of over fitting than deterministic algorithms	Simulated annealing, Randomized hill climbing, Genetic algorithms, Estimation of distribution algorithms

As we mentioned earlier, the majority of FSAs operate by selecting the maximum relevant features and ignoring any redundancy between the selected features. This redundancy may degrade the performance and reduce the accuracy. The mRMR is one of the most robust algorithms proposed to overcome this problem.

4.1.2 The minimum Redundancy, Maximum Relevance (mRMR):

The mRMR is an enhanced version of the general filter based feature selection. Its objective is to "expand the representative power of the feature set" [101] by maximizing the features discrimination properties. The mRMR was first introduced by Ding and Peng in 2003 in their original paper to improved class predictions of microarray genes data [51]. The benefits of this mRMR approach can be summarized in two ways [52]:

1) Generalization: with the same number of features, the mRMR feature set is more representative of the target classes; hence it leads to a better overview.
2) Equivalently: with smaller feature set, the mRMR effectively covers the same space as a larger conventional feature set does.

The mRMR is a simultaneous optimization of a minimum redundancy criterion and maximum relevance criterion. The aim of the former is to select the feature sets that are maximally dissimilar among themselves. While the goal of the latter is to select the feature sets that are maximally similar to the target classes. The mRMR is described using information theory concept by finding the features that are mutually apart from each other (minimum redundancy) while they are individually most comparable to the observation set (maximum relevance) [103].

To explain the concept of mRMR, suppose we have a group of people and we categorize them into classes of doctors, engineers, scientist, artist and basketball-players. To describe these classes, we list some features including age, tall, educational degree, educational period, white dress, monthly income, stethoscope, glasses, brush, computer, microscope and hat. We would like to use the mRMR process to select the features that best describe the class Doctors. We know that a doctor has a high educational degree, spends longer in education, uses stethoscope as a main tool, wears white robe, has high monthly income and he is usually quite old in age. These 6 features are the most features classify the class Doctor. This ordered list of features represents the maximum relevance part of mRMR. However, there are some dependant features like using stethoscope as a main tool and wearing white dress. Also spending longer in education and elderliness are dependant. So having both features in the list doesn’t provide additional information in classifying the classes. For more efficiency, this redundancy (dependant features) is minimized by removing the less relevant feature while keeping the other one. This process represents the minimum redundancy part of mRMR. Finally; to specify the class

Doctor, there are 4 extracted features that have the maximum relevance with the class Doctor and the minimum redundancy between themselves.

Mutual information is used to measure the correlation “similarity” between the features themselves as well as between each feature and the target classes. Maximum relevance is calculated by the mean value of the mutual information values between individual features and the target classes. And minimum redundancy is calculated also by the mean value of the mutual information values between each pair of features.

4.1.3 Mathematical Formulation of the mRMR Algorithm:

Let ***F*** represents the pool of all features; ***F*** = $\{f_i;\ i=1, \ldots, M\}$, let ***S*** represents the subset of features that we are seeking to find ($\boldsymbol{S} \subset \boldsymbol{F}$); ***S*** = $\{f_i;\ i=1, \ldots, m\}$, and let ***C*** = $\{c_1, \ldots, c_K\}$ represent the K target classes.

For discrete random variables, we use the mutual information $I(f_i, f_j)$ to measure the level of similarity among the features f_i and f_j. Also, we use mutual information $I(\boldsymbol{C}, f_i)$ to measure the level of relevance (discriminate) between the target classes ***C*** and the feature f_i.

The mutual information between two discrete random variables X and Y is defined based on their joint probability distribution function $p(x_i, y_j)$ and their respective marginal probability function s $p(x_i)$ and $p(y_j)$ and calculated by Eq. 3.6.

Based on the above, we define the minimum redundancy as:

$$\min_{\boldsymbol{S}} D, \qquad D = \frac{1}{|\boldsymbol{S}|^2} \sum_{i,j \in \boldsymbol{S}} I\left(f_i, f_j\right) \qquad (3.8)$$

where $I(f_i, f_j)$ is the mutual information between each pair of features f_i and f_j, and $|\boldsymbol{S}|$ is the number of features in $\boldsymbol{S}$.

The maximize relevance, on the other hand, is defined as:

$$\max_{S} V \ , \qquad V = \frac{1}{|\boldsymbol{S}|} \sum_{i \in \boldsymbol{S}} I(\boldsymbol{C}, f_i) \tag{3.9}$$

where $I(\boldsymbol{C}, f_i)$ is the mutual information between the class set $\boldsymbol{C}$ and the feature f_i.

To simplify the mRMR for discrete random variable, we will have to quantize the observations of the features.

Continuous Case:

For continuous random variables, the Pearson's correlation coefficient $C_{or}(i, j)$ or Euclidean distance $d(i, j)$ is used for the minimum redundancy condition.

Pearson's correlation coefficient between two variables X and Y is defined as the covariance of the two variables divided by the product of their standard deviations that is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X) - (Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{3.10}$$

For n observations, the Euclidean distance between sets $\{x_i\}$ and $\{y_i\}$ is defined as:

$$d(x,y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + ... + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{3.11}$$

For maximizing the relevance, a statistical test such as F-test is used to compare between class set $\boldsymbol{C}$ and the feature f_i. F-test; $F_t(f_i, C)$; is defined as:

$$F_t(\boldsymbol{C}, f_i) = \left[\sum_k n_k (\bar{f}_k - \bar{f}) \Big/ (K-1) \right] \Big/ \sigma^2 \tag{3.12}$$

where $\bar{f}$ is the mean value of all observations (k=1, 2, …,K) of all features, $\bar{f}_k$ is the mean value of the observations of f_i over the k^{th} class, and $\sigma^2 = \frac{\left[\sum_k (n_k - 1)\sigma_k^2\right]}{(n-K)}$, is the pooled variance (where n_k and σ_k are the size and the variance of the k^{th} class).

The minimum redundancy condition for the continuous case becomes:

$$\min_{\boldsymbol{S}} D,\ D = \frac{1}{|\boldsymbol{S}|^2} \sum_{i,j \in \boldsymbol{S}} \left| C_{or}\left(f_i, f_j\right) \right|, \qquad \textit{for correlation coefficient} \qquad (3.13)$$

$$\min_{\boldsymbol{S}} D,\ D = \frac{1}{|\boldsymbol{S}|^2} \sum_{i,j \in \boldsymbol{S}} d\left(f_i, f_j\right), \qquad \textit{for Euclidean distance} \qquad (3.14)$$

where $C_{or}(f_i, f_j)$ and $d(f_i, f_j)$ are the correlation coefficient and Euclidian distance (Eq. 3.10 and 3.11) between each pair of features f_i and f_j.

Moreover, the maximize relevance condition is:

$$\max_{\boldsymbol{S}} V\ , \qquad V = \frac{1}{|\boldsymbol{S}|} \sum_{i \in \boldsymbol{S}} F_t\left(\boldsymbol{C}, f_i\right) \qquad (3.15)$$

where $F_t(\boldsymbol{C}, f_i)$ is the F-test (Eq. 3.12) between the class set $\boldsymbol{C}$ and feature f_i.

The mRMR feature set is obtained by optimizing these two conditions simultaneously. The simplest way for simultaneous optimization of the 2 quantities is to maximize the difference between these quantities (additive combination). An alternative approach is to maximize the quotient (multiplicative combination) of the 2 cost functions.

1. Additive combination:

$$\max(V - D) \qquad (3.16)$$

The additive combination for discrete random variables based on mutual information is called Mutual Information Difference (MID). However; for continuous random variables based on F-test correlation, it is called F-test Correlation Difference (FCD).

2. Multiplicative combination:

$$\max\left(\frac{V}{D}\right) \qquad (3.17)$$

The multiplicative combination for discrete random variables based on mutual information is called Mutual Information Quotient (MIQ). On the other hand; for continuous random variables based on F-test with correlation, the multiplicative combination is called the F-test Correlation Quotient (FCQ). However, for continuous random variables based on F-test with distance, the algorithm is called the F-test Distance Multiplication (FDM) (for multiplication) or the F-test Similarity Quotient (FSQ) (for quotient).

The computational cost of evaluating correlations or mutual information in mRMR is *$O(NM^2)$*, where *M* is the number of feature set and *N* is the training set size [102].

4.1.4 The mRMR Algorithm Procedure:

The main steps of the mRMR consist of N-1 iterations, where N is the number of the desired selected features. The mRMR is carried through the following steps:

1) Maximum Relevance: the mutual information between each feature set and the target class is calculated, then, the features are ranking in a descending order according to the values of their mutual information. The feature of the highest mutual information value is declared as the maximum relevance feature (Eq. 3.9).

2) Minimum Redundancy: in each of N-1 iterations, we calculate the mutual information between the selected feature in the previous iteration (maximum relevance one at the first iteration) and the rest of the features in the order of the maximum relevance step. Then, we average the mutual information values of each of the remaining features with all the previous selected features as per Eq.3.8.
3) mRMR step: for N-1 iterations, the mRMR feature is selected by either additive combination (Eq. 3.16) or multiplicative combination (Eq. 3.17). In both cases, the mathematical operation is carried out between the mutual information values of the maximum relevance step and the average mutual information values resulted from the minimum redundancy step for the remaining features ordered as of maximum relevance order.

4.1.5 Numerical Example for The mRMR Algorithm Procedure:

Consider that we have a pool of 6 features and we need to select the best 3 features targeting to a specific classification task using the mRMR algorithm. Since N= 3 in this example, we need 2 iterations. We apply the steps mentioned above as follows: (the values below are just to explain the concept).

1) Maximum Relevance (*V*):

I(C, f1)=	0.50
I(C, f2)=	0.40
I(C, f3)=	0.49
I(C, f4)=	0.47
I(C, f5)=	**0.53**

⇨

f5
f1
f3
f4
f2

$I(C, f6)=$	0.36	*f6*

2) 1st Iteration:

2.1) Minimum Redundancy (D) with the 1st selected feature (*f5*):

$I(f5,\ f1)=$	0.16
$I(f5, f2)=$	0.14
$I(f5, f3)=$	0.20
$I(f5, f4)=$	0.10
$I(f5, f5)=$	0
$I(f5, f6)=$	0.15

	I_1	mean
mean $I(f5, f1)=$	0.16	0.16
mean $I(f5, f3)=$	0.20	0.20
mean $I(f5, f4)=$	0.10	0.10
mean $I(f5, f2)=$	0.14	0.14
mean $I(f5, f6)=$	0.15	0.15

2.2) mRMR (V-D) or (V/D):

Feature	*V*	*D*	*mRMR (V-D)*	*mRMR (V/ D)*
f1	0.50	0.16	0.34	3.13
f3	0.49	0.20	0.29	2.45
f4	0.47	0.10	**0.37**	**4.70**
f2	0.40	0.14	0.26	2.86
f6	0.36	0.15	0.21	2.40

Based on the above table, we select feature *f4***.**

3) 2nd Iteration:

3.1) Minimum Redundancy (D) with the 2nd selected feature (*f4*):

I(f4, f1)=	0.30
I(f4, f2)=	0.24
I(f4, f3)=	0.40
I(f4, f4)=	0
I(f6, f6)=	0.13

	I_1	I_2	mean
mean *[I(f5, f1) + I(f4, f1)]*	0.16	0.30	0.23
mean *[I(f5, f3) + I(f4, f3)]*	0.20	0.40	0.30
mean *[I(f5, f2) + I(f4, f2)]*	0.14	0.24	0.19
mean *[I(f5, f6) + I(f4, f6)]*	0.15	0.13	0.14

3.2) mRMR (V-D):

Feature	*V*	*D*	*mRMR (V-D)*	*mRMR (V/ D)*
f1	0.50	0.23	**0.27**	2.17
f3	0.49	0.30	0.19	1.63
f2	0.40	0.19	0.21	2.11
f6	0.36	0.14	0.22	**2.57**

Finally, the 3 optimal selected features using Mutual Information Difference mRMR algorithm are: *f5*, *f4* and *f1*. With the Mutual Information Quotient mRMR algorithm, the 3 optimal features selected are: *f5*, *f4* and *f6*. In contrast, the normal filter-base selection algorithm declares *f5*, *f1* and *f3* as the best features based on maximum relevance condition only.

4.1.6 <u>Advantages of mRMR Algorithm:</u>

The mRMR feature selection technique has a number of advantages over other feature selection algorithms, these include:

- Efficient coverage of space: obviously, the mRMR as a type of feature selection that reduces space dimension. However, mRMR efficiently do the reduction by minimizing the redundancy and consequently consider more "valuable" feature instead of the redundant ones. This careful selection of independent feature guarantees comprehensive space coverage.

- Low computational cost: that is a consequent result of the dimension reduction.
- Better reliability: the mRMR improves classification performance and reliability from two aspects: increasing accuracy and reducing noise. Accuracy improvement is a result of effective selection of features.
- Easier and faster since it is a low dimensional problem, it is faster and easier than other algorithm that involves multivariate density or high dimensional space.

These advantages especially the efficient reduction of space dimensions and the reliability prompted us to choose the mRMR as the optimization algorithm to best select the features that would improve speech recognition performance specifically at low SNR where the recognition accuracy is very low. In our research, we define the range from -10 dB to +5 dB as a low SNR area of interest.

4.2 Our Proposed Approach

For our implementation we used the MATLAB environment. For this research, we focus on developing an Isolated Word Recognition (IWR) system for both Arabic and English languages. The utterances were modeled using HMM and the feature selecting optimization algorithm is based on the mRMR MID algorithm. TI-46 is the English standard database [63] used in the English part of our experiments. While the Arabic database is a collection from local volunteers and from a database supported from Acoustics Center of King Abdulaziz City for Science and Technology (KACST) in

Riyadh. Samples of noise data under different environments were collected from NOISEX-92 database [63]. Figures 4.1 through 4.4 represent the flow charts of our approach.

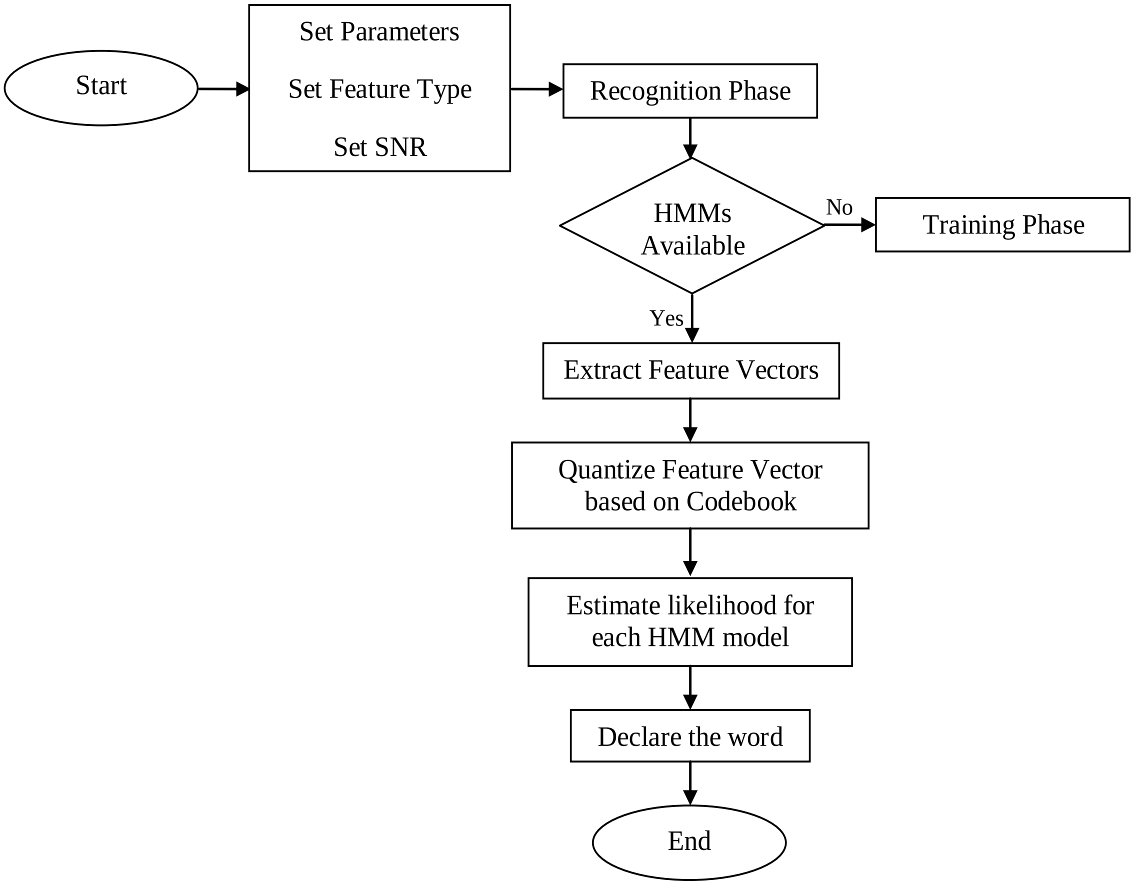

Figure 4.1: Flow Chart of Recognition Phase.

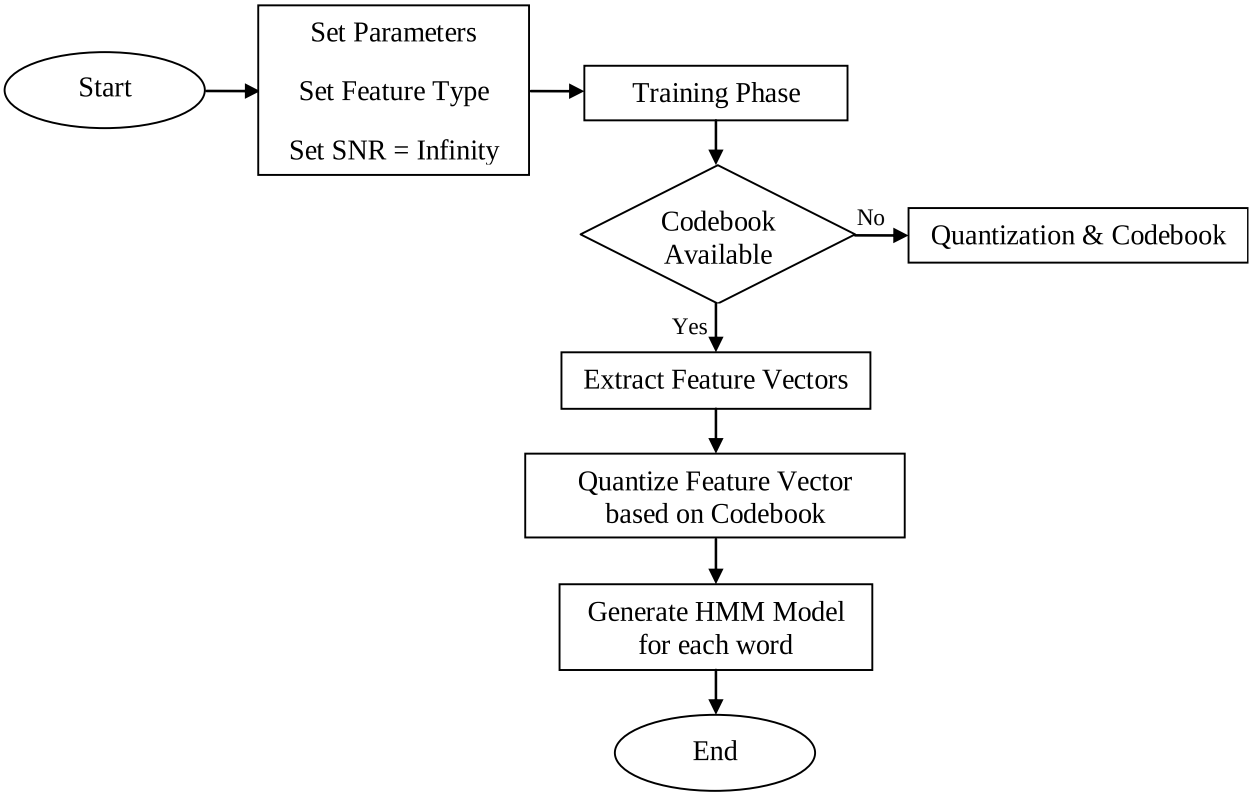

Figure 4.2: Flow Chart of Training Phase.

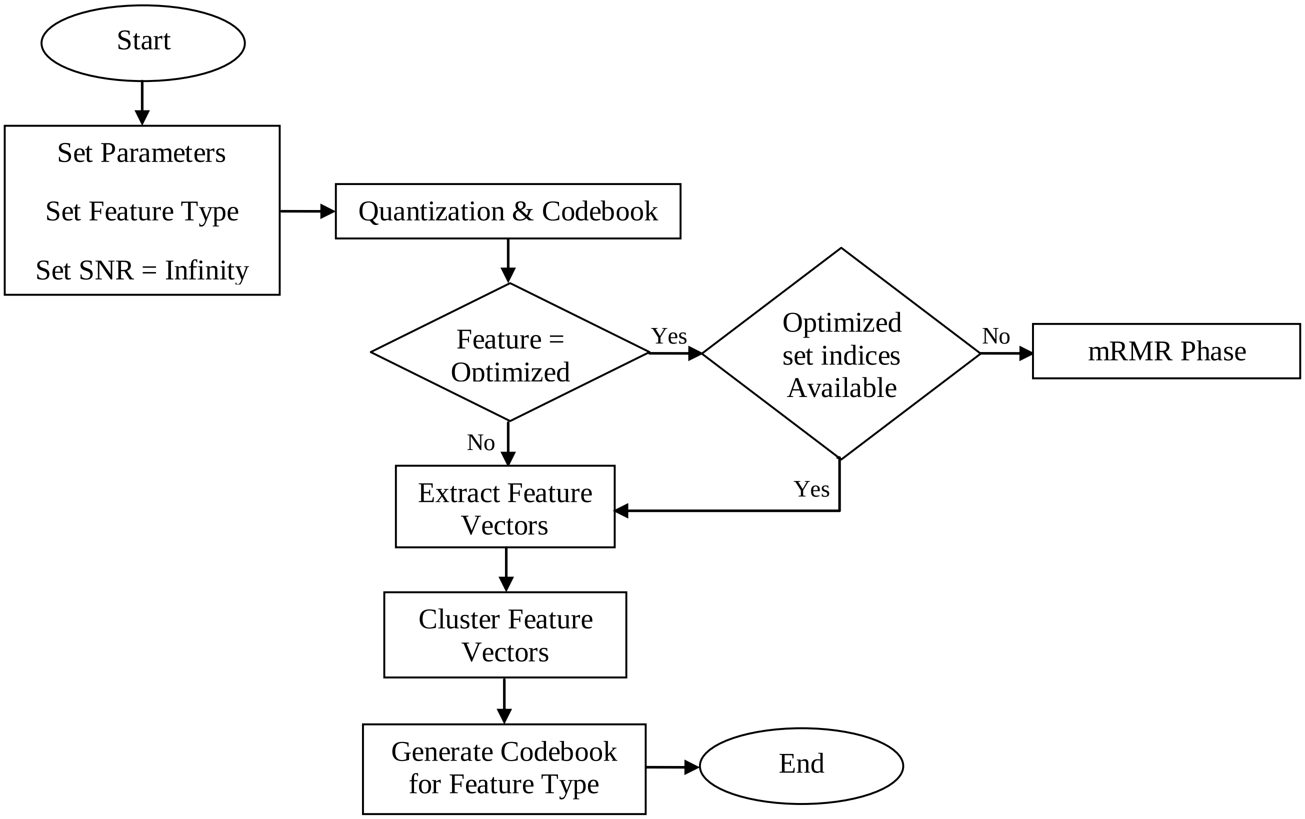

Figure 4.3: Flow Chart of Quantization & Codebook.

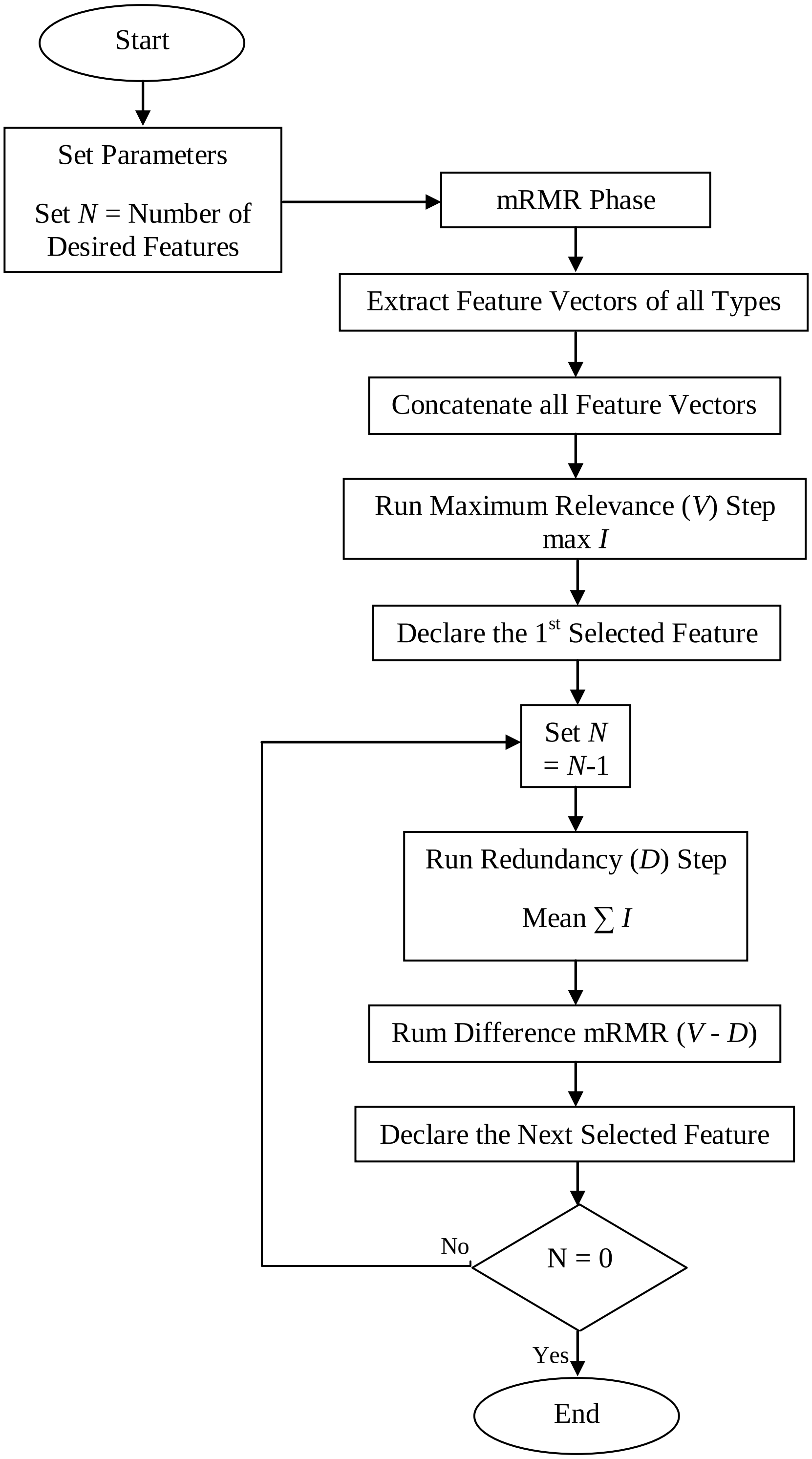
Start
Set Parameters
Set N = Number of Desired Features
mRMR Phase
Extract Feature Vectors of all Types
Concatenate all Feature Vectors
Run Maximum Relevance (V) Step max I
Declare the 1st Selected Feature
Set N = N-1
Run Redundancy (D) Step
Mean ∑ I
Rum Difference mRMR (V - D)
Declare the Next Selected Feature
N = 0
No
Yes
End

Figure 4.4: Flow Chart of mRMR Phase.

4.3 Outline of The Experimental Setup

Our proposed system is run in 6 main steps:

1. Preprocessing: First, the speech signal *s(n)* is digitized by sampling at a frequency of 10kHz. Then the resulted samples are segmented into frames of 320 samples/ frame or equivalently 32ms. The consecutive frames are overlapped with 80 samples. Then each frame is windowed by Hamming Window. Then the desired feature set is extracted in form of 16 feature vectors sequence for each word.
2. Feature extraction: The sequence of feature vectors are derived from the spectral analysis of the speech signal. In this work, 7 types of features were studied, 6 standard features and to the optimum one. The standard features are: LPC, LSP, RC, CC, Delta-CC and MFCC. The optimum feature is derived from the other 6 standard features.
3. Vector Quantization: The feature vectors are grouped into clusters of disjoint sets. Each observation vector is quantized into one of permissible sets of 16 scalar and discrete values. Then, we implemented the k-means vector quantization algorithm.
4. Training the HMM: Data for each digit is modeled into an m-state HMM. The HMM is trained using 10 observations for each digit. The model is trained under a very high SNR which is the ideal situation to generate the accurate reference models.
5. Testing/Recognition: Here, we calculated *P(y| M)*; the probability of getting the feature sequence *y* given model *M*. We then declared a digit whose model leads to the highest score. The computation of *P(y| M)* is based on forward/ backward recursions. The noises data used here are: White, Factory, Volvo and Pink.

The overall accuracy is determined for the range {-10 to +5} Signal to Noise Ratio (SNR). In our analysis, we segment the SNR range of interest into 2 parts:

1) The negative SNRs which cover the range from -10 dB where the signal power is 10 times less than the noise power to 0 dB where both powers are identical.
2) The positive SNRs that span from 0 dB to +5 dB at which the signal power is 5 times more than the noise power.

Recognition accuracy is evaluated every 5 dB step in the SNR in the concerned range for the four different types of noises mentioned above.

6. The Optimized Feature Set: Here, we used the MID optimization algorithm to select the mRMR features among 96 feature vectors (16 vectors per each of the 6 standard feature sets). The reference classes are the digits. Minimum redundancy is carried first by calculating the mutual information index between the classes and features. Then, maximum relevance is obtained by calculating the mutual information index between the features vectors themselves in the order of the previous step. Using this set of optimal feature, we calculated the recognition accuracy of 5 remaining set of utterances for the range {-10 to +5} SNR of four different types of noises: White, Factory, Volvo and Pink noises.

Finally, the whole experimental setup is repeated with Arabic digits. We will discuss in the next section our result of both Arabic and English languages. However, before presenting the experimental results, we will briefly discuss the different types of noises considered in our performance analysis.

TI46 Database:

The TI46 is a corpus of isolated spoken words collected at Texas Instruments (TI) in 1980. The words of this corpus were recorded in an isolated booth, using a cardoid dynamic microphone, positioned 2 inches from the speaker's mouth and out of his/her breathe stream. The TI46 database contains 46 words uttered 26 times from 16 speakers (8 males and 8 females).

Noise Database:

To cover as many types of noises as possible, we selected 4 types, these are:

White Noise: This noise is the most popular type of noise considered in the literature. The power spectrum density of white noise is constant over the whole frequency range (Figure 4.1).

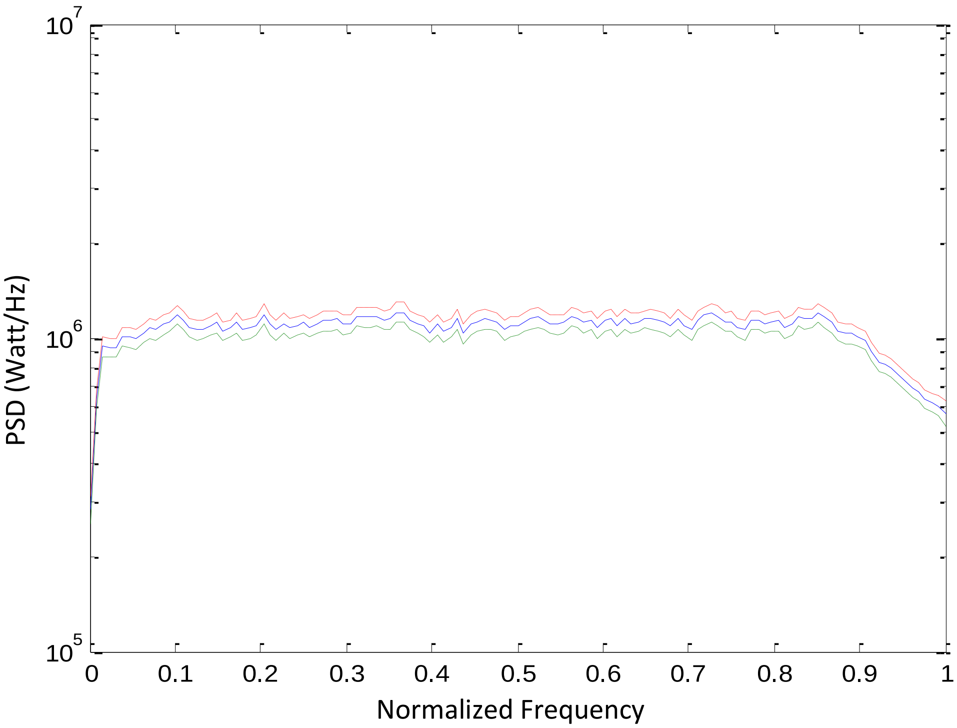

Figure 4.5: Power Spectral Density of White Noise.

Car Factory Noise: This noise was recorded in a car production hall. The power spectrum density of the factory noise decreases with frequency and also exhibits tones at low frequencies (Figure 4.2).

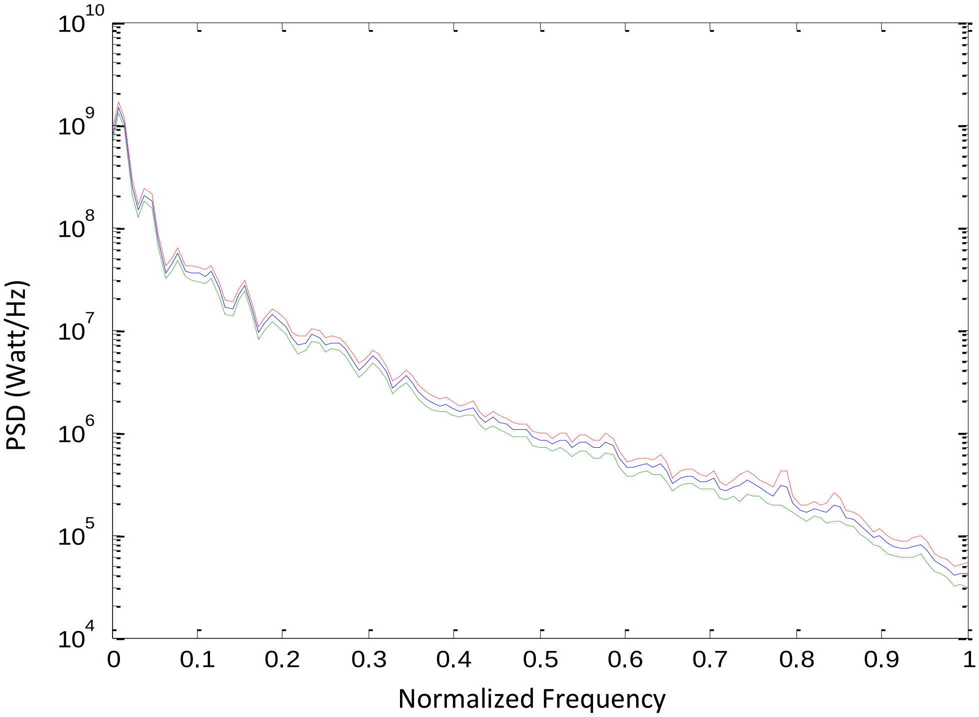

Figure 4.6: Power Spectral Density of Factory Noise.

Volvo Noise: This noise was recorder in a Volvo car at speed of 120 km/h, in the 4th gear, on an asphalt road, in rainy conditions. The power spectrum density of the Volvo noise decreases with frequency and exhibits very noticeable time domain impulses at low frequencies (Figure 4.3).

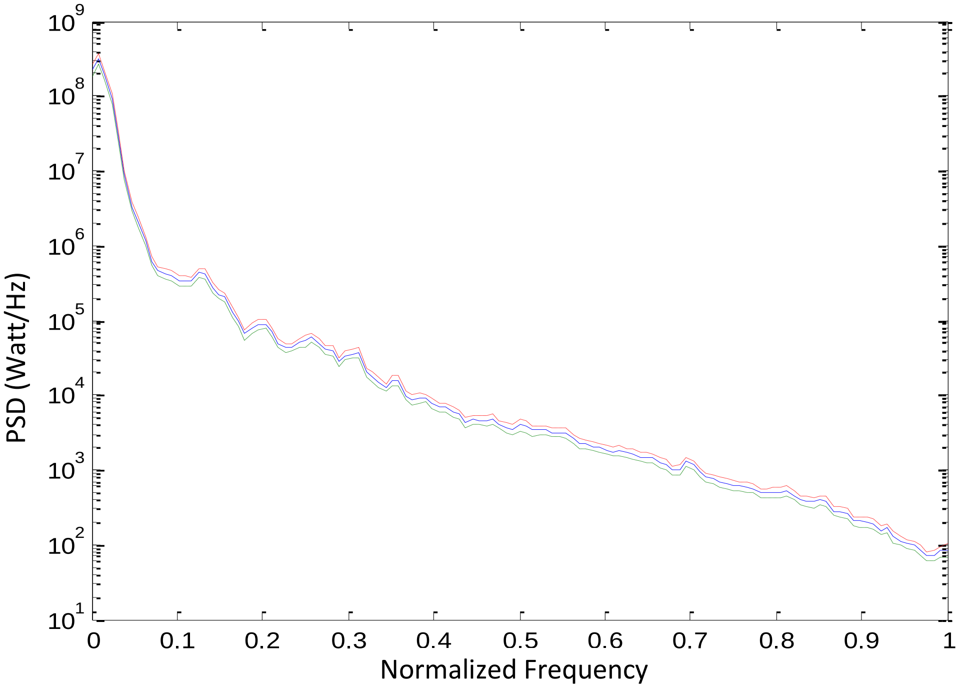

Figure 4.7: Power Spectral Density of Volvo Noise.

Pink Noise: This type of noise exhibits an equal energy per 1/3 octave. The power spectrum density of pink noise decreases 3dB per octave with frequency (Figure 4.4).

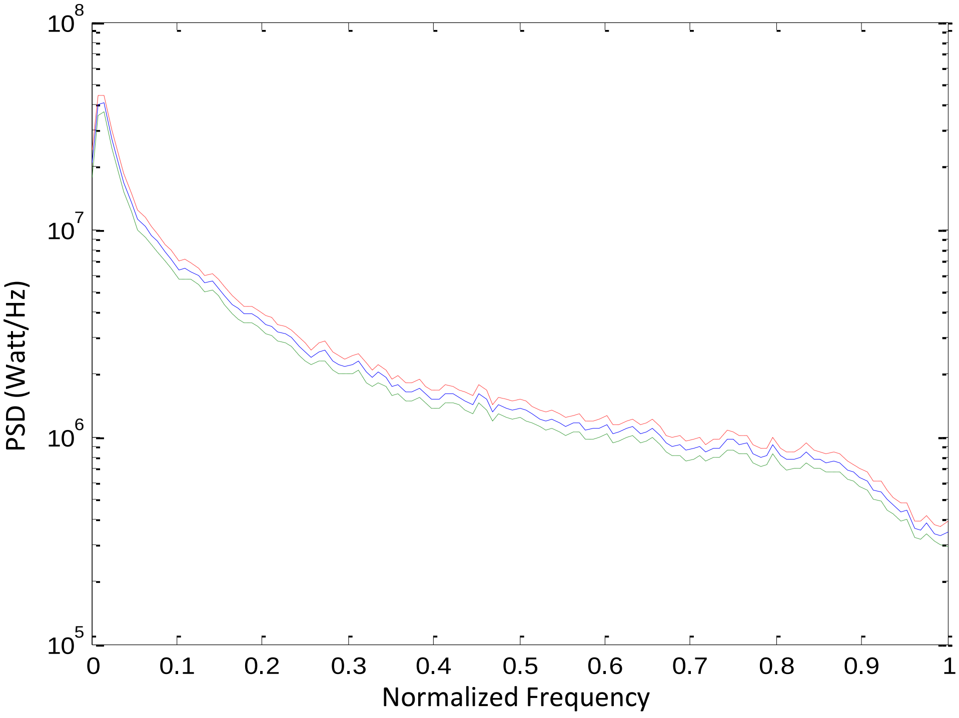

Figure 4.8: Power Spectral Density of Pink Noise.

4.4 Experimental Results

4.4.1Experimental Results for English Language Digits:

In our first experiment and assuming a noise free environment, we obtained a recognition accuracy of 98% for both the MFCC feature set and the Optimized feature set. The results emphasize the power of the MFCC coefficients in speech recognition. It is worth noting the Optimized feature set obtained using the mRMR algorithm consists of a large number of MFCC coefficients as shown in the table below:

Table 4.3: Optimized Feature Set for The English Language Numerals

1	MFCC01
2	LPC12
3	MFCC09
4	MFCC02
5	**MFCC13**
6	**MFCC08**
7	MFCC05
8	CC01
9	RF06
10	MFCC11
11	DeltaCC03

12	MFCC07
13	**MFCC06**
14	**MFCC15**
15	**LPC01**
16	**MFCC04**

The performance with English digits in terms of recognition accuracy with additive white noise for the standard and Optimized features; is shown in Table 4.4 and Figure 4.9.

The results show also that the MFCC coefficients dominate the other features over a wide range of SNR values. The Optimized feature set shows an improvement over the MFCC with an average of 15%.

Table 4.4: Recognition Accuracy in % for English Numerals with Additive White Noise

Case of Additive White Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	20	20	16	20	20	36	16
-5	20	20	20	28	32	40	8
0	24	32	24	24	36	48	12
5	36	28	48	44	52	76	24
10	64	56	72	60	80	92	12

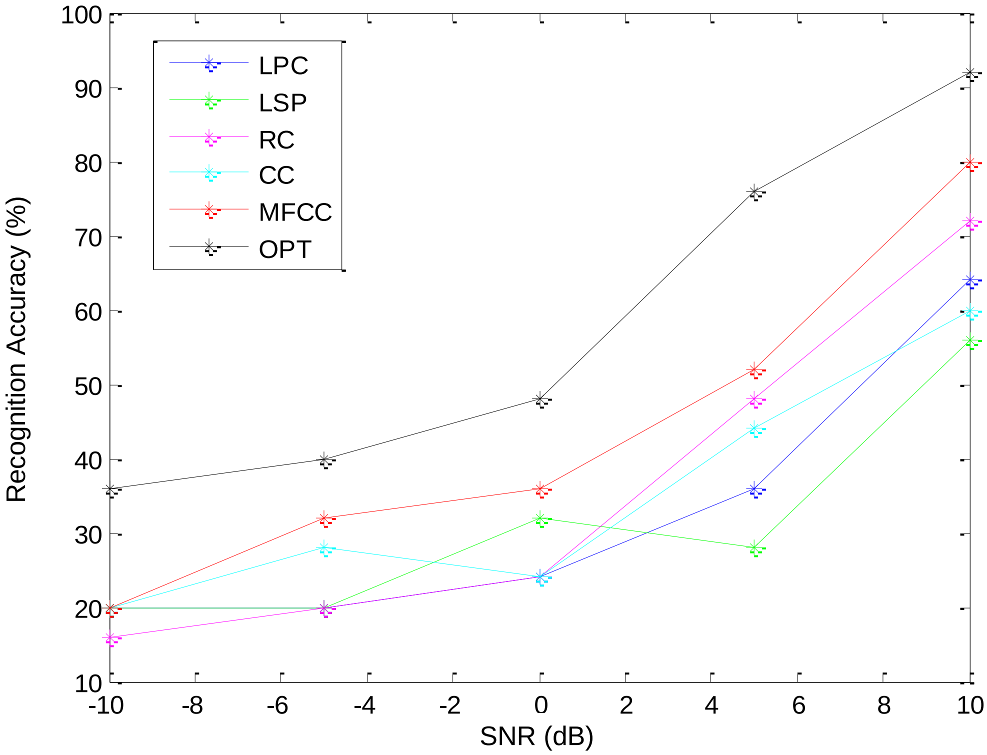

Figure 4.9: Results for English Numerals with White Noise.

The recognition accuracy performance in the case of additive factory noise; is shown in Table 4.5 and Figure 4.10.

Once again, the MFCC coefficients outperform other coefficients. The Optimized feature set improves the recognition accuracy for all SNRs. The important thing to notice is that the Optimized feature set exhibits a better improvement in accuracy at negative SNRs.

Table 4.5: Recognition Accuracy in % for English Numerals with Additive Factory Noise

Case of Additive Factory Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	12	14	10	8	14	44	30
-5	10	20	24	20	30	58	28

0	22	34	36	42	46	60	14
5	32	46	48	60	62	72	10
10	56	60	56	64	72	74	2

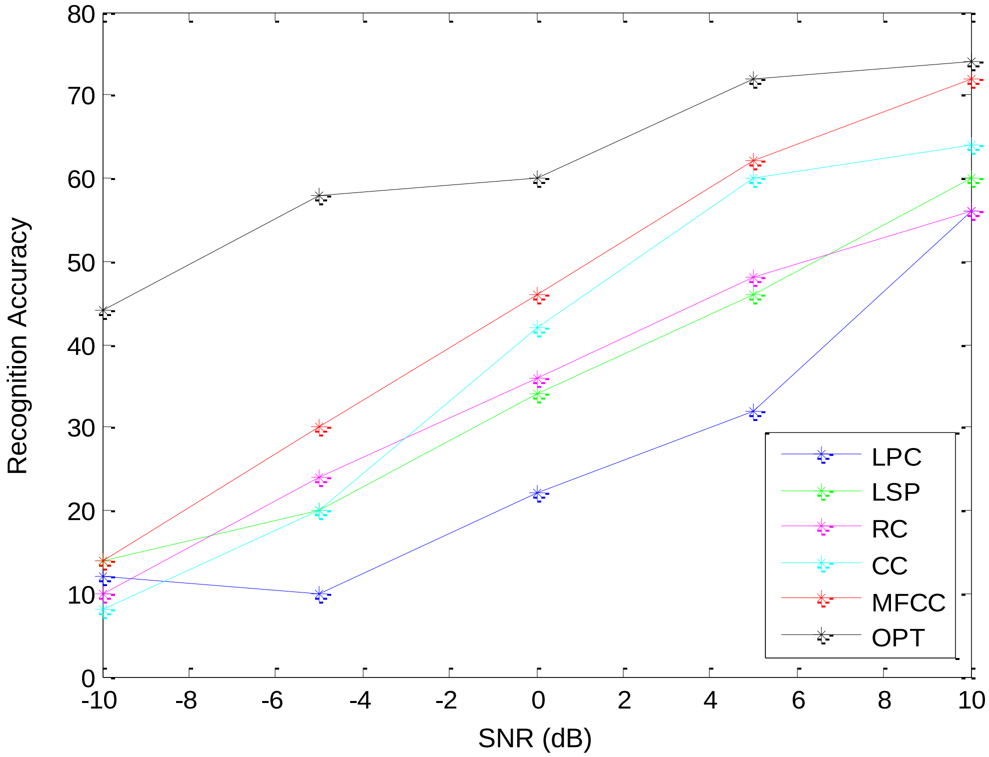

Figure 4.10: Results for English Numerals with Factory Noise.

The results shown in Figure 4.11 and Table 4.6 summarize our experiment for the case of additive Volvo noise. In this case, all features perform well in negative SNRs, but they progress slowly towards the positive region. While the MFCC coefficients dominate the other coefficients, the Optimized feature set shows an improvement in performance over the MFCC coefficients particularly above 0 dB SNRs.

Table 4.6: Recognition Accuracy in % for English Numerals with Additive Volvo Noise

Case of Additive Volvo Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	32	48	40	40	66	66	0
-5	44	64	52	60	68	70	2
0	42	62	50	64	70	78	8
5	50	66	58	66	70	80	10
10	50	66	60	70	74	84	10

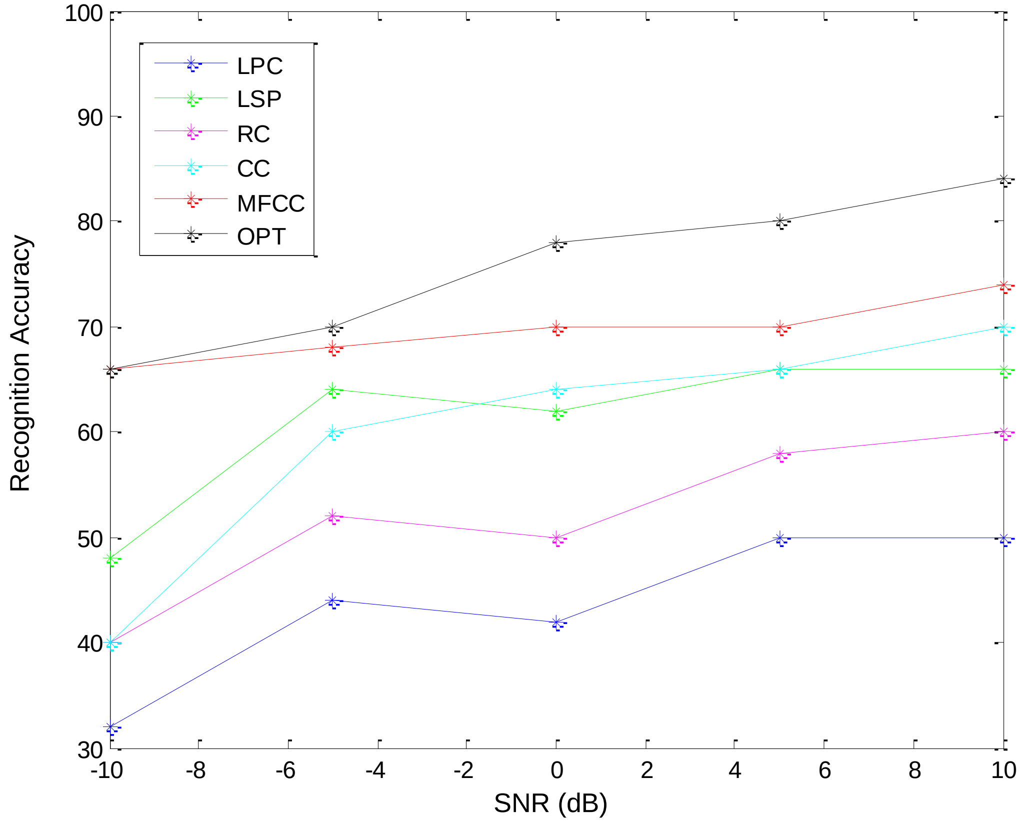

Figure 4.11: Results for English Numerals with Volvo Noise.

The case of additive pink noise was the most challenging one. In this case, the overall performance was low for all features between the range -10 dB and -5 dB SNRs. However, above -5 dB SNR, the performance of the MFCC features and the Optimized features are very similar (see Figure 4.12 and Table 4.7).

Table 4.7: Recognition Accuracy in % for English Numerals with Additive Pink Noise

Case of Additive Pink Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	16	14	20	16	10	20	10
-5	14	32	22	20	28	34	6
0	14	38	30	40	46	46	0
5	30	48	44	58	66	68	2
10	36	68	52	72	78	78	0

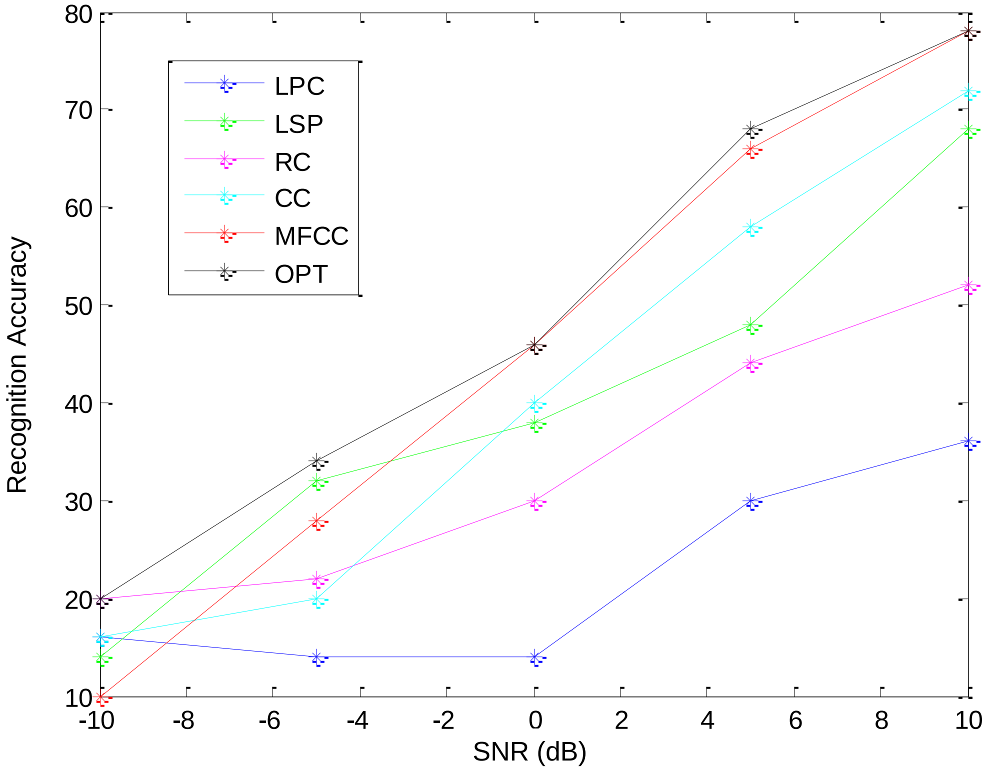

Figure 4.12: Results for English Numerals with Pink Noise.

4.4.2 Experimental Results for Arabic Language Digits:

Without noise, the performance reached was 92% in terms of recognition accuracy for both the MFCC feature set and the Optimized feature set. The Arabic Optimized feature set of size of 16 consists mainly of MFCC features as shown in the below table:

Table 4.8: Optimized Feature Set for The Arabic Language Numerals

1	**MFCC04**
2	DeltaCC02
3	MFCC10
4	MFCC03
5	**MFCC13**
6	MFCC14

7	MFCC11
8	**MFCC06**
9	**MFCC08**
10	CC04
11	**LPC01**
12	DeltaCC04
13	**MFCC15**
14	DeltaCC03
15	RF05
16	MFCC12

In our first experiment, we started by considering additive white noise. For this case, the Optimized feature set shows an improvement over the MFCC coefficients with an average of 10%. Note that this improvement is lower than the case of English language. This is due to the complicated structure of the Arabic language spectrum compared to English. The results are summarized in Figure 4.13 and Table 4.9.

Table 4.9: Recognition Accuracy in % for Arabic Numerals with Additive White Noise

Case of Additive White Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	26	18	26	20	32	44	12
-5	26	30	30	32	38	48	10

0	26	38	26	44	52	56	4
5	30	48	44	62	64	68	4
10	30	56	56	66	68	84	16

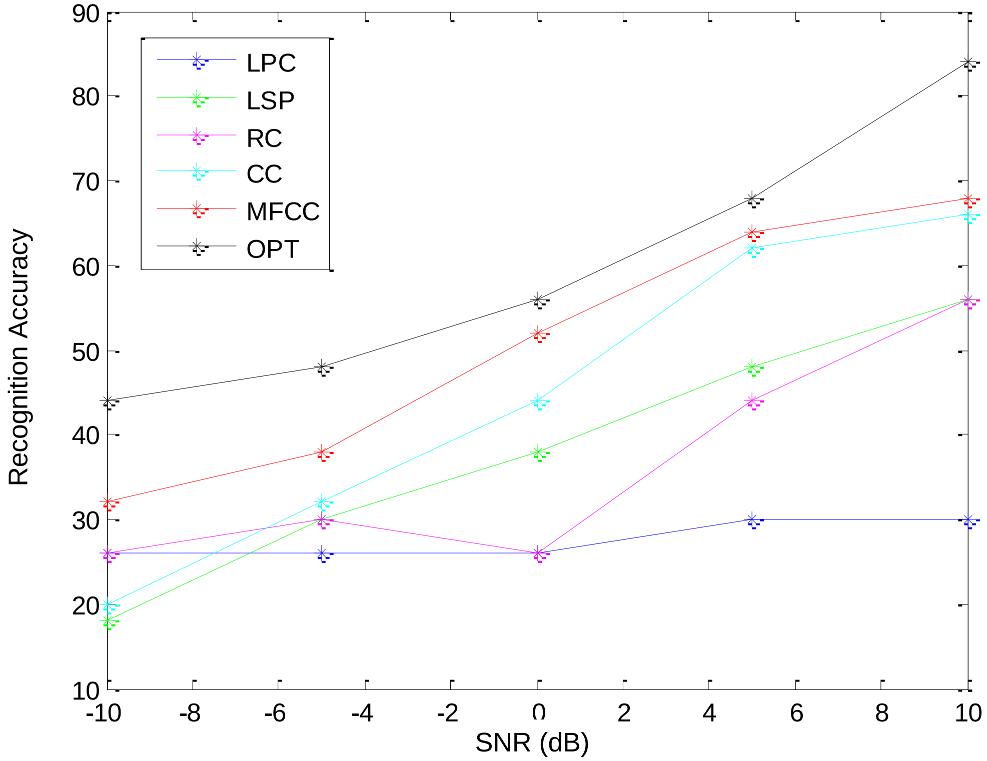

Figure 4.13: Results for Arabic Numerals with White Noise.

For the case of additive factory noise (Figure 4.14 and Table 4.10), the Optimized feature set gives an improvement in performance over MFCCs of about 11%. The major improvement appears at the negative range of SNRs.

Table 4.10: Recognition Accuracy in % for Arabic Numerals with Additive Factory Noise

Case of Additive Factory Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	14	22	18	22	20	30	10
-5	22	32	22	28	38	58	20
0	26	42	44	48	50	64	14
5	30	62	58	46	66	70	4
10	36	68	58	50	76	80	4

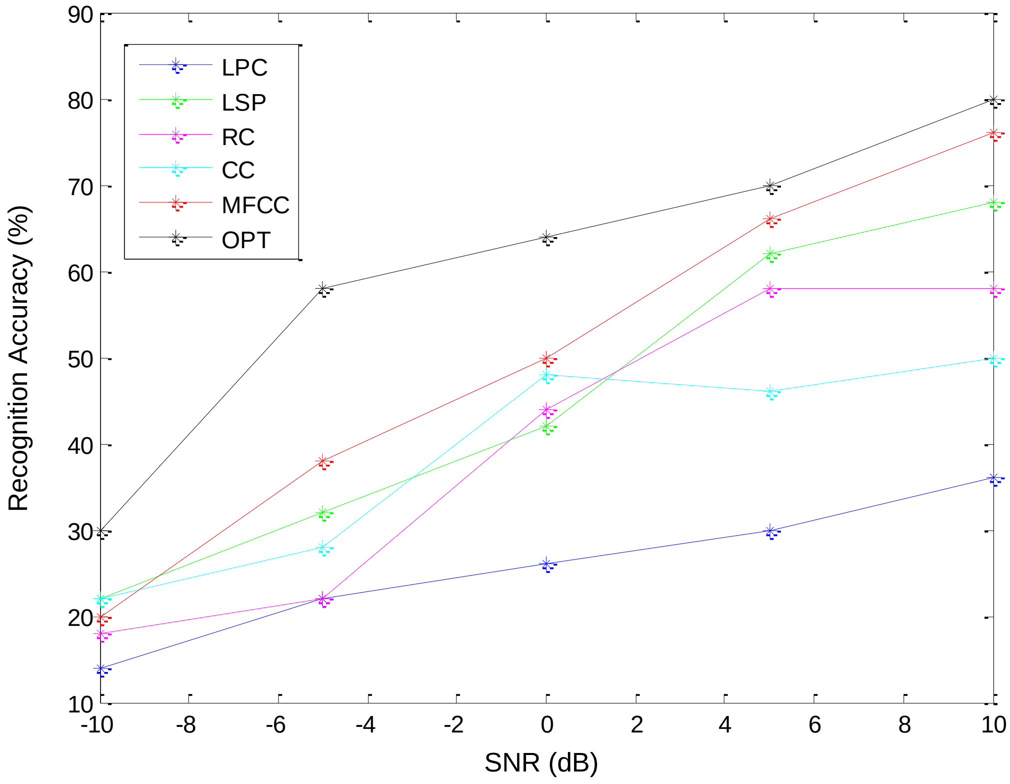

Figure 4.14: Results for Arabic Numerals with Factory Noise.

A reduced improvement is also noticed for the case of additive Volvo noise as can be seen in Table 4.11 and Figure 4.15. While all features perform well at very low SNRs, MFCC coefficients perform better overall SNRs. The improvement of the Optimized feature set over MFCC coefficients is noticeable below +5 dB SNRs with an overall average improvement of 7%.

Table 4.11: Recognition Accuracy in % for Arabic Numerals with Additive Volvo Noise

Case of Additive Volvo Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	34	40	42	48	50	64	14
-5	42	52	42	52	68	76	8
0	46	60	50	60	70	78	8
5	42	68	56	64	78	78	0
10	46	68	60	64	82	86	4

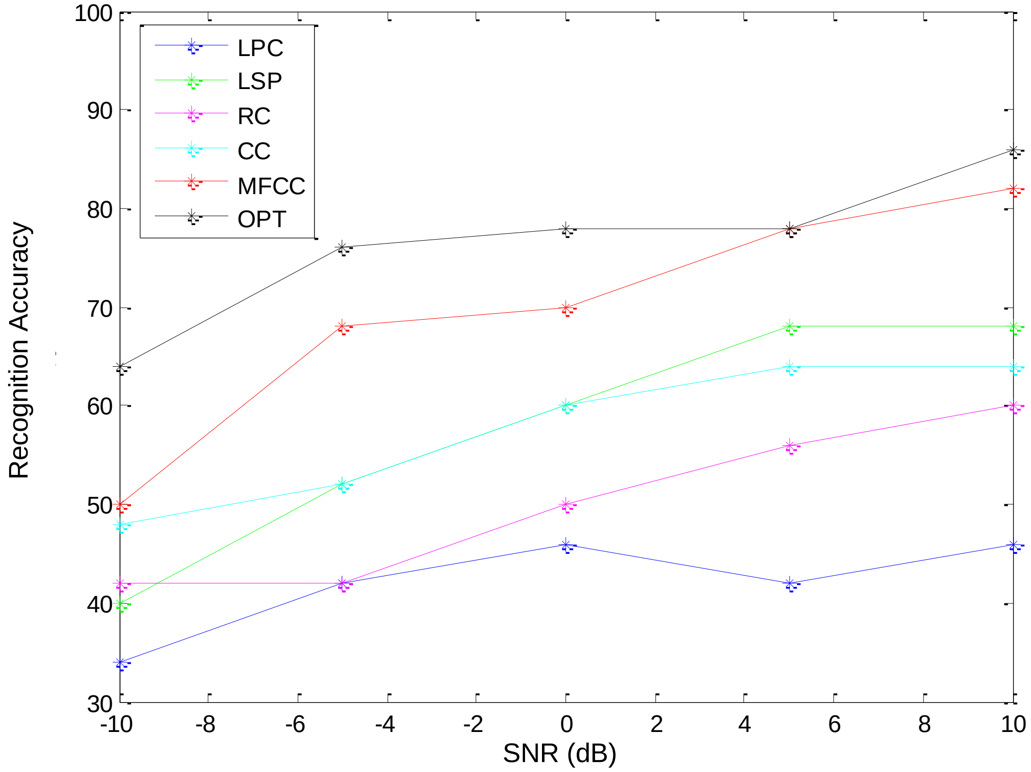

Figure 4.15: Results for Arabic Numerals with Volvo Noise.

The result we obtained in the case of additive pink noise indicates that the Optimized feature set and the MFCC feature set are comparable as shown in Table 4.12 and Figure 4.16.

Table 4.12: Recognition Accuracy in % for Arabic Numerals with Additive Pink Noise

Case of Additive Pink Noise							
Feat SNR	LPC (%)	LSP (%)	RFC (%)	CC/ DeltaCC (%)	MFCC (%)	Optimized (%)	Improvement in recognition accuracy (%) w.r.t best the feature set
-10	14	22	22	12	12	20	8
-5	12	22	22	16	28	40	12

0	12	34	20	30	40	44	4
5	12	42	26	34	58	54	-4
10	18	46	38	42	78	74	-4

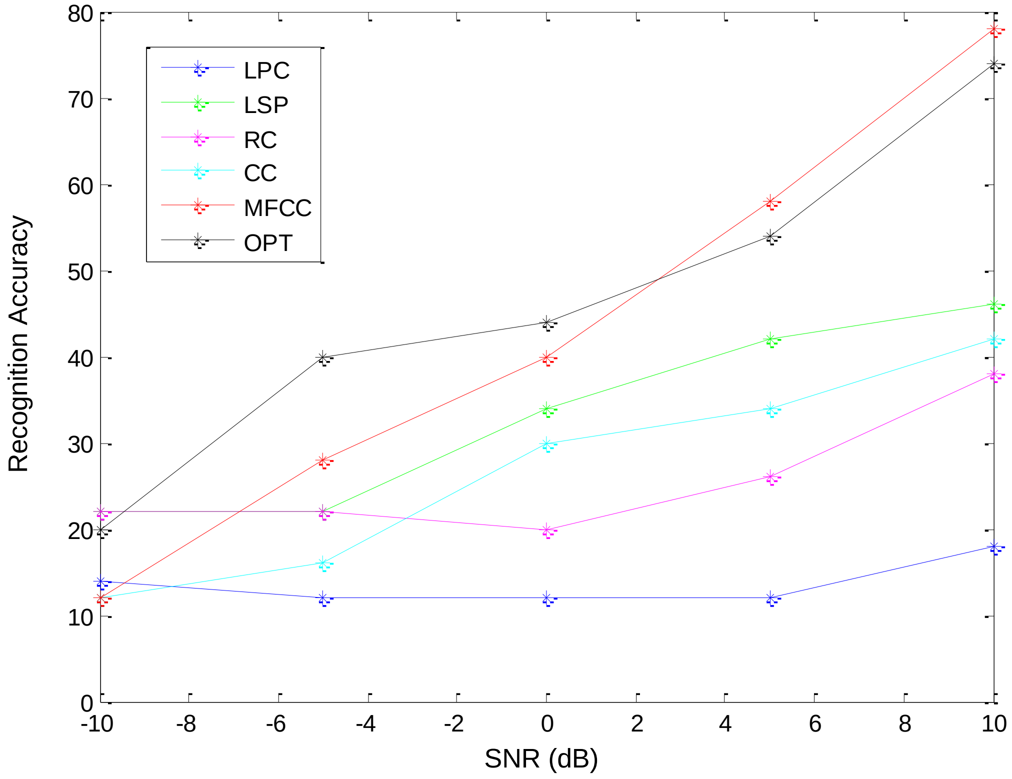

Figure 4.16: Results for Arabic Numerals with Pink Noise.

CHAPTER 5

CONCLUSION

5.1 Summary

In this work, we have discussed applying a new approach for Speech Recognition based on selecting optimal features and an HMM classifier. As introduction a background of the human speech communication process was given in Chapter 1. The importance and the objectives of this research were stated at the end of the introduction. Chapter 2 describes the concept of Speech Recognition and the implementation details of such systems. Feature Extraction is the entry process of speech recognition system upon which the overall system performance depends. Feature extraction and selection were explained also in this chapter. The Hidden Markov Model as a popular statistical modeling algorithm used in speech recognition was also discussed in this chapter. In Chapter 3, we briefly described some basis of information theory focusing on the concept of mutual information. In this chapter, we also introduced the concept of minimum Redundancy, Maximum Relevance (mRMR) algorithm for feature selection. In Chapter 4, we described the algorithm setup, implementation parameters, and the detailed results of experiments. In summary, the proposed algorithm of minimum Redundancy, Maximum Relevance selects the optimum features that perform better at low SNR for a range of noises. The majority of the optimized selected features by the mRMR were MFCC coefficients. In 16-features English experiment, the mRMR selects 11 MFCCs, 2 LPCs, 1 RFC, 1 CC and 1 Delta-CC features. The Optimum features exhibit an improvement at low SNR (-10 to +10) for white, factory, and Volvo noises. The average improvements over the studied low SNR range are 15% and 17% for white and factory noises respectively. This improvement is reduced to about 4% for Volvo noise case. On the other hand, the Optimum feature performs as well as the MFCC in a pink noise. In 16-features Arabic experiment, the mRMR selects 10 MFCCs, 3 Delta-CCs, 1 LPC, 1 RFC and 1 CC features. The Arabic Optimum feature set provides the same performance results as English for all types of tested noises. The average improvements achieved for white and factory noises are 10% and 11% respectively. For Volvo noise case, the average improvement is 7% which is the only case that Arabic language results overcome the English language results. Compared to the results discussed by Korba et al. [106], our algorithm gives a substantial improvement at low SNR for different type of noises.

5.2 Future Research Directions

The work we started in this thesis is the first step in a new direction of research for enhancing speech recognition algorithms. To further enhance this work, we list below a number of ideas:

1. A Parameterized mRMR algorithm: one could introduce a weighted factor for each part of mRMR equation and study the performance. Modified mRMR optimization algorithm becomes:

$$\max\left(\alpha V-\beta D\right), \text{for additive combination} \tag{5.1}$$

$$\max\left(\frac{\alpha V}{\beta D}\right), \text{for multiplicative combination} \qquad (5.2)$$

where α and β are constant cumulative values sum up to 1. The objective of the parameterized algorithm is to find the values of α and β that provide better recognition performance.

2. Changing the optimization technique used for maximum relevance part is another area for research. The summary of this idea is to combine another optimization technique like gradient descent or genetic algorithm for maximum relevance part with the mutual information technique for minimum redundancy.

3. An interesting research is to evaluate the speech recognition performance using mRMR with other classifier e.g. Bayesian and Neural Network. In our research, we used HMM as a common and successful classifier for speech recognition systems. In the proposed research, one can concludes if mRMR can boost up the performance of other classifiers to the level of HMM or better.

4. Acknowledged outcome of mRMR algorithm are based on precise measurements for relevance and redundancy. The type of measurements is another field for research. Studying how different statistical measuring function such as correlation function impacts the efficiency of mRMR.
5. Another area of research is the study of the relationship between the performance improvement and the size of the selected feature. Our algorithm chooses the first 16 optimized features. This number is chosen as the same number used for standard features. The suggested study will find the optimum mRMR feature set size that provides the best performance.
6. Another study is to consider the performance of the Quotient mRMR. Along with that is the performance study of the combination of Differential mRMR and Quotient mRMR. The aim of such a research is to evaluate different performance of different types of mRMR and to find the optimal cost function.

7. Finally, evaluating the performance of mRMR algorithm among different standard and new standalone features is a nice topic for research. In this research, we selected the most popular standard features from which the selected optimized feature score a recognized improvement. The proposed study of different features will help to find a criteria for features on which mRMR can operate well.

REFERENCES

[1] J. R. Deller, J. H. L. Hansen and J. G. Proakis, "Discrete-Time Processing of Speech Signals*", IEEE Press,* 1993.

[2] B. Gold and N. Morgan, "Speech and Audio Signal Processing", *John Wiley & Sons, Inc,* 1999.

[3] L. R. Rabiner and B. Juang, "Fundamental of Speech Recognition", *Prentice-Hall,* 1993.

[4] F. J. Owens, "Signal Processing of Speech", *The MacMillan Press Ltd,* 1993.

[5] C.E. Shannon, "A mathematical Theory of Communication", *Bell System Technical Journal*, vol. 28, pp 379-423, 623-656, 1948.

[6] T. M. Cover and J. A. Thomas, "Elements of Information Theory", 2nd Edition, *John Wiley & Sons,* 2006.

[7] Y. Y. Yao, "Entropy Measures, Maximum Entropy Principle and Emerging Applications", *Springer*, pp. 115-136, 2003.

[8] Guiasu and Silviu, "Information Theory with Applications", *McGraw-Hill*, 1977.

[9] J. Proakis and D. Manolakis, "Digital Signal Processing – Principles, Algorithms, and Applications", 3rd Edition, *Prentice-Hall Inc*, 1996.

[10] L. Zhao and Z. Han, "Speech Recognition System Based on Integrating Feature and HMM", *International Conference on Measuring Technology and Mechatronics Automation*, vol. 3, pp.449-452, 2010.

[11] F. A. Elmisery, A. H. Khalil, A. E. Salama and H. F. Hammed, “A FPGA-Based - HMM For A Discrete Arabic Speech Recognition System”, *ICM*, 2003.

[12] M. A. Mokhtar and A. Z. El-Abddin, “A Model For The Acoustic Phonetic Structure Of Arabic Language Using A Single Ergodic Hidden Markov Model”, *Institute of Electrical and Electronics Engineers (IEEE)*.

[13] M. Thomae, G. Ruske and T. Pfau, “A New Approach to Discriminative Feature Extraction Using Model Transformation”, *Institute of Electrical and Electronics Engineers (IEEE),* 2000.

[14] M. Shoaib, F. Rasheed, J. Akhtar. M. Awais, S. Masud and S. Shamai1, “A Novel Approach to Increase the Robustness of Speaker Independent Arabic Speech Recognition”, *Proceedings IEEE International Multi-topic Conference (INMlC)*, 2003.

[15] T. Nitta, “A Novel Feature-Extraction for Speech Recognition Based on Multiple Acoustic-Feature Planes”, *Institute of Electrical and Electronics Engineers (IEEE)*, 1998.

[16] H. Bahi and M. Sellami, “Combination of Vector Quantization and Hidden Markov Models for Arabic Speech Recognition”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2001.

[17] I. Shafran and K. Hall, “Corrective Models for Speech Recognition of Inflected Languages”, *Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2006.

[18] S. Dharanipragada, “Feature Extraction for Robust Speech Recognition”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2002.

[19] T. Nitta, “Feature Extraction for Speech Recognition Based on Orthogonal Acoustic-Feature Planes and LDA”, *Institute of Electrical and Electronics Engineers (IEEE)*, 1999.

[20] R. Sarikaya, Y. Gao and G. Saon, “Fractional Fourier Transform Features Speech Reconnection”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2004.

[21] S. A. Selouani and D. O’Shaughnessy, “Hybrid Architectures for Complex Phonetic Features Classification: A Unified Approach”, *IEEE International Symposium on Signal Processing and its Applications (ISSPA)*, 2001.

[22] J. Hai and E. M. Joo, “Improved Linear Predictive Coding Method for Speech Recognition”, *IEEE International Conference on Information and Communications Systems_ Pacific-rim Conference on Multimedia (ICICS-PCM)*, 2003.

[23] H. Satori, M. Harti and N. Chenfour, “Introduction to Arabic Speech Recognition Using CMU Sphinx System*”, Baywood Publishing Company Inc,* 2007.

[24] M. Padmanabhan and S. Dharanip, “Maximizing Information Content in Feature Extraction”, *IEEE Transaction on Speech and Audio Processing,* vol. 13, No. 4, 2005.

[25] A. Klautuu, “Mining Speech: Automatic Selection of Heterogeneous Features Using Boosting”, *Institute of Electrical and Electronics Engineers (IEEE),* 2003.

[26] S. J. Russell, P. Norvig, “Artificial Intelligence: A Modern Approach”, *Pearson Education*, 2003.

[27] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schone, R. Schwartz and D. Vergyri, “Novel Approaches to Arabic Speech Recognition”, *Institute of Electrical and Electronics Engineers (IEEE),* 2003.

[28] C. S. Yip, S. H. Leung and K. K. Chu, “Optimal Root Cepstral Analysis for Speech Recognition”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2002.

[29] J. Hung, “Optimization of Filter-Bank to Improve the Extraction of MFCC Features in Speech Recognition”, *International Symposium on Intelligent Multimedia, Voice and Speech Processing*, 2004.

[30] G. L. Sarudu, T. Nagarujan, and H. A. Murthy, “Multiple Frame Size and Multiple Frame Rate Feature Extraction for Speech Recognition”, *IEEE International Conference on Signal Processing & Communications (SPCOM),* 2004.

[31] L. R. Rabiner, “A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of Institute of Electrical and Electronics Engineers (IEEE),* vol. 77, pp: 257-286, 1989.

[32] B. Gold and N. Morgan, “Speech and Audio Signal Processing: Processing and Perception of Speech and Music”, *John Wiley & Sons, Inc,* 1st edition, 1999.

[33] A. Biem, “Optimizing Features and Models Using the Minimum Classification Error Criterion”, *Institute of Electrical and Electronics Engineers (IEEE),* 2003.

[34] B. Gajic and K. Puliwal, “Robust Speech Recognition Using Features Based on Zero Crossings with Peak Amplitudes”, *Institute of Electrical and Electronics Engineers (IEEE),* 2003.

[35] N. Wada, N. Hayasaka, S. Yoshizawa and Y. Miyanaga, “Robust Speech Recognition with Feature Extraction Using Combined Method of RSF and DRA”, *International Symposium on Communications and Information Technologies (ISClT),* 2004.

[36] L. R. Rabiner and S. E. Levenson, “Isolated and Connected Word Recognition – Theory and Selected Application”, *IEEE Transaction on Communication,* vol. No. 5, pp.621-659, 1981.

[37] Z. Rakaz, N. J. Ibrahim, M. Y. Idna Idris, E. M. Tamil, M. Yakub, Z. M. Yousef and N. N. AbdulRahman, “Quranic Verse Recitation Recognition Module for

Support in j-QAF Learning: A Review", *International Journal of Computer Science and Network Security (IJCSNS),* vol.8 No.8, 2008.

[38] Z. Rakaz, N. J. Ibrahim, M. Y. Idna Idris, E. M. Tamil, M. Yakub, Z. M. Yousef and N. N. AbdulRahman, "Quranic Verse Recitation Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC)". *IEEE International Colloquium on Signal Processing and its Application (CSPA)*, 2008.

[39] H. Tabbal, W. El-Falou and B. Monla, "Analysis and implementation of a "Quranic" verses delimitation system in audio files using speech recognition techniques", *Institute of Electrical and Electronics Engineers (IEEE)*, 2006.

[40] H. Tabbal, W. El-Falou and B. Monla, "Analysis and Implementation of an Automated Delimiter of "Quranic" Verses in Audio Files using Speech Recognition Techniques", *Open Access Database www.i-techonline.com*, 2007.

[41] Y. Alhaj, M. Alghamdi, M. Alkanhal and A. Alansari, "Quranic Speech Recognition", *International Computing Conference in Arabic. Alhammamat, Tunisia*, 2010.

[42] M. M. Azmi and H. Tolba, "Syllable-Based Automatic Arabic Speech Recognition in Noisy Environment", *Institute of Electrical and Electronics Engineers (IEEE)*, 2008.

[43] M. Siafarikas, I. M., T. Ganchev and N. Fakotakis, *"*Speech Recognition using Wavelet Packet Features*", Journal of Wavelet Theory and Applications*, vol. 2 pp. 41–59, 2008.

[44] L. R. Rabiner and B. H. Juang, “An Introduction to Hidden Markov Models”, *IEEE Acoustics Speech and Signal Processing (ASSP) Magazine*, pp: 4-16, 1986.

[45] L. A. Raming and R.L. Ringal, “Effects of Physiological Ageing on Selected Acoustic Characteristic of Voice”, *Journal of Speech and Hearing Research*, vol. 26, no. 1, pp. 22-30, 1985.

[46] S. Furui, F. Itakura, and S. Saito, “Talker Recognition by Long Time Averaged Speech Spectrum”, *Electronic and Communications in Japan*, vol. 55-A, No. 10, pp.54-61, 1972.

[47] J. Fokes, Z. S. Bond, and M. Steinberg, “Patterns of English Word Stress by Native and Non- native Speakers”, *Proceedings of the Tenth Annual International Congress of Phonetic Science*, pp. 682-686, 1983.

[48] M. Yeou, M. Embarki and S. Al-Maqtari, “Contrastive Focus and F0 Patterns in Three Arabic Dialects”, *New Papers on French linguistics*, pp.317-326, 2007.

[49] C. Lee and B. Juang, “A Survey on Automatic Speech Recognition with an Illustrative Example on Continuous Speech Recognition of Mandarin”, *Computational Linguistics and Chinese Language Processing*, vol. 1, no.1, pp. 01-36, 1996.

[50] D. Colton, “Automatic Speech Recognition Tutorial”, 2003.

[51] C. Ding and H. Peng, “Minimum Redundancy Feature Selection From microarray Gene Expression Data”, *Journal of Bioinformatics and Computational Biology,* vol. 3, no. 2, pp. 523–529, 2005.

[52] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection”, *Journal of Machine Learning Research*, 2003.

[53] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[54] R. Cilibrasi, and P. Vitányi, "Clustering by compression". *IEEE Transactions on Information Theory,* vol. 51, no.4, pp. 1523–1545, 2005.

[55] Y. Saeys, I. Inza and P. Larranaga, “A review of feature selection techniques in bioinformatics”, *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[56] A. Zolnay, R. Schluter and H. Ney, “Acoustic Feature Combination for Robust Speech Recognition”, *International Conference on Acoustics Speech and Signal Processing (IEEE- ICASSP),* pp 457-460, 2005.

[57] R.G. Gallager, “Information Theory and Reliable Communication”, *John Wiley & Sons*, 1968.

[58] Y. Peng, W. Li, and Y. Liu, “A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification”, *Published online*, 2007.

[59] Y. Lu, I. Cohen, X. Zhou and Q. Tian, “Feature Selection Using Principal Feature Analysis”, *Association of Computing Machinery annual conference on Multimedia (ACM Multimedia),* 2007.

[60] J. Jeong, J.C. Gore, and B.S. Peterson, “Mutual Information Analysis of the EEG in Patients with Alzheimer’s Disease”, *Clinical Neurophysiology*, vol. 112, no. 5, pp. 827-835, 2001.

[61] P. Viola, and W.M. Wells, “Alignment by maximization of mutual information”, *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137-154, 1997.

[62] Y. Normandin, R. Cardin, and R. De Mori, “High-performance connected digit recognition using maximum mutual information”, *IEEE Transaction Speech and Audio Processing*, vol. 2, no. 2, pp. 299-311, 1994.

[63] The Linguistic Data Consortium site for TI46 corpus database: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S9.
And The Signal Processing Information Base (SPIB) site for NOISEX-92 database: http://spib.rice.edu/spib/select_noise.html.

[64] S. Kaza, Y. Wang, H. Chen, “Enhancing border security: Mutual information analysis to identify suspect vehicles”, *Decision Support Systems*, vol. 43, pp. 199-210, 2007.

[65] M. Alghamdi, F. Alharqan, M. Alkanhal, A. Alkhayri and M. Addusooqiee, “Saudi Accented Arabic Voice Bank”, *Computer and Electronics Research Institute, King Abdulaziz City for Science and Technology*, (in Arabic), 2003.

[66] M. Alghamdi, A. Alhumayid and M. Addusooqee, “Arabic Sound Database: Sentences”, *Computer and Electronics Research Institute, King Abdulaziz City for Science and Technology*, (in Arabic), 2003.

[67] A. Papoulis. “Probability, Random Variables, and Stochastic Processes”, *McGraw-Hill*, 2nd Edition, chapter 15, 1984.

[68] A. S. Ribeiro, S. A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. Socolar, "Mutual Information in Random Boolean models of regulatory networks", *Physical Review E*, vol.77, no. 1, 2008.

[69] A. Khodaei, H. Lee, F. Banaei-Kashani, C. Shahabi and I. Ershaghi, “A Mutual Information-Based Metric for Identification of Nonlinear Injector Producer Relationships in Waterfloods”, *Society of Petroleum Engineers (SPE)*, 2009.

[70] J. P. W. Pluim, J. B. A. Maintz and M. A. Viergever, “Mutual Information Based Registration of Medical Images: a Survey”, *IEEE Transactions on Medical Imaging*, pp. 986-1004, 2003.

[71] Z. Yana, Z. Wanga, H. Xieb, “The Application of Mutual Information-based Feature Selection and Fuzzy LS-SVM-based Classifier in Motion Classification”, *Computer Methods and Programs in Biomedicine*, vol. 90, pp. 275-284, 2008.

[72] G. Choueiter, D. Povey, S. Chen, and G. Zweig, "Morpheme-based Language Modeling for Arabic LVCSR", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[73] I. Titze, "Psychological and Acoustic Difference between Male and Female Voices", *Journal of Acoustic Society of America*, vol. 85, no. 4 pp.1699-1707, 1989.

[74] S. Dharanipragada, "Feature Extraction for Robust Speech Recognition", *Institute of Electrical and Electronics Engineers (IEEE)*, 2002.

[75] R. Schluter and H. Ney, "Using Phase Spectrum Information for Improved Speech Recognition Performance", *Institute of Electrical and Electronics Engineers (IEEE)*, 2001.

[76] 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, 2002.

[77] D. Vergyri, K. Kirchhoff, K. Duh and A. Stolcke, "Morphology-based Language Modeling for Arabic Speech Recognition", *International Speech Communication Association (InterSpeech)*, pp. 2245-2248, 2004.

[78] mRMR website: http://research.janelia.org/peng/proj/mRMR

[79] D. Vergyri and K. Kirchhoff, “Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition”, *Association for Computational Linguistics*, pp. 66–73, 2004.

[80] M. Al-Zabibi, “An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition”, *The British Library in Association with UMI*, 1990.

[81] Muhammad, “Alaswaat Alaghawaiyah”, *Daar Alfalah, Jordan*, (in Arabic), 1990.

[82] H. Hammady, O. Badawy, S. Abdou and M. Rashwan, “An HMM System for Recognizing Articulation Features for Arabic Phones”, *Institute of Electrical and Electronics Engineers (IEEE)*, 2008.

[83] S. Lemmetty, “Review of Speech Synthesis Technology”, *Master Thesis in Helsinki University of Technology*, *Finland*, 1999.

[84] Y.A. El-Imam, “An Unrestricted Vocabulary Arabic Speech Synthesis System”, *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, no. 12, pp.1829–1845, 1989.

[85] G. Pullum, W. Ladusaw, “Phonetic Symbol Guide”, *The University of Chicago Press*, 1996.

[86] Y. A. Alotaibi, “Investigating Spoken Arabic Digits in Speech Recognition Setting”, *Information and Computer Science*, pp.173, 115, 2005.

[87] C. Mokbel, M. Alghamdi and M. Mrayati, “Arabic Language Resources and Tools for Speech and Natural Language: KACST and Balamand”, *International Conference on Arabic Language Resources & Tools*, Cairo, Egypt, 2009.

[88] D. Liu, X. Wang, J. Zhang and X. Huang, “*Feature Extraction Using Mel Frequency Cepstral Coefficients for Hyperspectral Images Classification*”, *Journal of the Optical Society of America*, 2010.

[89] H. Almuhtaseb, M. Elshafei and M. Alghamdi, “Arabic Broadcast News Transcription System”, *International Journal of Speech Technology,* 2009.

[90] A. Araúzo-Azofra, J. M. Benítez, “Empirical Study of Feature Selection Methods in Classification”, *Proceedings of the 2008 8th International Conference on Hybrid Intelligent Systems*, pp. 584-589, 2008.

[91] D. J. Mashao, Y. Gotoh, and H. F. Silverman, “Analysis of LPC/DFT Features for an HMM-based Alpha-digit Recognizer,” *IEEE Signal Processing Lett*er, vol. 3, pp. 103–106, 1996.

[92] L. Gu and R. Liu, “The Application of Optimization in Feature Extraction of Speech Recognition,” *International Conference on Signal Processing*, pp. 745–748, 1996.

[93] Y. A. Alotaibi, M. Alghamdi and F. Alotaiby, “Speech Recognition System Based on a Telephony Speech Corpus”, *International Conference on Image and Signal Processing*, 2010.

[94] J. Zhou, and H. Peng, "Automatic Recognition and Annotation of Gene Expression Patterns of Fly Embryos," *Bioinformatics,* vol. 23, no. 5, pp. 589-596, 2007.

[95] L. R. Rabiner and R. W. Schafer, "Introduction to Digital Speech Processing", *Foundations and Trends in Signal Processing*, vol. 1, no. 1-2, pp 1-194, 2007.

[96] Constance Holden, "The Origin of Speech", *Science*, vol. 303, no. 5662, pp. 1316 – 1319, 2004.

[97] E. Choi, D. Hyun and C. Lee, "Optimizing Feature Extraction for English Word Recognition*", Institute of Electrical and Electronics Engineers (IEEE)*, 2002.

[98] B.H. Juang and L. R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development", *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[99] K.K. Paliwal, B.J. Shannon, J.G. Lyons and K.K. Wojcicki, "Speech-Signal-Based Frequency Warping", *IEEE Signal Processing Letters*, vol. 16, pp. 319-322, 2009.

[100] C. Kim, and R. M. Stern, "Feature Extraction for Robust Speech Recognition using a Power-Law Nonlinearity and Power-Bias Subtraction", *INTERSPEECH*, pp. 28–31, 2009.

[101] J. Liu, S. Ranka and T. Kahveci, “Classification and Feature Selection Algorithms for Multi-Class CGH Data”, *bioinformatics,* vol. 24, pp. i86–i95, 2008.

[102] I. R-Lujan, R. Huerta, C. Elkan and C. S. Cruz, “Quadratic Programming Feature Selection”, *Journal of Machine Learning Research*, pp. 1491-1516, 2010.

[103] J. Zhou and H. Peng, “Automatic Recognition and Annotation of Gene Expression Patterns of Fly Embryos”, *bioinformatics,* vol. 23, no. 5, pp 589–596, 2007.

[104] M. J. F. Gales and S. J. Young, “Robust Continuous Speech Recognition using Parallel Model Combination,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[105] C. Kim, and R. M. Stern, “Feature Extraction for Robust Speech Recognition Based on Maximizing the Sharpness of the Power Distribution and on Power Flooring”, *International Conference on Acoustics Speech and Signal Processing (IEEE-ICASSP),* pp 4574-4577, 2010.

[106] M. Korba, D. Messadeg, R. Djemili and H. Bourouba, “Robust Speech Recognition Using Perceptual Wavelet Denoising and Mel-frequency Product Spectrum Cepstral Coefficient Features”, *Informatica-32,* pp 283-288, 2008.

[107] T. Ganchev, P. Zervas, N. Fakotakis, G. Kokkinakis, “Benchmarking Feature Selection Techniques on the Speaker Verification Task”, International

Symposium on Communication Systems, Networks and Digital Signal Processing, (CSNDSP), pp. 314-318, 2006.

[108] P. A. Estéve, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized Mutual Information Feature Selection”, *IEEE Transaction on Neural Networks*, vol. 20, no. 2, 2009.

[109] C. Lee, D. Hyun, E. Choi, J. Go, and C. Lee, “Optimizing Feature Extraction for Speech Recognition”, *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 1, 2003.

[110] M. P. Lewis, “Ethnologue: Languages of the World”, *SIL International,* 16th edition, 2009.

[111] H. Hermansky, “Perceptual linear prediction analysis of speech”, *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[112] L. Ladha and T. Deepa, “Feature Selection Methods and Algorithms”, *International Journal on Computer Science and Engineering (IJCSE),* vol. 3, no. 5, pp. 1787-1797, 2011.

VITAE

- Esam Abid Al-Mashabi.
- Nationality: Saudi.
- Born in Taif, Saudi Arabia.
- Mailing Address: P.O Box: 12605, Dhahran: 31311.
- Contact No. +966505790146.
- Email: emashabi@gmail.com.
- Received a Bachelor of Science degree in Electrical Engineering at KFUPM in June 2001.
- Received a Master of Science degree in Telecommunication Engineering at KFUPM in June 2011.