Rt
441
c.1

CASRA: Colloquial Arabic Speech Recognition Application

by

Omar Aref El - Ariss

Submitted in partial fulfillment of the requirements
for the Degree of Master of Science

Thesis Advisor: Dr. Ramzi A. Haraty

Division of Computer Science & Mathematics
LEBANESE AMERICAN UNIVERSITY
May 2005

LEBANESE AMERICAN UNIVERSITY

GRADUATE STUDIES

We hereby approve the thesis of

Omar Aref El-Ariss

Candidate for the *Master of Science* degree*.

[redacted]

Dr. Ramzi A. Haraty
Associate Professor
Chairman, Division of Computer Science and Mathematics
Lebanese American University
Beirut-Lebanon

[redacted]

Associate Professor
Division of Computer Science & Mathematics
Lebanese American University
Beirut-Lebanon

[redacted]

Dr. Faisal Abu Khzam
Assistant Professor
Division of Computer Science & Mathematics
Lebanese American University
Beirut-Lebanon

*We also certify that written approval has been obtained for any proprietary material contained therein.

I grant to the LEBANESE AMERICAN UNIVERSITY the right to use this work, irrespective of any copyright, for the University's own purpose without cost to the University or to its students, agents and employees. I further agree that the University may reproduce and provide single copies of the work, in any format other than in or from microforms, to the public for the cost of reproduction.

CASRA: Colloquial Arabic Speech Recognition Application

Abstract

by

Omar Aref El-Ariss

Although there was, and still continues, extensive research and advancements in speech recognition on English language, there has been little research done on Arabic language. In addition, most of the research done is either for the standard Arabic language or the Egyptian colloquial language. Commercial applications related to this field are mostly based on telephony technology. In this thesis, the implementation of a Lebanese colloquial Arabic discrete speech recognition is described. According to our knowledge, the thesis proposed here is the first attempt to implement a speech recognition system for the Lebanese colloquial Arabic language.

Praise to the creator, who has made us with inquisitive minds able to seek knowledge and comprehend the mysteries of the universe.

ACKNOWLEDGMENTS

I would like to start thanking every body I know or met during the thesis period because they had an impact on me or on my work. Academically, I would like to thank Dr. Ramzi Haraty and Dr. Ahmad Kabbani for everything. Family wise, I would like to thank my parents, I owe them and will still owe them a lot in this life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech communication is one of the basic and most essential capabilities possessed by human beings. It can be considered the single most important method through which people can easily pass on information without the need for any aid of the senses (like the eyes). The natural form of communication among humans through speech is to be sought into computer technology, and if it is successfully imitated then the human-computer interaction will be more transparent. The advancement of computer technology seen through the growing usage of personal digital assistant (PDA) and tablet PC, makes speech a central, if not the only, means of communication between the human and the machine [5, 6].

In some graphical user interface (GUI) systems a simple operation, example of that is changing the resolution of the screen, requires opening one or more windows or menus, and manipulating sliders, check-boxes, or other graphical elements. This task requires the user to have some knowledge of the system's interface conventions and structures, making the novice user uncomfortable. Spoken language processing allows the operation to be natural and directly accessible. Taking the example of screen resolution, the user can say “increase/decrease the resolution” to change the resolution of the screen. Spoken language processing refers to technologies related to speech recognition, text-to-speech conversion (speech synthesis), and spoken language understanding. Spoken language processing is a diverse subject that relies on knowledge of many levels including acoustics, phonology, phonetics, linguistics, semantics, pragmatics, and discourse. The thesis addressed here concentrates on the automatic speech recognition of the Arabic language.

The speech wave itself conveys linguistic information that includes the meaning the speaker wishes to impart, the speaker's vocal characteristics, and the speaker's emotion. Speech recognition is the process of automatically extracting and determining linguistic information conveyed by a speech wave using computers or electronic circuits. Only the linguistic information is needed from the speech wave, while the rest of the information are used in other fields of signal processing. Data-driven statistical approaches are applied in the development of the speech recognition systems. Those approaches are usually based on modeling the speech signal using well-defined statistical algorithms that can automatically extract knowledge from the data. The data-driven approach can be viewed fundamentally as a pattern recognition problem. The patterns are either recognized during the runtime operation of the system or identified during system construction to form the basis of runtime generative models. A speech recognition system performs three primary tasks [5, 6, 12]:

- Preprocessing: converts the spoken input into a form the recognizer can process.
- Recognition: identifies what has been said by comparing the input with the built-in reference models.
- Communication: sends the recognized input to the software systems that needs it.

1.1 Advantages of Speech Recognition

The increase of importance to the field of automatic speech recognition is due to the fact that implementing computer software that supports speech brings with it many advantages [5, 6]:

1) Speech input is easy to perform because it does not require a specialized skill as typing or pushbutton operations.
2) Speech can be used to input information three to four times faster than typewriters and eight to ten times faster than handwriting.
3) Information can be input even when the user is moving or doing other activities involving the hands, legs, eyes, or ears.
4) Speech as input is more suitable for individuals challenged with a variety of

physical disabilities, such as loss of sight or limitations in physical motion and motor skills.

1.2 Classification of Speech Recognition

Speech recognition can be classified according to how the user of the system speaks. There are two types of speech flow:

- **Isolated word recognition** (called discrete word or discrete utterance): words uttered in isolation are recognized, that means the user must pause between words, and this pause serve two purposes:
    - Preventing cross-word coarticulation from distorting the acoustic pattern of the word to be recognized. In other words, due to the pause between each word, there is no effect between neighboring words (cross-word coarticulation). In addition, the detection of the start and end of a word will be a lot easier than continuous word recognition.
    - Allowing the processor time to accomplish its analyses.

  The development of faster PC processors made the reduction of the length of the pause possible. Isolated word recognition requires a clear beginning and termination for a word or a phrase.
- **Continuous speech recognition**: continuously uttered sentences are recognized, this means that the user speaks in a natural manner without unnatural pauses. The challenges involved in recognizing continuous speech are:
    - The number of words in a block of input speech is generally unknown, due to the fact that the word boundaries are usually unclear. Therefore, all the uttered words seem to be one continuous word.
    - The locations where each word begins and ends are unknown.
    - Cross-word coarticulation effects blur word boundaries.

  These three issues function together to make continuous speech recognition extremely difficult. Continuous speech recognition can be further classified into transcription speech recognition and understanding speech recognition (also called conversational speech recognition). The former aims at recognizing each word correctly, while the latter focuses on understanding the meaning of sentences.

From a different point of view, speech recognition can also be classified according to the speaker modeling. There are two approaches to create speaker models [5, 12]:

- **Speaker-dependent modeling**: creates a special speaker model for every user of the system. Recognition proceeds in two stages; first training, then recognition. The recognition step depends on the type of model used (HMM, template matching, or other types). The training step (also called enrollment), which is the data collection process, creates stored templates (also called word models or reference patterns). Training involves gathering spoken samples for all the items in the vocabulary for each user to use the recognition system. Therefore, new reference templates (models) must be created for every new speaker.
- **Speaker-independent modeling**: can recognize speech uttered by any user without the need of training. The reference models used must recognize large, heterogeneous populations of speakers. Consequently, reference models developed fore speaker-independent applications are more complex and difficult to construct than reference models created for speaker-dependent applications.

Although speaker-independent recognition is much more difficult than speaker dependent modeling, its importance lies in the ease and broad usage.

1.3 Difficulties in Speech Recognition

It is difficult to develop computer programs that are sufficiently sophisticated to understand continuous speech by a random speaker. Recognition of speech becomes possible by computers only when the difficulties are simplified by applying some constraints and rules as isolating words, limiting the vocabulary or number of speakers, or constraining the way in which sentences may be formed. Speech signals can be very different even for the same word depending on speakers, stress, and noise. Furthermore, the length of speech signal, which consists of thousands of samples, even for a simple word, shows a large variation among individuals. Therefore, speech recognition can be considered one of the most difficult pattern recognition tasks due to many difficulties; some of these difficulties are [5, 7, 9, 12]:

1. ***The voluminous data in the speech sound wave***

   Although it may seem as if we speak using a single tone, the quantity of data in the sound wave is overwhelming. Within the range of human hearing, speech sounds can span more than 20,000 frequencies.

2. ***The paucity of information in the speech sound wave***

   Speech is more than acoustic sound patterns, additional knowledge, as word meanings, is needed in order to recognize exactly the intended speech. Therefore, words with widely different meanings may share the same sequence of sound patterns. For example:

   - The word'كَلَّ' that means exhausted, and the word 'كَلاَّ' that means no or never.
   - The word 'جَرَّ' that means to drag, the word 'جَرَّى' that means to make something to stream, and the word 'جَرَّة' that means a jar.

3. ***The continuous flow of speech***

   Speech is uttered as a continuous flow of sounds and even when words are spoken distinctly there is no inherent separation between the words. To illustrate this idea, any unfamiliar foreign language can be heard as a continuous stream of sound without any distinction or identification of the word boundaries.

4. ***Variability***

   A person's voice and speech patterns can be entirely different from those of another person. The elements that cause the difference are many: size and shape of the mouth, length and width of the neck, age, sex, regional dialect, health, and personal style of speech. An example of this variability is: speakers in Egypt pronounce the phoneme 'ج' in the word 'جمال' different than the speakers in Lebanon. Another variability is that some speakers talk more slowly or more nasally.

5. ***More variability***

Even a single speaker will exhibit variability. The sound pattern of a word changes when speakers whisper, shout, and become angry, sad, tired, or ill. Even when speaking normally, individual speakers rarely say a word the same way twice. Variability is a basic characteristic of speech.

6. ***Coarticulation effects***

The acoustic realization of a phoneme may heavily depend on the acoustic context in which it occurs. This effect is usually called coarticulation. Thus, the acoustic feature of a phoneme is affected by the neighboring phonemes, the position of a phoneme in a word, and the position of this word in a sentence. Such acoustic features are very different from those of isolated phonemes, since the articulatory organs do not move as much in continuous speech as in isolated utterances. We can see the effect of coarticulation in the following phrase 'و في الأيام'. Here the phoneme 'ي' in the word 'في' is affected by the neighboring phoneme 'فـِ' and by the phoneme 'لْ' in the word 'الأيام'. Therefore, the acoustic realization is different from the stand alone phoneme 'ي'.

7. ***Insufficient linguistic knowledge***

With the help of linguistic knowledge, such as syntactic and semantic constraints, the listener can usually predict the next word. Unfortunately, this kind of knowledge is not applied in the field of speech recognition due to the difficulty to model this mechanism.

8. ***Arabic language difficulties***

   Arabic language presents problems that are not encountered in mainstream languages like English or Spanish. The cause to such difficulties is the extreme dialectical variation and non-standardized speech representations. Some of the dialectical variation for the word 'رحم' are: 'رَحِمَ', 'رُحِمَ', 'رَحَّمَ','رَحِمْ'.

9. ***Noise***

   Background noise is not the only intrusion speech recognition systems must combat. They must handle:

   - Noise produced by the input device (telephone or microphone).
   - Sounds made by the speaker; such as lip smack, nervous breathing.
   - Non-communication vocalizations made by the speaker such as "uh", a cough.

1.4 Scope of the Thesis

The research proposed is for an Arabic speech recognition application, concentrating on the Lebanese dialect for the six digits from number one to number six. The speech recognition system is a small vocabulary based system that is speaker dependent and accepts discrete speech, that is the user has to pause between words in order to identify the word boundaries. The system starts by sampling the speech, which is the process of transforming the sound from analog to digital, and then extracts the features by using the Mel-Frequency Cepstral Coefficients (MFCC). The extracted feature is then compared with the system's stored model; in this case the stored model chosen is a word-based model. The reference model used is template matching. Dynamic time warping is applied in comparing the input sound with the stored templates to improve the difference in duration. The major problem that was encountered during the implementation phase of the thesis was the vast theoretical material with no implementation basis. In other words, there were not enough material showing the theory being applied, and how each process or theory interact with the other processes. Other problems encountered were the variety of speech, and the distortion of the sound signal due to various types of noise.

1.5 Organization of the Thesis

The thesis is organized as follows. Chapter2 gives a brief background to the field of speech recognition, followed by perspective of the application of speech recognition to the Arabic language. Chapter 3 describes the human perception and production of speech. Chapter 4 discusses the digital signal processing methods that are popular and mostly used in the field of speech recognition. Chapter 5 is concerned with the description of the implementation process of the proposed thesis. Chapter 6 gives the theory to evaluate a speech recognition system, then proceeds with experimentations done to evaluate the proposed system. Chapter 7 concludes the thesis. An appendix of the definition of the technical terms used in the chapters is supplemented.

Chapter 2

Background

The chapter starts with a brief historical overview of signal processing, an essential part in a speech recognition system. The chapter proceeds with a historical perspective in the general field of automatic speech recognition. The chapter concludes with an up-to-date perspective in the field of Arabic speech recognition research and development.

2.1 Brief Historical Overview of Digital Signal Processing

The assumptions by electronics engineers on the application of digital hardware techniques on the many problems that are currently solved by digital signal processing started since World War II, if not earlier. It was not until the mid 1960's that the first major contributions to the field of Digital Signal Processing in the area of digital filter design and synthesis were proposed by Kaiser (at Bell Laboratories). Kaiser's work showed clearly how to design useful digital filters using the bilinear transform. At about the same time (1965), a great motivation to the field was given by James W. Cooley and John W. Tukey. Cooley-Tukey paper proposed a fast method of computing the Discrete Fourier Transform (DFT), which has come to be known as the Fast Fourier Transform (FFT). The FFT method reduces the computation time of the DFT by one to two orders of magnitude. The great importance of the FFT was that it showed quite strikingly that digital methods, opposed to analog methods, could be basically more economic to employ for spectrum analysis [3, 14].

2.2 Historical Overview of Speech Recognition

The first documented attempts to construct an automatic speech recognition system occurred long before the digital computer was invented. The invention of the telephone in 1871 by Antonio Meucci, later accredited to Alexander Graham Bell in 1876, constituted not only the most important period in history of communications, but it also represented the first step in which speech began to be dealt with as an engineering target. In 1913, the Russian scientist A. Markov described a network model capable of analyzing the letter sequence in the text of a literary work. This model, the Markov chain, was used as the basis for computational models in a wide range of fields, including models of human language. Then in the 1922, the first machine to recognize speech was manufactured. The machine was a commercial dog toy called Radio Rex. The dog would only respond to the name "Rex", and when it is uttered the dog jumps from his home to the feet of his loving owner [5, 6, 12,].

The hope of early researchers at Bell Laboratories, RCA Laboratories, and elsewhere was that speech recognition would be straightforward and easy. In 1952 the first machine capable of recognizing speech was built at AT&T Bell Laboratories. The system used template matching; it compared stored reference patterns of the ten English digits with utterances of individual digits. It required extensive tuning to recognize the speech of a person [12].

By the mid 1960's, most researchers have realized that speech recognition was far more complex and broader than they had predicted. As a result of this realization, the field was narrowed to systems that can handle speech of one person, the input contains pauses between words, and the vocabulary is small. Also during the 1960's, the recognition systems started to include techniques to minimize differences in the speed with which a person might say a word. The systems no longer sought exact matches, instead they chose the acoustic pattern that most closely resembled the utterance. From the year 1967 through 1970, the basic HMM theory, the most successful method for acoustic modeling, was published in a series of classical papers by Baum and his colleagues [6, 12].

In 1971, the first speech recognition product, VIP 100 system, was developed by Threshold Technology Inc. In the same year, the Advanced Research Projects Agency (ARPA) of the United States Department of Defense launched the Speech Understanding Research project (ARPA SUR, which ended in 1976). ARPA SUR had a profound effect on the course of speech recognition research and development. Researchers began investigating using HMM's for speech recognition in the early 1970's. In 1975, James Baker of Carnegie Mellon University (CMU), used HMM's for speech recognition to develop CMU's DRAGON system for the ARPA SUR project. While in 1976 Frederick Jelinek and his colleagues at IBM Research pioneered widespread applications using HMM. In 1982, James and Janet Baker founded Dragon Systems, and soon developed the DragonScribe system, one of the first commercial products using HMM technology. HMM technology did not gain widespread acceptance for commercial systems until the late 1980's, but by 1990 HMM has become one of the most powerful statistical methods for modeling speech signals, and most of the state-of-art speech recognition systems on the market are based on it. In 1986, Speech Systems Inc. introduced the first very large vocabulary, commercial system. The PE100 was a twenty thousand words, phoneme-based, continuous speech, speaker-independent. In 1994, Philips Dictation Systems launched the first PC-based, very large vocabulary dictation system with continuous speech recognition [6, 12].

2.3 Arabic Speech Recognition Research and Development

Although there was, and still continues, extensive research and advancements in speech recognition on English language, there has been little research done on Arabic language. The Department of Electrical Engineering in the University of Washington funded by DARPA (previously called ARPA), NSF, and CIA did research and development on automatic speech recognition of dialectal Arabic. The importance of this research is the concentration on the Egyptian dialect rather than formal Arabic [9, 10].

Supervised by one of the researchers from the University of Washington, a workshop at the University of Johns Hopkins University was also done on the dialectal Arabic (Egyptian) speech recognition [10]. Two researches were done in the Arab world. The first was done in Algeria, the speech recognition developed recognizes isolated formal Arabic words. The system was based on a hybrid HMM model [11]. The latter research was done at the American University of Beirut. The system developed is based on recurrent neural networks rather than HMM. This speech recognition system also recognizes isolated words, and the vocabulary is composed of six formal Arabic words [4]. Finally, in the commercial industry, Sakhr has developed an Arabic Automatic Speech Recognition system. The system is a telephony-based application composed of a very large vocabulary, phoneme-based, continuous speech, and speaker-independent [16].

Chapter 3

Fundamental Characteristics of Speech

The first two chapters gave a brief introduction to the field of speech recognition, but to further understand this concept, a description of the human speech organs, the production and perception of speech, and sound types must be described. This information clarifies the recognition process by showing the distinct properties of speech that the recognition system searches for in a sound wave in order to identify the utterance. The chapter starts with acoustic and articulatory information general to all languages, and then proceeds with a specific description for the Arabic language.

3.1 The Speech Communication Pathway

The speech communication pathway describes the process of speech production from the speaker's side to the process of perception from the listener's side. The communication pathway can be decomposed into five levels [6, 13]:

1. **Linguistic level at the speaker's side**
   Speech starts in the brain with a thought, and intent to speak this thought. This thought or idea is then transformed into words and sentences according to the grammatical rules of the language present in the speaker's mind.

2. **Physiological level at the speaker's side**
   The brain transforms the linguistic units (words and sentences) into electric signals or nerve impulses that move along the motor nerves. Those electric signals initiate the movement of the muscles in the vocal tract and vocal cords.

3. **Acoustic level at both the speaker's side and the listener's side**

   The movement of the vocal tract and vocal cords causes a pressure change forming a sound wave. The sound wave propagates through space as a chain reaction among the air particles, resulting in a pressure change at the ear canal, of both the speaker and listener, and thus vibrating the eardrum.

4. **Physiological level at both the speaker's side and the listener's side**

   The vibration of the eardrum causes the production of electric signals that move along the auditory nerve system to the brain.

5. **Linguistic level at both the speaker's side and the listener's side**

   The brain of both the listener and speaker performs speech recognition and understanding. Here the speaker listens to the speech to check for the correctness and if the speech is audible.

3.2 The Vocal Organs

The human speech apparatus, as shown in the schematic diagram in figure 3.1, consists of [6]:

- **lungs:** from it the source of air during speech is produced.
- **vocal folds (larynx):** if the folds are close together, they will vibrate as air passes through them; if they are far apart, they won't vibrate. The place where the vocal folds come together is called the glottis.
- **velum (soft palate):** operates as a valve, opening to allow passage of air through the nasal cavity.
- **hard palate:** a long relatively hard surface at the roof inside the mouth, which, when the tongue is placed against it, enables consonant articulation.
- **tongue:** flexible articulator, shaped away from the palate for vowels, placed close to or on the palate or other hard surfaces for consonant articulation.
- **teeth:** another place of articulation used together with the tongue for certain

consonants.

- **lips:** can be rounded or spread to affect vowel quality, and closed completely to stop the oral air flow in certain consonants.

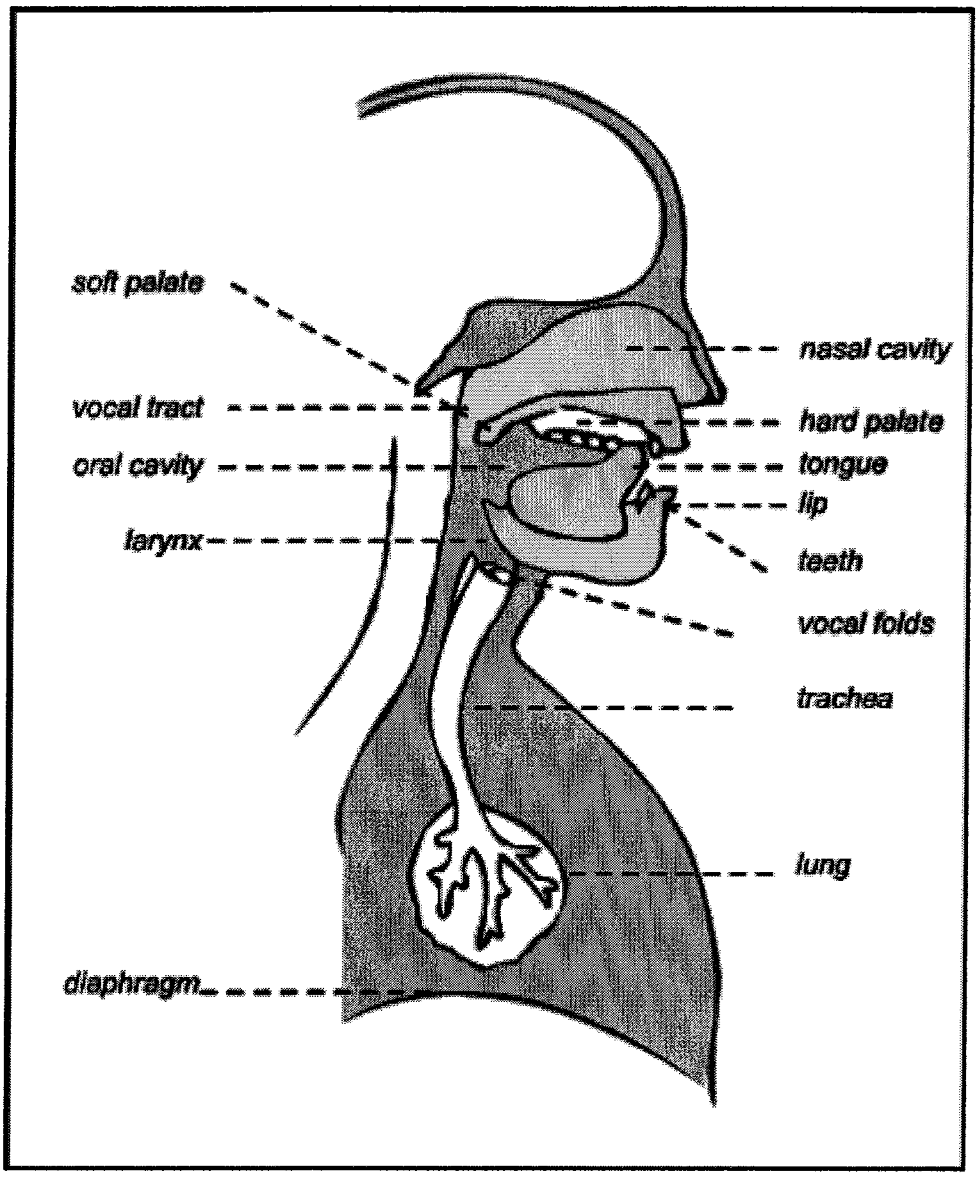


Figure 3.1: ***The Human Voice Organs***

3.3 Speech Production

In order for the sound to be produced, air must pass through the vocal organs. The release of air happens when the abdominal muscles force the diaphragm up. The movement of the diaphragm upwards causes the air to be pushed up and out from the lungs. As the air passes through the windpipe (or trachea), it passes through the larynx, which is commonly known as the Adam's apple or voice box. The larynx has two small folds of muscle called the vocal folds (often referred to non-technically as the vocal cords), which can be moved together or apart. The glottis, which is the gap between the left and the right vocal cords, is usually open during breathing, and becomes narrower during the production of sound. The airflow through the glottis is then periodically interrupted by the opening and closing of the gap in accordance with the interaction between the airflow and the vocal cords. Then the airflow passes through the vocal tract, which consists of the oral tract and the nasal tract. Finally, the airflow can exit the body either through the mouth or through the nose.

Most of the sounds produced are made by air passing through the mouth. Sounds that are made by air passing through the nose are called nasal sounds. Nasal sounds use both the oral and nasal tracts as resonating cavities [5, 8].

3.4 The Voicing Mechanism

Speech can be viewed as a sequence of units that are linked in time. The smallest unit that has a unique articulatory and acoustic characteristic is called a phoneme. Phonemes in most languages are considered to be the alphabet letters. Other forms of unit, as syllables (an example of that is 'ال') or words, that also have a unique articulatory and acoustic characteristic can be used to describe a speech. Combining phonemes together form other larger units as syllables or words. Therefore, the knowledge of the distinctive characteristics of every phoneme in a particular language is essential for recognition [6].

Unfortunately, every person has a unique vocal anatomy and also due to the effect of articulation (discussed in the introductory chapter) the same phoneme is not pronounced the same way. Therefore, the term phone is used to describe a phoneme's acoustic realization. For example, the phoneme 'كْ' in the word 'كَلامْ' and the word 'كِلابْ' has a different realization [6, 8].

The most fundamental distinction between sound types in speech is whether the sound is considered to be voiced or unvoiced; an example of this distinction is shown in figure 3.2. Voiced sounds have more energy than voiceless sounds, and have in their time and frequency structure a regular pattern. On the other hand, voiceless sounds have no particular pattern. The reason for this difference is due the vibration of the vocal chords during the production of voiced sounds. An example of voiced and unvoiced sounds is: vowels for the former, and noise and some consonants for the latter. Other forms of sound distinction can be based on articulators and their effect depending on their positions [6].

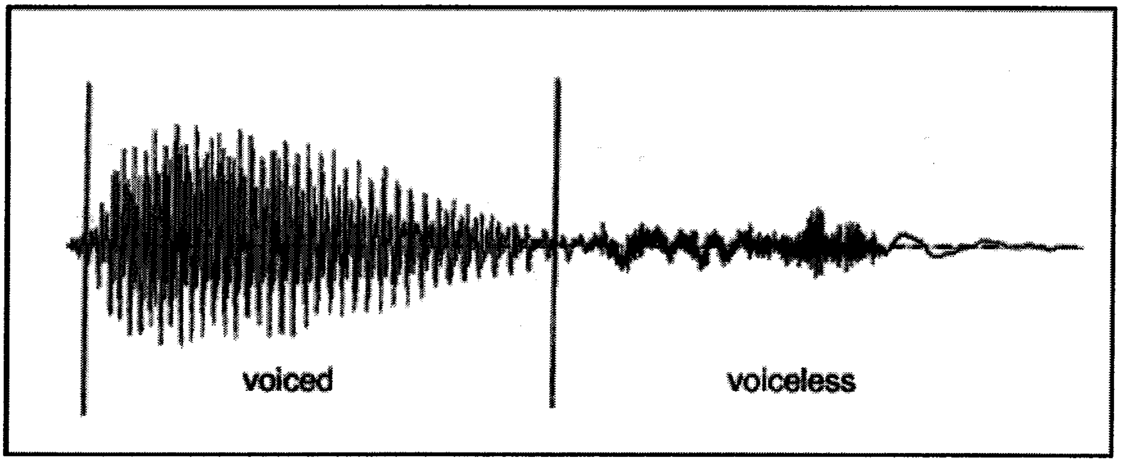


Figure 3.2: ***Voiced phoneme followed by a voiceless phoneme (adapted from Markowitz, Using Speech Recognition, Copyright 1996, Prentice Hall)***

Phonemes or Phones are divided into two main classes [6, 8]:

- **Consonants:** are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced.
- **Vowels:** the tongue shape and positioning in the oral cavity do not form a major constriction of airflow during the phase of articulation. However, variations of tongue placement give each vowel its distinct character by changing the resonance.

Vowels are usually voiced, and are generally louder and longer lasting than consonants.

3.5 Articulation of Vowels

The two important mechanisms for vowel articulation are the tongue and the lips. Although the tongue's shape and position do not form a major limitation of airflow during articulation, but these variations are responsible for giving each vowel a distinctive characteristic. The shape and rounding of the lips add more distinction to each vowel [5, 6, 8].

3.6 Articulation of consonants

Consonants can be distinguished either by the place of articulation, or by the manner of articulation.

3.6.1 Place of Articulation

Consonants can be distinguished by the place were the restriction of airflow happens. The point of maximum restriction is called the place of articulation of a consonant. Places of articulation, shown in figure 3.3, are often used in automatic speech recognition as a useful way of grouping phones together into equivalence classes:

- **labial:** Consonants whose main restriction is formed by the two lips coming together have a bilabial place of articulation. An example of bilabial in Arabic is the two consonants 'بْ' and 'مْ' in the word 'إِبْهَمْ'. The labiodental consonants, as the phoneme 'فْ' in the word 'حَذِفْ', are made by pressing the bottom lip against the upper row of teeth and letting the air flow through the space in the upper teeth.
- **dental:** Sounds that are made by placing the tongue against the teeth are dentals. Example of that in Arabic language is the 'ثْ' as in 'بَحِثْ', which is made by placing

the tongue behind the teeth with the tip slightly between the teeth.

- **alveolar**: The alveolar ridge is the portion of the roof of the mouth just behind the upper teeth. The phones 'تْ', 'نْ', and 'دْ' by placing the tip of the tongue against the alveolar ridge.
- **palatal**: The roof of the mouth (the palate) rises sharply from the back of the alveolar ridge. The palato-alveolar sound 'شَ' 'شَمِس' is made with the blade of the tongue against this rising back of the alveolar ridge. The palatal sound 'يَ' of 'يَلْعَبْ' is made by placing the front of the tongue up close to the palate.
- **velar**: The velum or soft palate is a movable muscular flap at the very back of the roof of the mouth. The sounds 'كَ' in 'كَهفْ', 'ءْ' in 'نَبَاء', and 'ءِ' in 'إلتِقَافْ' are made by pressing the back of the tongue up against the velum.
- **glottal**: is made by closing the glottis, that is by bringing the vocal folds together.

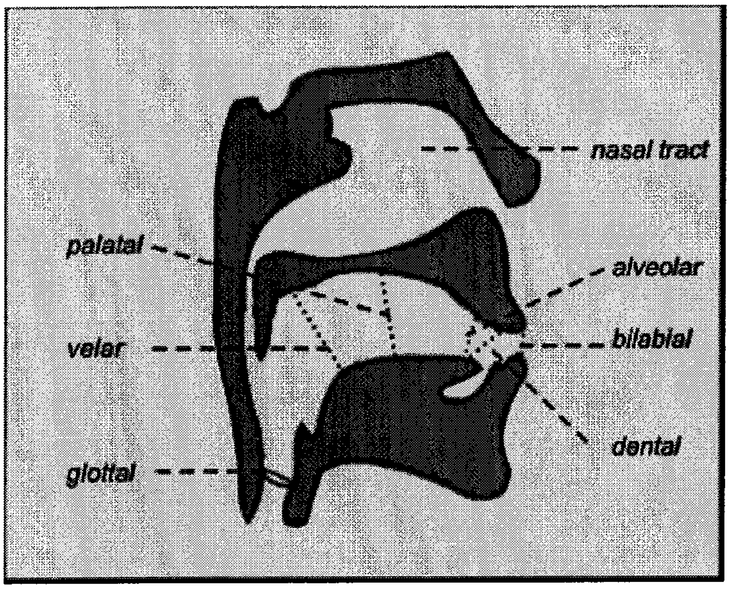


Figure 3.3: ***Places of Articulation***

3.6.2 Manner of Articulation

Consonants are also distinguished by how the restriction in airflow is made, for example whether there is a complete stoppage of air, or only a partial blockage, etc. This feature is called the manner of articulation of a consonant. The combination of place and manner of articulation is usually sufficient to uniquely identify a consonant. Here are the major manners of articulation for English consonants:

- **stop:** A stop, or a plosive, is a consonant in which airflow is completely blocked for a short time. This blockage is followed by an explosive sound as the air is released. The period of blockage is called the closure and the explosion is called the release. Examples of Arabic stops are ‘ب’, ‘د’, ‘ض’, ‘ط’, and ‘ك’.
- **nasals**: The nasal sounds as the consonants ‘ن’, ‘م’, and ‘و’ are made by lowering the velum and allowing air to pass into the nasal cavity.
- **fricative**: In fricatives, airflow is constricted but not cut off completely. The turbulent airflow that results from the constriction produces a characteristic "hissing" sound. The labiodental fricatives, as ‘ف’, are produced by pressing the lower lip against the upper teeth, allowing a restricted airflow between the upper teeth. The dental fricatives, as ‘ث’, allow air to flow around the tongue between the teeth. The alveolar fricatives, as ‘س’ and ‘ز’, are produced with the tongue against the alveolar ridge, forcing air over the edge of the teeth. In the palato-alveolar fricatives, as ‘ش’, the tongue is at the back of the alveolar ridge forcing air through a groove formed in the tongue. The higher pitched fricatives (‘س’, ‘ز’, ‘خ’) are called sibilants. Stops that are followed immediately by fricatives are called affricates.
- **approximant**: In approximants, the two articulators are close together but not close enough to cause turbulent airflow. This can be seen in the Arabic letter ‘ي’ in the word'يَهْمِسْ', the tongue moves close to the roof of the mouth but not close enough to cause the turbulence that would characterize a fricative.
- **tap**: A tap or flap is a quick motion of the tongue against the alveolar ridge. The consonant ‘تِ’ in the middle of the word 'التِبَاس' is a tap [8].

3.7 The Arabic Language

Linguistically speaking, Arabic language does not have a normalized form that is used in all circumstances of speech and writing. Arabic used in daily informal communication is not the same form of Arabic that is used on TV to broadcast the news. The forms of Arabic are as follows:

- **Classical or formal Arabic:** is the old form of the language. It can be seen in the Jahelia poetry.
- **Modern Standard Arabic (MSA):** is a version of classical Arabic with modernized vocabulary. It is considered to be the formal language that is common in all Arabic speaking countries. Modern Standard Arabic is the form of Arabic used in all written texts.
- **Colloquial or dialectical Arabic:** there are many different dialects that differ considerably from each other and from the Modern Standard Arabic. According to the paper [10], colloquial Arabic can be divided into the following subgroups: Gulf Arabic, Egyptian Arabic, Levantine Arabic, and North African Arabic. This categorization is too general and wrong at the same time. Dialectical forms of Arabic can be many even in one country. For example Lebanon, dialects are different in the south, north, Beirut, and the mountains, further dialectical subdivisions can also be made. Another example, in Oman the dialect spoken is similar to the dialect spoken in Sudan and not to the other Gulf countries. The regional dialects of Arabic are spoken languages; very little written dialectical material exists [9].

Although some consider the alphabet to consist of twenty-eight letters (excluding the hamza) [10, 17], the Arabic alphabet consists of twenty-nine letters, shown in the table 3.1. Additional symbols or letters can be introduced for certain phones that are not present in the Arabic alphabet (like the English phonemes [p] and [v]).

Table 3.1: The Arabic Alphabet

Letter	*Name*	*Consonant (sokoon)*	*Short vowel (fatha)*	*Short vowel (damma)*	*Short vowel (kasra)*	*Long vowel (ا)*	*Long vowel (و)*	*Long vowel (ي)*
ب	Baa	بْـ	بَـ	بُـ	بِـ	با	بُو	بِيـ
ت	Taa	تْـ	تَـ	تُـ	تِـ	تا	تُو	تِيـ
ث	Thaa	ثْـ	ثَـ	ثُـ	ثِـ	ثا	ثُو	ثِيـ
ج	Gym	جْـ	جَـ	جُـ	جِـ	جا	جُو	جِيـ
ح	Haa	حْـ	حَـ	حُـ	حِـ	حا	حُو	حِيـ
خ	Khaa	خْـ	خَـ	خُـ	خِـ	خا	خُو	خِيـ
د	Daal	دْ	دَ	دُ	دِ	دَا	دُو	دِيـ
ذ	Thal	ذْ	ذَ	ذُ	ذِ	ذا	ذُو	ذِيـ
ز	Zayn	زْ	زَ	زُ	زِ	زا	زُو	زِيـ
ر	Raa	رْ	رَ	رُ	رِ	را	رُو	رِيـ
س	Seen	سْـ	سَـ	سُـ	سِـ	سا	سُو	سِيـ
ش	Sheen	شْـ	شَـ	شُـ	شِـ	شا	شُو	شِيـ
ص	Saad	صْـ	صَـ	صُـ	صِـ	صا	صُو	صِيـ
ض	Daad	ضْـ	ضَـ	ضُـ	ضِـ	ضا	ضُو	ضِيـ
ط	Taa	طْ	طَ	طُ	طِ	طا	طُو	طِيـ
ظ	Zaa	ظْ	ظَ	ظُ	ظِ	ظا	ظُو	ظِيـ
ع	Ayn	عْـ	عَـ	عُـ	عِـ	عا	عُو	عِيـ
غ	Ghayn	غْـ	غَـ	غُـ	غِـ	غا	غُو	غِيـ
ك	Kaaf	كْـ	كَـ	كُـ	كِـ	كا	كُو	كِيـ
ق	Qaaf	قْـ	قَـ	قُـ	قِـ	قا	قُو	قِيـ
ف	Faa	فْـ	فَـ	فُـ	فِـ	فا	فُو	فِيـ
ل	Laam	لْـ	لَـ	لُـ	لِـ	لا	لُو	لِيـ
ن	Noon	نْـ	نَـ	نُـ	نِـ	نا	نُو	نِيـ
م	Meem	مْـ	مَـ	مُـ	مِـ	ما	مُو	مِيـ

ه	Haa	هْـ	هَـ	هُـ	هِـ	ها	هُو	هِيـ
و	Waw	وْ	وَ	وُ	وِ	وا	وُ	وِيـ
ي	Yaa	يْـ	يَـ	يُـ	يِـ	يا	يُو	يِـ
ء	Hamza	ئْـ	أَ	أُ	إِ	آ	أُوُ	إِيـ
ا	Alif	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Arabic doesn't have letters for vowels; all the alphabets are consonants. Diacritics play an important role in forming short vowels. The fatha, kasra, damma, and tanween all form different short vowels for the same letter. Long vowels can also be produced by adding an 'ا' after a short vowel. Also the madda diacritic form a long vowel for the letter 'ا'. The sokoon means that the letter is a consonant, while the shadda doubles the letter (the first is a consonant while the other letter is a vowel). The lack of diacritics in a word might cause considerable ambiguities, leading the vocabulary to be used in a speech recognition system to give wrong results. The word 'كتب' as an example has a possibility of 21 diacritizations. Therefore in order for a word-based speech recognition system to recognize those diacritization, the system must have at least one model for every diacritization form. Table 3.2 lists all the Arabic diacritics [1, 10].

Table 3.2: Arabic Diacritics

Symbol	Name	Meaning	Example
ْ	Sokoon	Consonant letter	حبْس
َ	Fatha	Short vowel	كَذَبَ
ُ	Damma	Short vowel	كُل
ِ	Kasra	Short vowel	عنِد
ّ	Shadda	Letter doubling	شدَّة
ً	Tanween el-fatha	Adds [an] to the letter	ايضاً
ٌ	Tanween ed-damma	Adds [on] to the letter	نصرٌ
ٍ	Tanween el-kasra	Adds [in] to the letter	بيتٍ

~	Madda	Turns the hamza into a long vowel	آدم

The Modern Standard Arabic has at least one hundred twelve phonemes. Every letter except the letter 'ا', which is not included because it just changes the vowel duration from short to long, are affected by the four diacritics: fatha, damma, kasra, sokoon. Therefore, every letter has four phonemes.

3.7.1 Letter production in the Arabic language:

Arabic letters can be divided into subgroups depending on the place and manner of articulation. Figure 3.4 shows a detailed diagram for the articulation of the Arabic letters. Unlike the groupings (for consonants and vowels) described above, the categorization differs. Some of the relevant categorizations are mentioned below:

- According to Ibn Sina [17], letters can be either single or composite. Letters that are produced by complete blockage of airflow are considered to be single letters. The single letters are: ‘ب’, ‘ت’, ‘ج’, ‘د’, ‘ط’, ‘ض’, ‘ق’, ‘ك’, ‘ل’, ‘م’, ‘ن’. The rest of the letters are considered to be composite.

- According to Yousof Bin Abi Beker El Sakaki [17], letters can also be loud or quiet. The loud letters are the product of the restriction of the airflow, while the quiet letters have no restriction of airflow during production of the letter. The loud letters are: 'ء', 'ا', 'ق', 'ك', 'ج', 'ي', 'ر', 'ن', 'ط', 'د', 'ت', 'ب', 'م', 'و'. The rest of the letters are considered to be quiet.

- According to Abdallah Bin Mohammed El Khafaji [17], the letters can be dense, intermediate, or loose. The letters are considered to be dense when the airflow is obstructed during production, and considered to be loose if there was no obstruction of the airflow. The dense letters are: 'ء', 'ق', 'ك', 'ج', 'ط', 'د', 'ت', 'ب'. The intermediate letters are: 'ا', 'ع', 'ر', 'ل', 'ي', 'ن', 'م', 'و'. The rest of the letters are considered to be loose.

- Other categorizations are [17]: Closed letters are the letters that are produced by closing the lips and raising the tongue, the letters are: 'ص', 'ض', 'ط', 'ظ'. The tip letters are the letters that are produced by using the tip of the tongue, and the letters are: 'ل', 'ر', 'ن', 'ف', 'ب', 'م'.

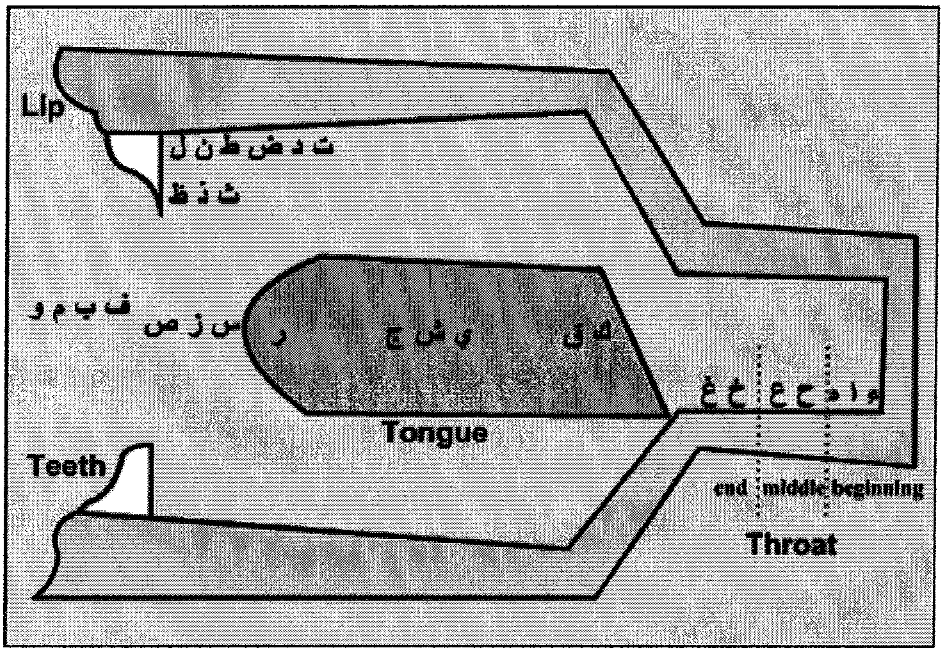


Figure 3.4: ***Places of articulation for the Arabic letters***

Chapter 4

Digital Signal Processing

A speech recognition system, using the typical state-of-the-art technology, is not able to watch the user during the utterance of speech, or to analyze the user's articulatory organs, but has the stream of speech as its only source of input. Therefore, a thorough understanding of the characteristics of the speech signal and the methods used to convert the signal and to extract useful information, play an essential part in the development of the recognition process. The chapter starts by a description of the speech signal, then proceeds with Digital Signal Processing (DSP) methods that are useful and implemented in this research for speech recognition.

4.1 The Signal

Speech is an analog signal, a pressured wave of air particles, that is composed of a continuous flow of sound waves and silence. Some speech sounds, such as voiced phonemes, have regular patterns of air movement. The simplest sound pattern (a simple sine wave), which is shown in figure 4.1, can help in describing the sound signal. The vibratory movement of the wave, shown in wave (a), is composed of high air pressure and low air pressure. The high air pressure causes the eardrum to move from its rest position to an inward position, while a low air pressure causes the eardrum to move downward [12].

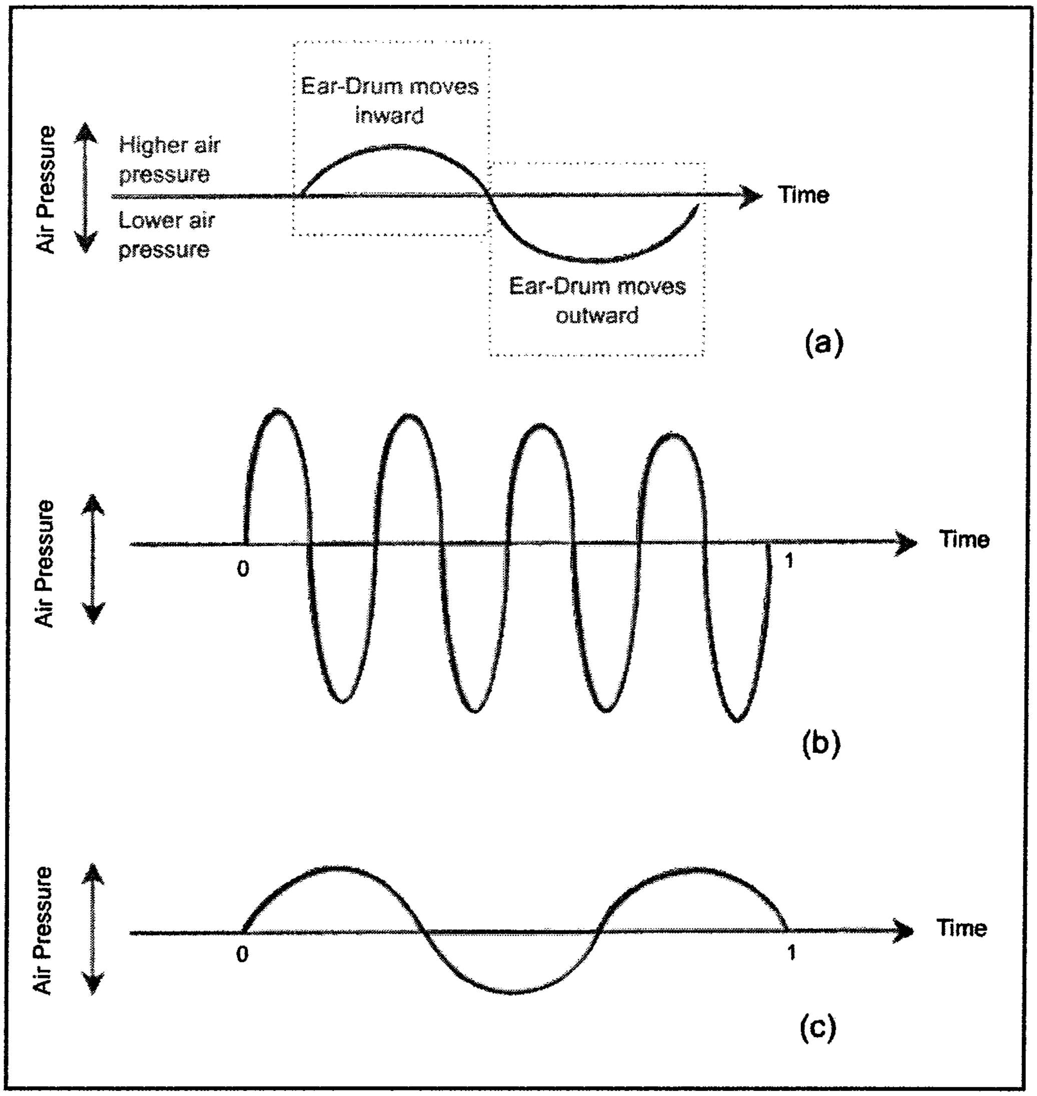


Figure 4.1: ***Pure tone sound patterns (adapted from Markowitz, Using Speech Recognition, Copyright 1996, Prentice Hall)***

The frequency of a wave is the number of vibratory cycles during one second. Therefore, a wave that vibrates 3000 times per second is said to have a frequency of 3000 Hz or 3 kHz. The pitch of sound depends on the frequency, the higher the frequency the higher the pitch. The loudness of the sound, on the other hand, depends on the amplitude of the wave (measured in decibels dB). In figure 4.1, wave (b) and (c) cover the same length of time, but wave (b) has a frequency of 4 Hz and an amplitude that is double the amplitude of wave (c) that has a frequency of 2 Hz. Therefore, wave (b) is considered to be higher in pitch and louder.

Most of the speech sounds are more complex in form than the waves shown in figure 4.1. They are not only composed of one frequency, but can contain many, having one frequency as a dominant (or primary) frequency . The dominant frequency is called the fundamental frequency, acoustically defined as the rate (rate of cycling) at which the vocal cords flap against each other when producing a voiced phoneme. The fundamental frequency is responsible for the pitch of the sound, while the secondary frequencies are responsible for the quality or timbre of the sound. The secondary frequencies help in recognizing the voice of specific individuals, and in distinguishing phonemes. These frequencies are called formants and are produced through resonance [12].

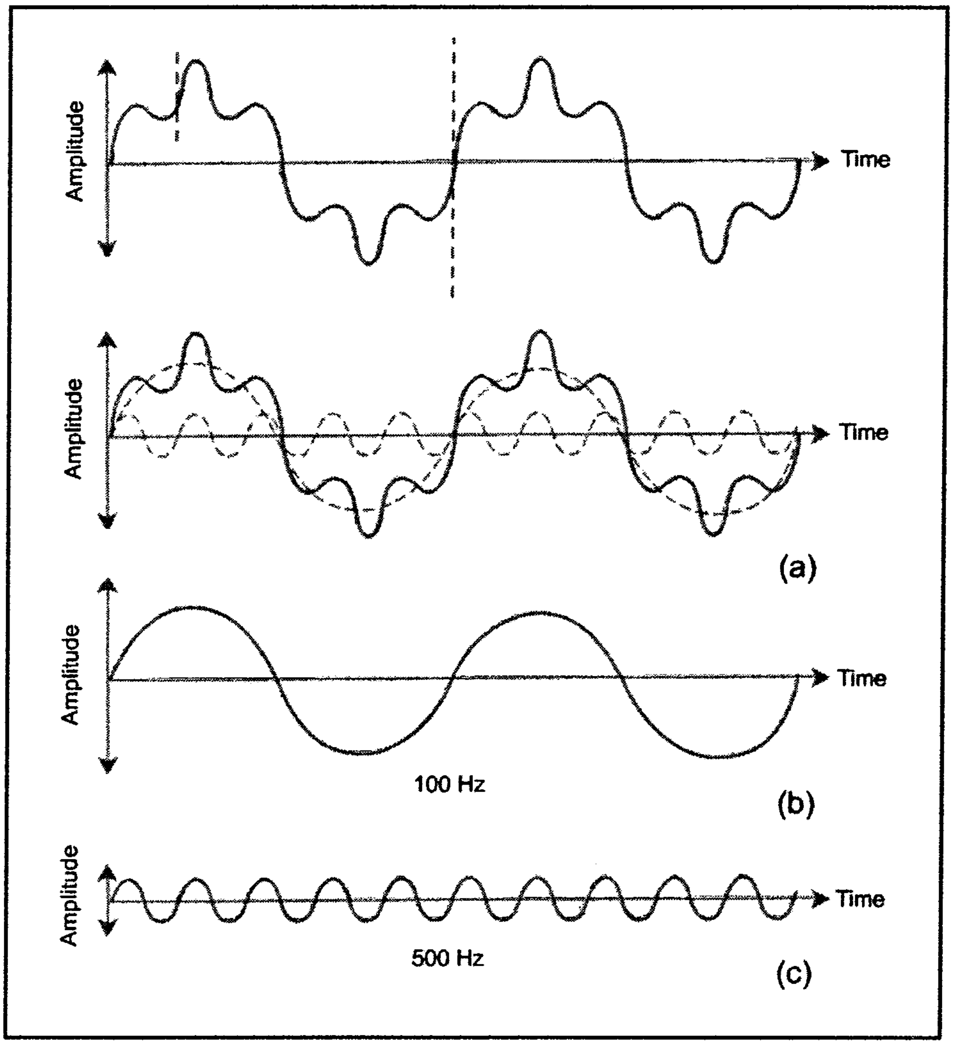


Figure 4.2: ***A complex wave representation (adapted from Markowitz, Using Speech Recognition, Copyright 1996, Prentice Hall)***

An example of a multi-frequency sound signal composed of two frequency waves is shown in figure 4.2. The fundamental frequency can be seen as the cycle that ends at the blue line, while the secondary frequency can be seen as the cycle that ends at the red line. Wave (a) shows the original waves (the fundamental and the secondary frequencies) being displayed as dotted line to help to locate them. Wave (b) and (c) show the original waves of the two frequencies. Wave (b) depicts the fundamental frequency (100 Hz), and wave (c) shows the secondary frequency (500 Hz). Phonemes are usually richer in secondary frequencies than figure 4.2 shows. The figure below shows the structure of two phonemes (خَ), signal (a), and (بُ), signal (b). Both phonemes have a fundamental frequency and many secondary frequencies. The repeating patterns and the fundamental frequencies are more clear in phoneme (خَ) than phoneme (بُ).

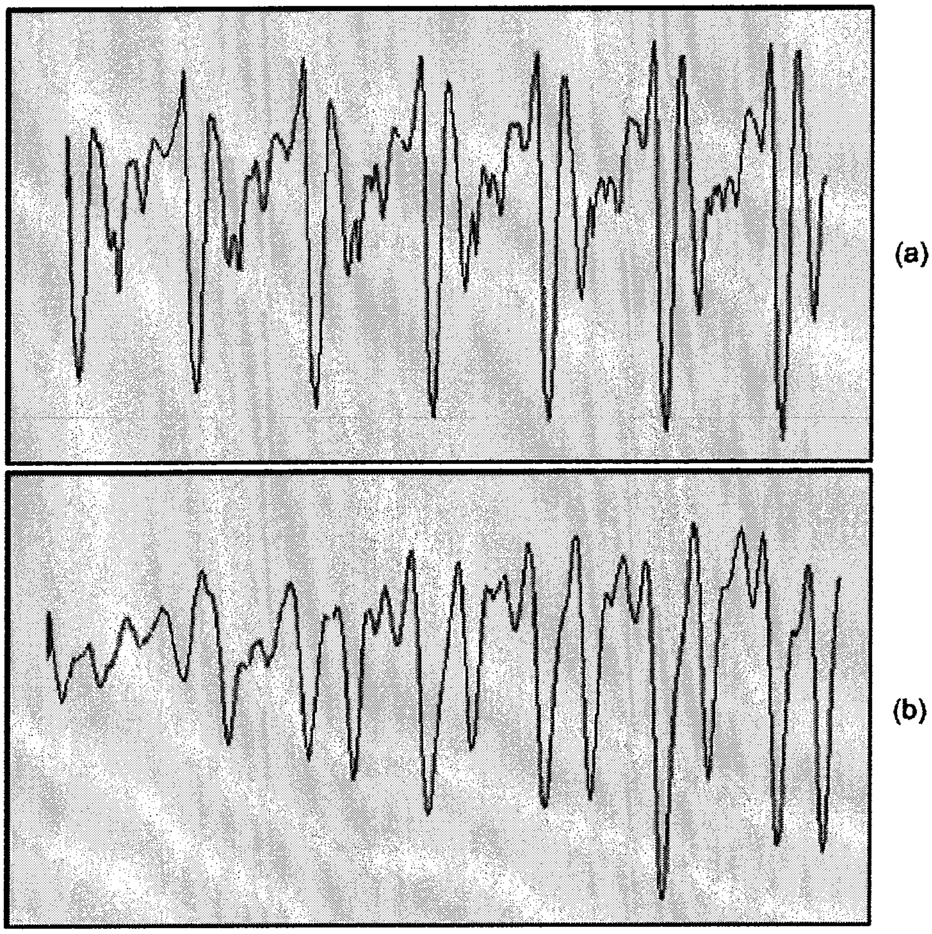


Figure 4.3: ***Complex wave representation for phonemes (خَ) and (بُ)***

Figures 4.1 to 4.3 depicted only cyclic waves, or voiced phonemes. Figure 4.4 shows an acyclic wave. Acyclic sound waves do not have repeating patterns, in other words there are no fundamental frequencies. Sounds that produce acyclic waves, as voiceless phonemes are often called noise. The nature and frequencies of acyclic wave patterns provide critical information about the identity of the phoneme being produced [12].

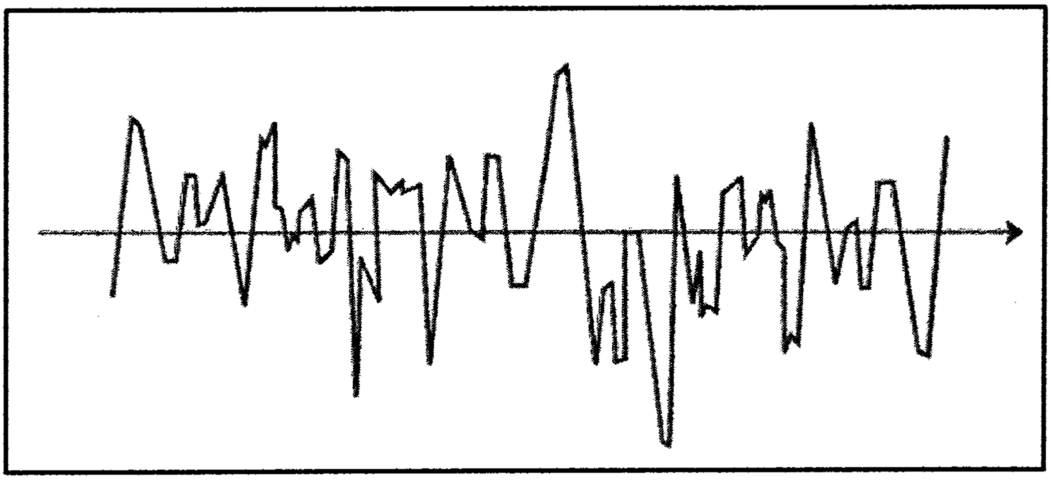


Figure 4.4: ***A noise signal (from Markowitz, Using Speech Recognition, Copyright 1996, Prentice Hall)***

4.2 Capturing the Speech Signal

Capturing the speech wave is the first step to be done by a speech recognition system. The system starts by transforming the speech signal into a processable form, using a microphone, by converting it into an electrical signal. This electrical signal, which is an analog signal, is then changed into digital form using digitization. The reason for digitizing the speech signal is that digital techniques achieve a guaranteed accuracy and facilitate highly sophisticated signal processing which cannot be realized by analog techniques. The digitization process could be done using special digital signal hardware, but in this research digitization is done using the audio sound card. Figure 4.5 shows an analog signal and its corresponding digital signal.

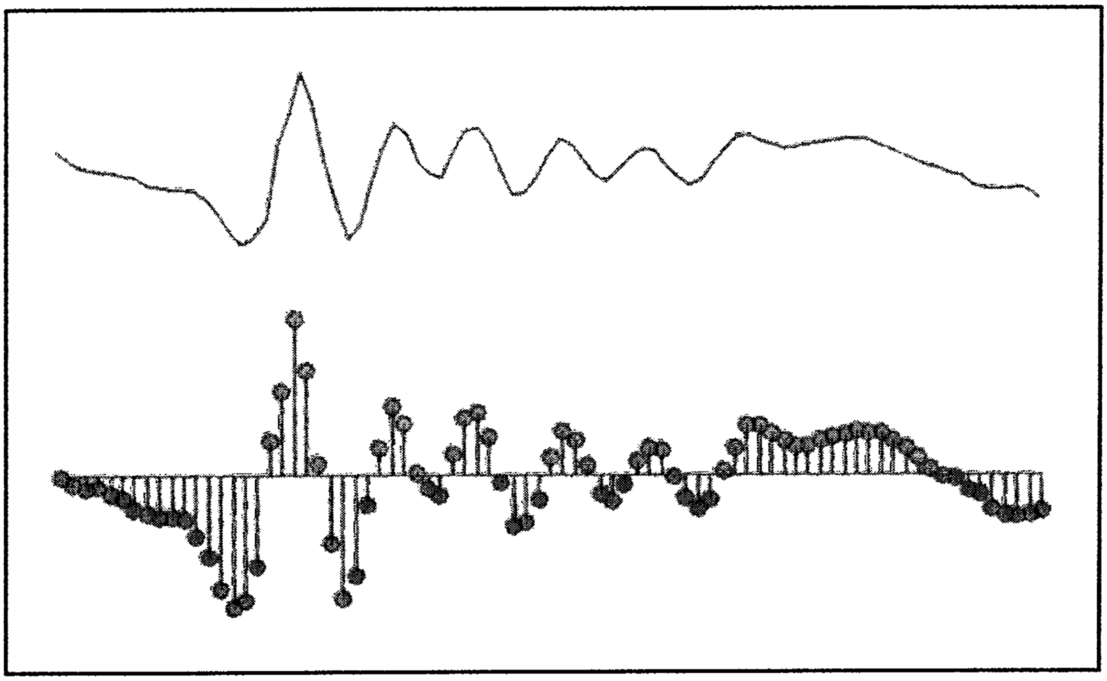


Figure 4.5: *Analog signal and its corresponding digital signal (adapted from Huang, Spoken Language Processing, Copyright 2001, Prentice Hall)*

Without the use of digitization, the quantity of speech data would be so great that the processing and storage requirements would be prohibitive. In order for a speech recognition system to function at an acceptable speed, the amount of data must, be reduced. Speech in addition to sounds contains also noise patterns and silences. Therefore, some data in the speech signal are redundant, some are irrelevant to the recognition process, and some need to be removed from the signal because they interfere with accurate recognition. The challenge is to eliminate these useless components from the signal without losing or distorting critical information contained in the data. This is done by choosing appropriate parameters for the digitization process [12].

Digitization is the process of converting the electrical speech signal into numerical values, in other words representing it mathematically as a function of a continuous variable t that represents time. Digitization can be divided into the following sub processes:

- Sampling
- Quantization
- Coding

Sampling is the process for depicting a continuously varying signal as a periodic sequence of values (or samples). It is similar to taking snapshots of a motion picture at regular intervals. So, if we define an analog signal $x_a(t)$ as a function varying continuously in time, it can be converted into a sequence (sampled sequence) of values $\{x_i\} = \{x(iT)\}$ at a periodic time $t_i = iT$ (i is an integer), as depicted in figure 4.6. Here, T is called the sampling period, and its reciprocal F = 1/T (Hz) is the sampling frequency. If T is too large; the original signal cannot be reproduced from the sampled sequence. For example, F = 8 kHz corresponds to a sampling period T of 125 microseconds [5, 6, 12].

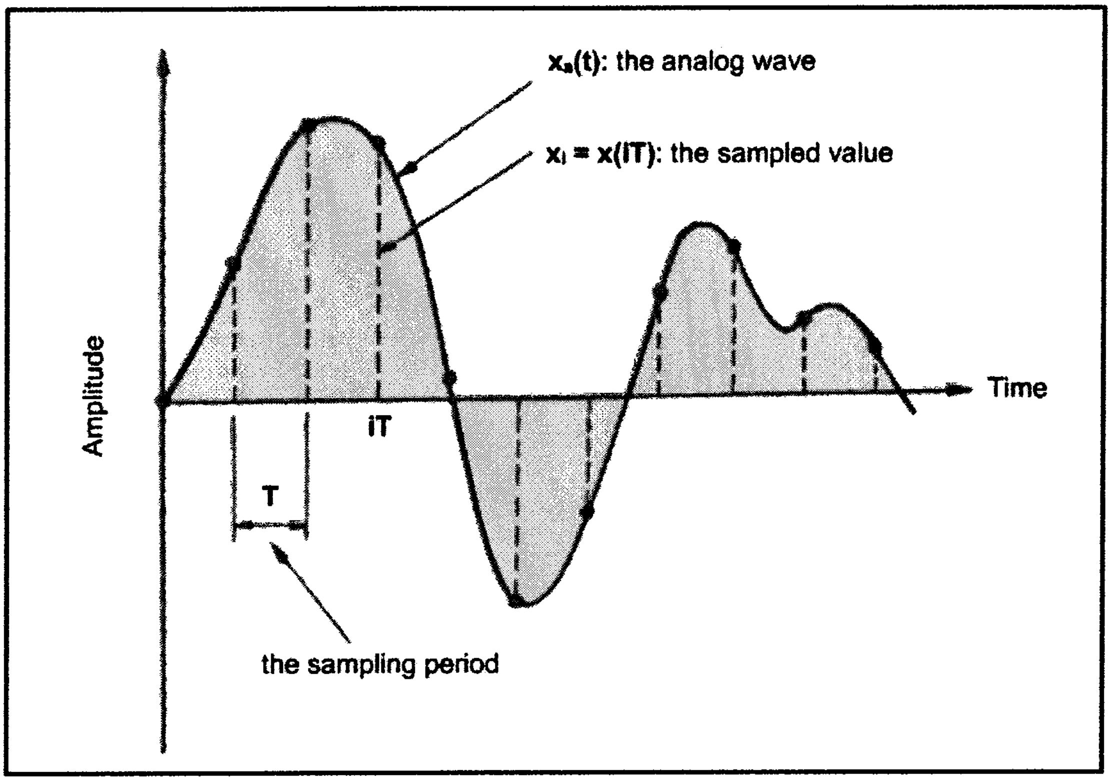


Figure 4.6: ***Sampling of the signal in the time domain (adapted from Furui, Digital Speech Processing, Synthesis, and Recognition, Copyright 2001, Marcel Dekker)***

The number of samples that must be extracted from the signal depends upon the sampling period to be included in the analysis. If T is too small, useless samples for the original signal reproduction is included in the sampled sequence. Also, if T is too large, the original signal cannot be reproduced from the sampled sequence. Along these lines and following Shanon-Someya's Sampling theorem , the minimum sampling rate is set at twice the rate of the highest frequency of interest. This insures that the start, middle, and end of the wave cycle at that frequency will be captured. Shanon-Someya's Sampling theorem says that when the analog signal x(t) is band restricted between 0 and W Hz and when x(t) is sampled at every T = 1/(2W) (s), the original signal can be completely reproduced. For example, a regular telephone signal can be sampled every T = 1/8000 (s), since its bandwidth is restricted under 4 kHz. In this research the sampling frequency used is 16 kHz [5, 6, 12].

The digitization process continues with quantization which involves approximately representing a waveform value by one of a finite set of values. While the coding process is concerned by assigning an actual number to each value. The whole digitization process, which is depicted in figure 4.7, thus enable a continuous analog signal to be converted into a sequence of codes selected from a finite set [5].

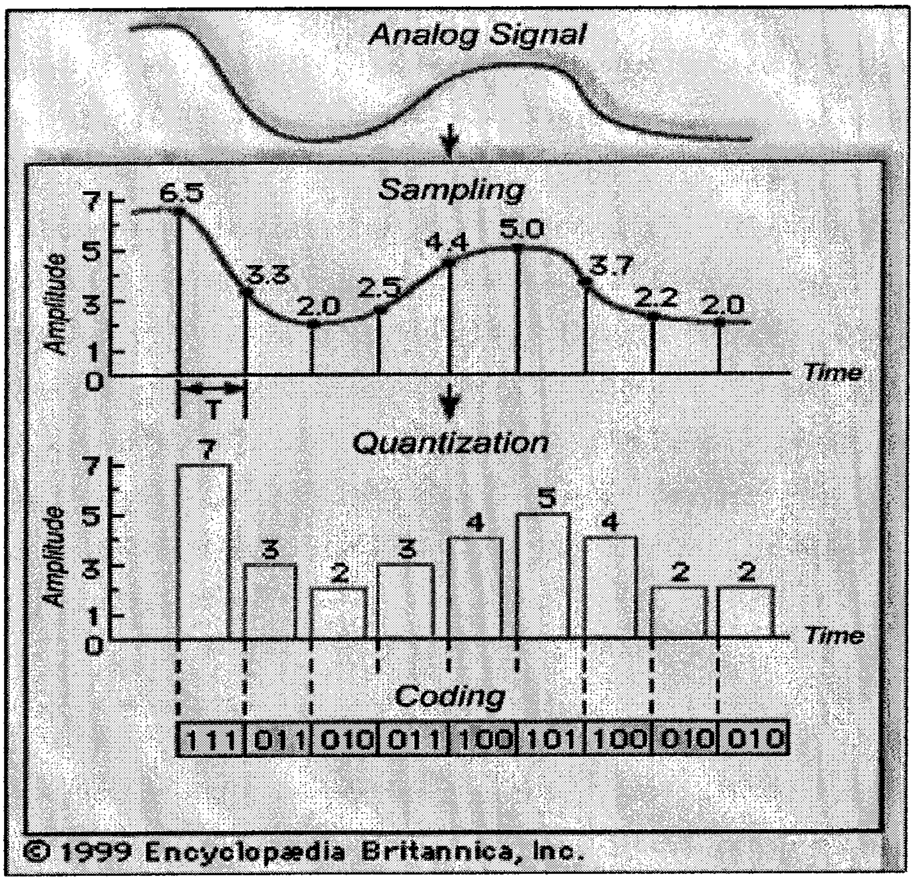


Figure 4.7: ***Digitization of a speech signal***

4.3 Signal Processing

The role of a signal processing module in a speech recognition system, is to reduce the data rate, to remove noises, and to extract noticeable features that are useful for subsequent acoustic matching. To perform speech recognition, a number of components such as digitizing speech, feature extraction and transformation, and acoustic matching can be pipelined time-synchronously from left to right. In this section, the components needed to recognize speech are presented in the order of how they are applied to the speech signal. For every component, a general description and comparison, when possible, of the popular research methods currently used will be given.

4.3.1 Signal Acquisition or Digitization

Most of the necessary speech signal acquisition tasks can be handled in these days through software. For example, most PC sound cards have direct memory access, and the speech can be digitized to memory without burdening the CPU with input/output interrupts. The operating system can correctly handle most of the necessary AD/DA functions in real time.

The used memory buffer to digitize speech typically ranges from 4 to 64KB with 16kHz sampling rate and 16-bit A/D (analog to digital) precision. 16 kHz sampling rate is sufficient for the speech bandwidth (8 kHz). Table 4.1 shows some empirical relative word recognition error increase using a number of different sampling rates. So, if we take the 8kHz sampling rate as our baseline, the word recognition error can be reduced by about 10% if we increase the sampling rate to 11 kHz. A further increase to the sampling rate making it 16 kHz, will further reduce the word recognition error rate by 10%. Further increase to the sampling rate does not have any additional impact on the word recognition errors [6].

Table 4.1: ***Sampling rates and their effect on error reduction***

Sampling Rate	Relative Error-Rate Reduction
8 kHz	Baseline
11 kHz	+10%
16 kHz	+10%
22 kHz	+0%

4.3.2 Speech Period Detection

Detection of the speech period is the first stage, after the signal acquisition, of a speech recognition system. This is a particularly important stage because it is difficult to detect the speech period correctly in noisy surroundings. Its importance lies in the fact that any detection error usually results in a serious recognition error. For instance, the inclusion of the speaker's breath with the word uttered might cause a possible error in recognition. The continuously listening model listens all the time and automatically detects whether there is a speech signal or not. It needs a so-called speech end-point detector, which is used to filter out obvious silence. This model can be realized by using both log-energy and delta log-energy respectively. Figure 4.8 shows a speech period detection for a signal. The black lines show the correct beginning and end of speech, while the blue line show a speech detection error [5, 6].

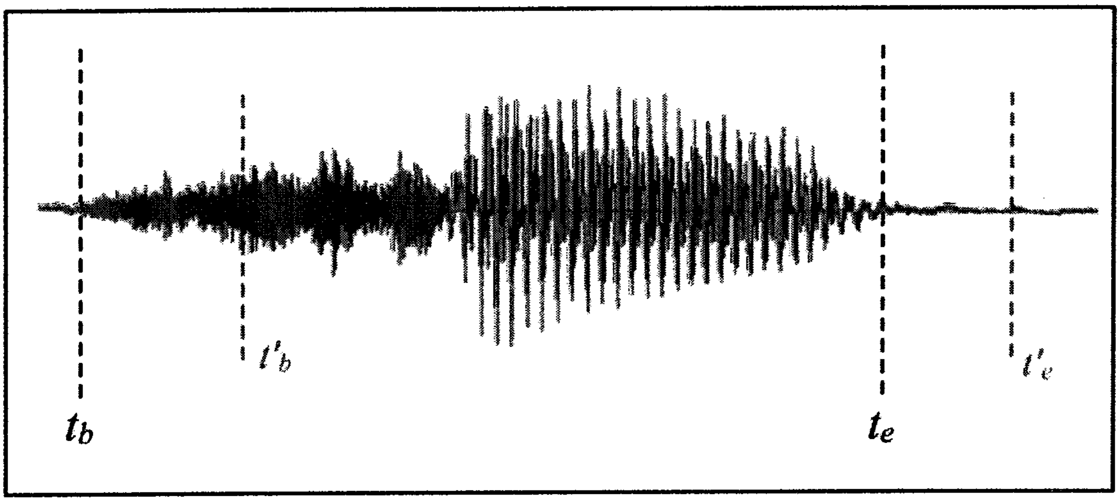


Figure 4.8: ***Speech period detection***

4.3.3 Speech Signal Representations and Feature Extraction

The main requirement for a speech recognition system is the extraction of speech features, which may distinguish different phonemes of a language. Therefore, the extraction of reliable features is one of the most important issues in speech recognition. The number of features to be extracted is crucial. Limiting the number of features might cause a decrease in performance of the recognition process. However, increasing the number of features does not necessary mean an improvement in performance. Increasing the number of features means increasing the training data set, causing more computations and memory usage.

The central theme is to decompose the speech signal into frames, and then pass these frames into a linear time varying filter. The recognition system extracts acoustic patterns contained in each frame and captures the changes that occur as the signal shifts from one frame to the next. The filter could be represented by a speech production model (such as Linear Predictive Coding) or a speech perception model (such as Mel-Frequency cepstrum). A speech production model is a parametric analysis model, where a model that fits the objective signal is selected and applied to the signal by adjusting the feature parameters representing the model. On the other hand, the speech perception model is a nonparametric analysis model, where the model used do not model the signals and could be applied to various signals. Parametric analysis models usually perform better than nonparametric analysis models if the model used by the parametric analysis model thoroughly fits the objective signal.

Two of the most commonly used spectral methods for feature extraction are the

- Mel-Frequency Cepstrum Coefficient
- Linear predictive coding

The Mel-Frequency Cepstrum Coefficient (MFCC) is a representation defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal. The FFT (Fast Fourier Transform) defines the speech signal in terms of its component wave patterns. The FFT is applied to the speech signal during its periodic form. Usually speech signals are periodic, when the phonatory mechanism articulators are in a stable position during the production of a phoneme, at a time period shorter than 100 milliseconds.

Linear Predictive Coding (LPC) also known as Auto-Regressive (AR) modeling, is a very powerful method for speech analysis. This method is predicated on the idea that it is possible to estimate the values of important acoustic parameters from an incoming sample by using the parameter values from previous samples. The popularity of LPC is because it is fast and simple and also to its ability to provide accurate estimates of acoustic parameters with less computation and storage than most other approaches.

In general, time-domain features are much less accurate than frequency-domain features such as the Mel-Frequency Cepstral Coefficients (MFCC). This is because many features such as formants, useful in discriminating vowels, are better characterized in the frequency domain. In addition to that, the calculation of the delta coefficients, that measure the change in the MFCC over time, captures the temporal changes in the spectra. These temporal changes play an important role in human perception.

The relative error reduction with a typical speech recognition system is shown in table 4.2. The 13th-order MFCC outperforms the 13th-order LPC cepstrum coefficients, which indicates that perceptually motivated Mel-scale representation helps recognition. Increasing the number of MFCC (a higher order MFCC) doesn't help in reducing the error rate. This shows that the first 13 coefficients already contain the important information needed for speech recognition. Taking the temporal changes into consideration by calculating the first- and second-order delta features can significantly reduce the word recognition error rate by about 20% [2, 5, 6, 12].

Table 4.2: ***Number of features and their effect on error reduction***

Feature Set	Relative Error Reduction
13th-order LPC cepstrum coefficient	Baseline
13th-order MFCC	+10%
16th-order MFCC	+0%
+1st-and 2nd-order dynamic features	+20%
+3rd-order dynamic feature	+0%

4.3.4 Recognition - Reference Models

Once the preprocessing of a user's input is complete the recognizer is ready to perform its primary function, that is to identify what the user has said. To accomplish this task, recognition systems store models (sometimes called reference models) of the words in an application. During recognition they compare the stream of acoustic parameters uttered by the user with the stored models. The competing recognition technologies found in commercial speech recognition systems are:

- Templates
- Hidden Markov Models
- Sequence of phonemes
- Sequences of sub-words

4.3.4.1 Templates

Template matching is a form of pattern recognition, where each word or phrase in an application is stored as a separate template. The input is then compared with the stored templates, and the template that most closely match the incoming speech pattern is identified as the recognized word or phrase. The selected template is called the best match for the input. Template matching is performed at the word level and contains no reference to the phonemes within the word. The representation is simple, straightforward, and easy to generate. There is no internal structure beyond the sequence of vectors generated by the coding process. No attempt is made to analyze linguistic or acoustic relationships that might exist within the word. The advantages and disadvantages of template matching are shown below:

Advantages:

- Performs well with small vocabularies of phonetically distinct words.
- Midsize vocabularies in the range of 1000-10000 words are possible if the number of vocabulary choices at a one time is kept minimal.

Disadvantages:

- Must have at least one template for each word in the application vocabulary. Therefore, for every diacritization of an Arabic word a distinct template should be included. For instance, a template for 'كَتَبَ' and another template for 'كُتِبَ'.
- Not good with large vocabularies containing words that have similar

sounds (confusable words., e.g., 'جَرَّة' and 'جَرّ').

4.3.4.2 HMM

They are designed to capture and represent patterns of variation. The statistical information embedded in the states and transitions of the HMM contain probability information extracted from multiple tokens of a word. These tokens may be supplied by individual application users, group of speakers, or pre-existing database (called a corpus) of digitized speech.

4.3.4.3 Phoneme model

This approach views words as a sequence of phonemes. It analyzes the input into a string of sounds, which is then converted to words through a pronunciation-based dictionary. Theoretically, it is an attractive approach to speech recognition because it limits the number of representations that must be stored to the number of phonemes needed for a language.

4.3.4.4 Sub-words

Phoneme based approaches provide little assistance handling coarticulation effects. A number of alternative sub-word units have been proposed that derive their information from spoken data rather than theoretical constructs. The most successful sub-word unit is the triphone.

4.3.4.4.1 Triphone

A triphone is not a phoneme, it consists of a phoneme surrounded by contextual information on both sides. They are generally represented as HMM's of three states representing [6, 12]:

- Transition from the preceding phoneme
- The phoneme
- Transition from the phoneme to the following phoneme.

Chapter 5

The Implementation Process

This chapter is concerned with the implementation process of the Colloquial Arabic Speech Recognition Application (CASRA). The first section gives a specific description of the essential components for a speech recognizer. Only the research methods used during implementation are discussed here, more information can be found in the previous chapter or through the resources used there. The chapter then proceeds with a brief description of the tools used to implement the system and the environment needed to deploy it. The chapter ends by showing screens of the graphical user interface of CASRA, demonstrating how the system works.

5.1 Speech Signal Analysis

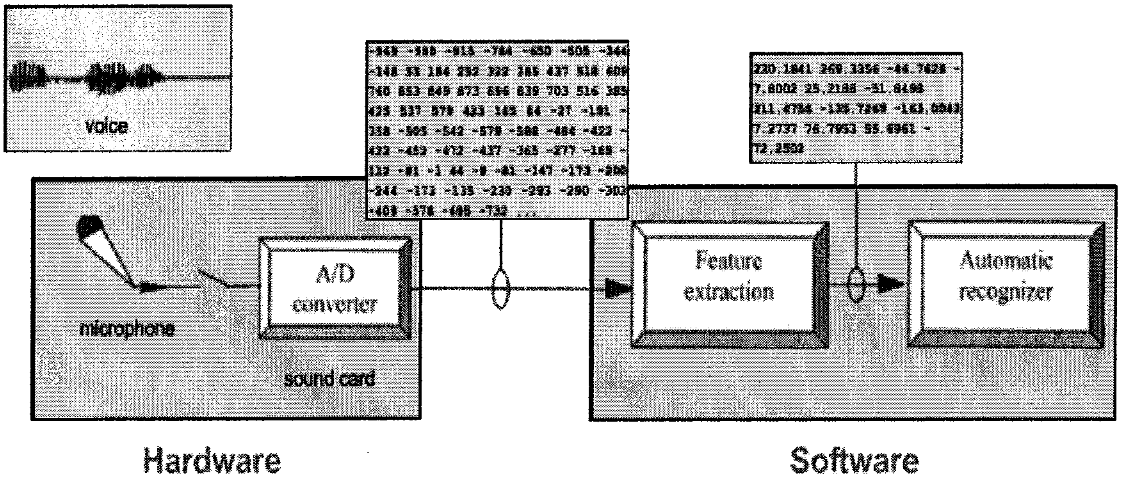


Figure 5.1: ***Structure of a speech signal analysis (adapted from Becchetti, Speech Recognition, Copyright 1999, John Wiley & Son)***

Figure 5.1 shows the structure of a speech signal analysis component in an automatic speech recognition system. The speech analysis, as shown, can be summarized into three main stages, the first is done through hardware while the remaining two are implemented through software. The first stage can be shown as the movement of speech through the microphone, followed by the passage of the microphone output through the A/D converter. The microphone transforms the pressure wave into an electrical analog signal, while the A/D converter digitizes or transforms the analog signal into a digital signal. The second stage is the extraction of the features from a digitized speech signal. The third stage recognizes the word uttered from the features extracted from the speech signal. The rest of this section describes the implementation process of all the steps needed for the analysis of a speech signal.

5.1.1 Signal Acquisition

First, the speech signal is changed into an electrical signal, and this is done by the microphone. Then, this electrical signal, which is an analog signal, should be changed into a digital signal. The digitization of a speech signal, or the transformation of a signal form analog to digital, could either be done through the use of a specific digital signal processor or through the usage of a digital sound card. Due to the ease and availability of the latter, a digital sound card is used in this research.

Setting the parameters of the digitization process has a major effect on the relative error rate of the recognition process, as shown in the previous chapter. A sampling rate of 16kHz with a sampling precision of a 16-bit are chosen. That means, for every second the sound card returns a 16,000 samples or numbers, each number is a double byte integer. The size of the memory buffer used for digitization is 4,096 double bytes or 8KB. In other words, for every seconds there are four memory interrupts.

5.1.2 Speech Period Detection

The detection process, a continuously listening model, is done for a discrete word recognition system, which is easier than continuous speech recognition. An energy and log energy for the speech samples are first calculated. Then, the speech period is detected by the fact that the energy of the speech samples exceeds a threshold for longer than a predetermined period. In addition to the energy level, a zero-crossing number, which is the spectral difference between the input signal and the reference noise spectrum, is also used to help in the detection process [5, 6]. The energy of the speech samples can be calculated using the following mathematical equation:

$$E = \sum(x_i^2)/N \quad \text{(where i: } 0 \leq i < N\text{)}$$

where E stands for the Energy, x_i the speech sample at time i, and N the number of speech samples or the duration of speech used for the detection of energy. The value for N used is 512.

5.1.3 Feature Extraction

In the previous chapter, the most commonly used methods for feature extraction were discussed and compared. This section will only focus on the method that is to be applied to this research. The Mel-Frequency Cepstrum Coefficient (MFCC) is chosen to be the feature extraction method due to the better performance, and the ability of the frequency domain to model adequately the sound. Figure 5.2 shows the components of an MFCC process; the rest of this section will define each component and how it is implemented.

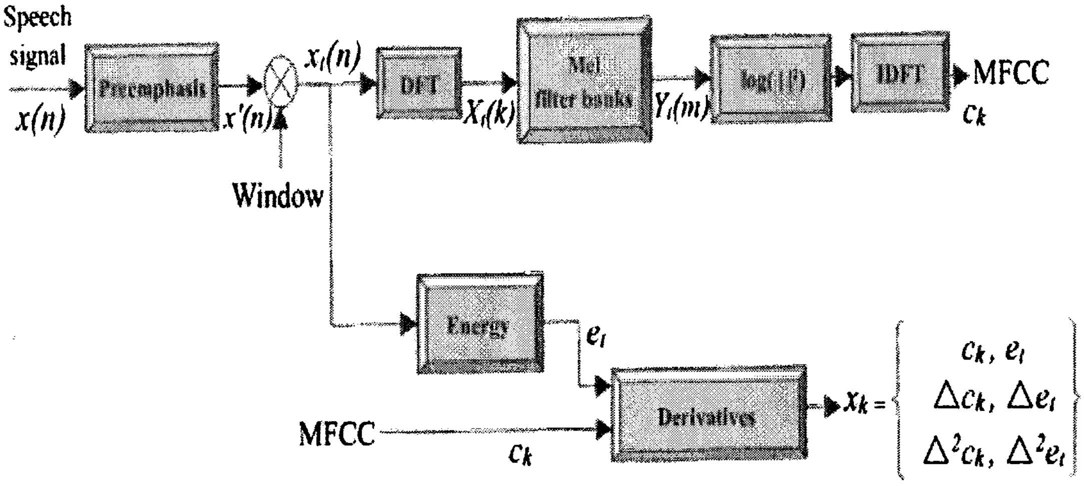


Figure 5.2: ***MFCC processing (adapted from Becchetti, Speech Recognition, Copyright 1999, John Wiley & Son)***

5.1.3.1 Preemphasis

Formants, which are the peaks that result from the resonance of the vocal tract, usually define the structure of a phoneme. The high frequency formants carry with them relevant information, but they have smaller amplitude with respect to law frequency formants. Therefore, an amplitude that is the same for all formants should be attained. This can be done through the use of a Preemphasis filter, which flattens the spectral tilt. Preemphasis can be accomplished after the digitization of a speech signal through the application of the first-order Finite Impulse Response (FIR) filter [2, 5, 6]:

$$H(z) = 1 - \alpha z^{-1}$$

where α is the Preemphasis parameter set to a value close to 1, in this case 0.95. Applying the FIR filter to the speech signal, the preempahsized signal is related to the input signal by the relation:

$$x'(n) = x(n) - \alpha x(n-1)$$

here x' stands for the speech sample after Preemphasis, x the speech sample before Preemphasis, and α the Preemphasis parameter.

5.1.3.2 Windowing

Fourier transform, which will be discussed in the next section, is reliable only when the signal is in a stationary position. For voice, this holds only within a short time interval usually less than 100 milliseconds. Therefore, the speech signal is decomposed into a series of short segments, called analysis frames, then each frame will be analyzed and useful features will be extracted from it. A 512 points frame is chosen in this research, this can be seen in figure 5.3 [5, 6, 7].

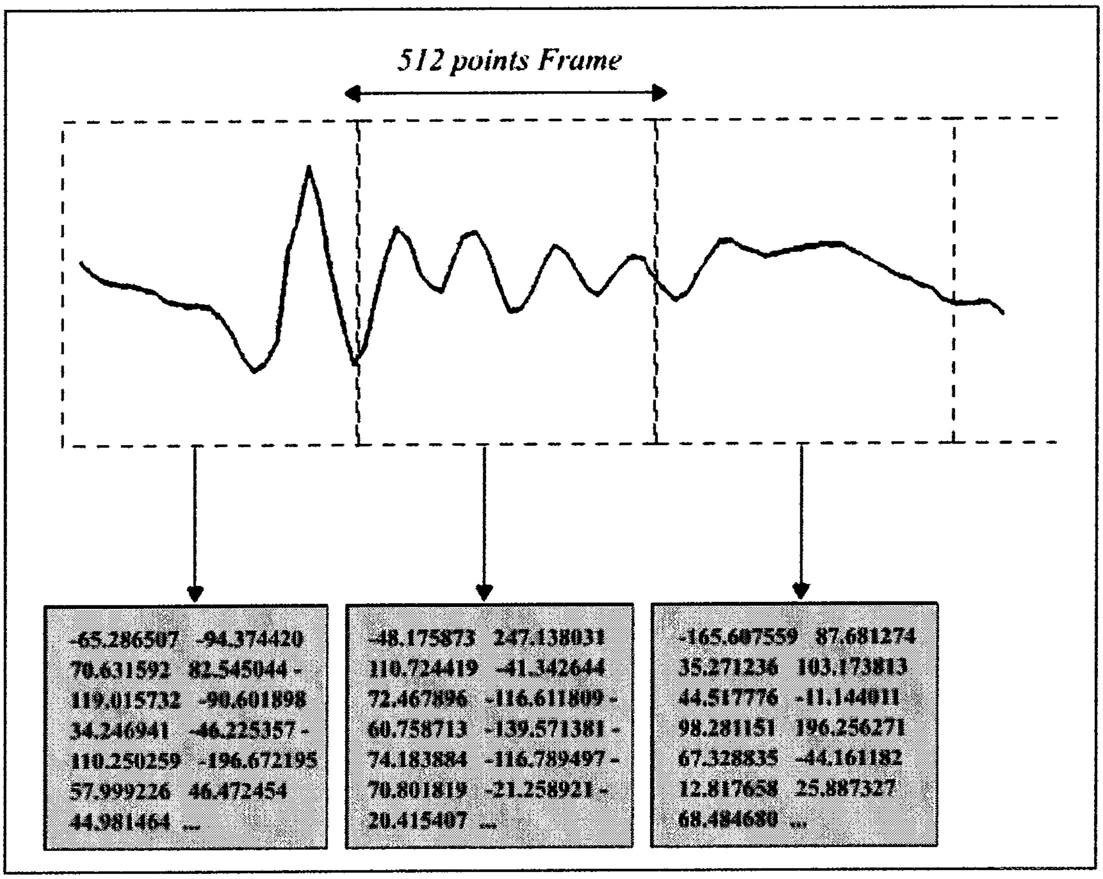


Figure 5.3: ***Frame segmentation of a speech signal***

To minimize the discontinuity and therefore preventing spectral leakage of a signal at the beginning and end of each frame, every frame is multiplied by a window function. Window functions are signals that are concentrated in time, often limited in duration, that consist of a central lobe which contains most of the energy of the window and side lobes which decay rapidly. There are many different window functions, like rectangular, hanning, hamming, triangular, Kaiser, and many others, that can be applied to a speech signal. Here, the hamming window will be used. The characteristics and the application of this window to the speech signal can be seen in figure 5.4 [2, 5, 6, 14].

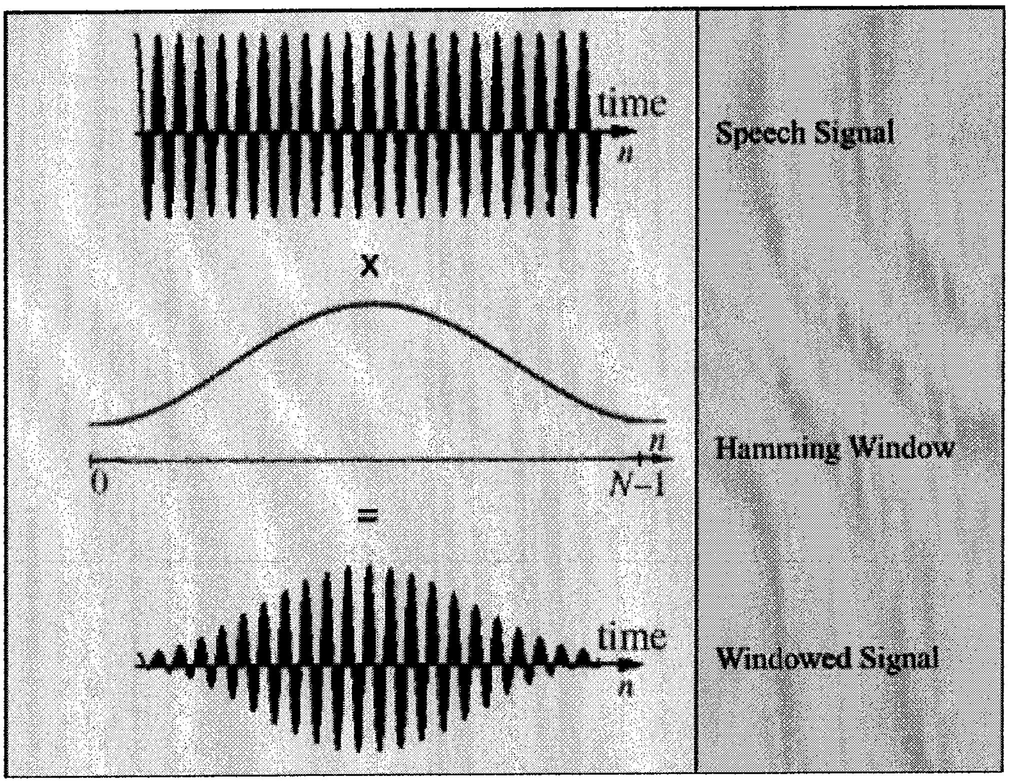


Figure 5.4: Characteristics of a Hamming Window

The hamming window is defined as

$$W_H(n) = 0.54 - 0.46 \cos(2n\pi/N-1)$$

and the application of this window function to the speech signal is

$$x_t(n) = W_H(n).x'(n)$$

where $x_t(n)$ stands for the speech sample at time n after applying the window function, W_H is the hamming window function, and x' is the sampled speech after Preemphasis.

5.1.3.3 Fast Fourier Transform

Discrete Fourier Transform (DFT) is considered to be the basis of spectral analysis, and spectral analysis reveals speech features that are due to the shape of the vocal tract. The Discrete Fourier Transform of a finite duration sequence $\{x(n)\}$ where $0 \leq n \leq N - 1$ is defined as:

$$X(k) = \sum x(n) e^{-j(2\pi/N)nk} = \sum x(n) W^{nk} \qquad \text{where } (0 \leq n, k \leq N - 1)$$

It can be easily seen that W^{nk} is periodic of period N, and this periodicity is the key to the Fast Fourier Transform. The Fast Fourier Transform (FFT) is an algorithm that consists of variety of tricks for reducing the computation time required to compute a DFT. Although FFT algorithms are well known and widely used, they are rather intricate and often difficult to grasp due to the great variety of different FFT algorithms such as radix-4, split-radix, radix-8, radix-16, and decimation-in-time (DIF) algorithms [2, 14].

This research implements the radix-2 algorithm. The idea behind this algorithm is to break the original N point sequence into two shorter sequences. This process continues by iterating, as long as N is an integer power of 2, until two point DFT's are left to be evaluated. The algorithm described here has been called the decimation-in-time (DIT) algorithm, since at each stage of the process, the input sequence is divided into smaller sequences; that is the input sequence is decimated at each stage [6, 14].

5.1.3.4 Mel Filter bank processing

This procedure has the role of smoothing the spectrum, closely modeling the sensitivity of the human ear. The Mel frequency scale is composed of a set of band-pass filters, generally 24 filters are used. The part of the spectrum which is below 1 kHz is usually processed by more filter banks since it contains more relevant information. Mel filters are linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above [2, 6].

5.1.3.5 Log energy and IDFT

After smoothing the spectrum, the logarithm of the square magnitude of the coefficients $y_t(m)$ are computed. The final step in MFCC consists of performing the Inverse Discrete Fourier Transform (IDFT) on the logarithm coefficients. The IDFT can be calculated using the FFT procedure.

Figure 5.5 shows the number of input and output coefficients for every component of the Mel Frequency Cepstrum Coefficients (MFCC). MFCC takes as an input 512 samples and yields 12 Mel cepstrum coefficients plus the zero order MFCC coefficient ck[0], which is approximately equivalent to the log energy of the frame. Figure 5.6 shows the original speech signal and how it is transformed after applying preemphasis, hamming window and the rest of the MFCC components on it.

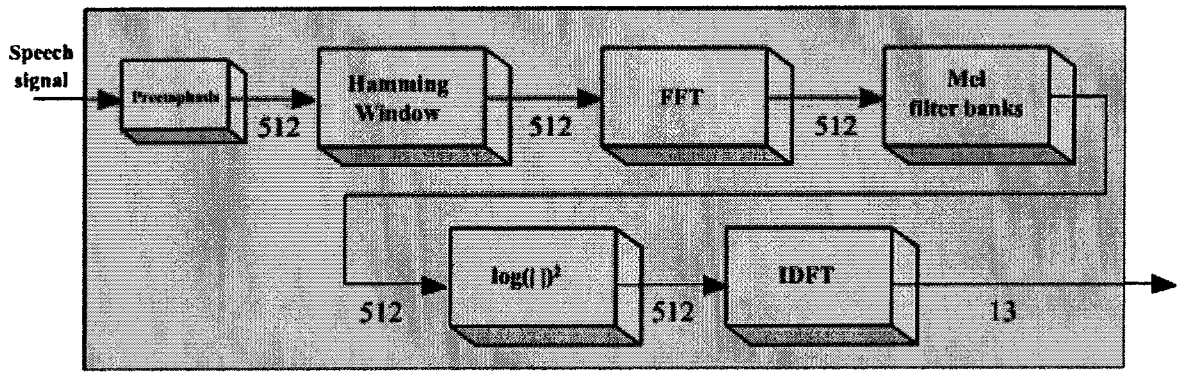


Figure 5.5: ***The MFCC components***

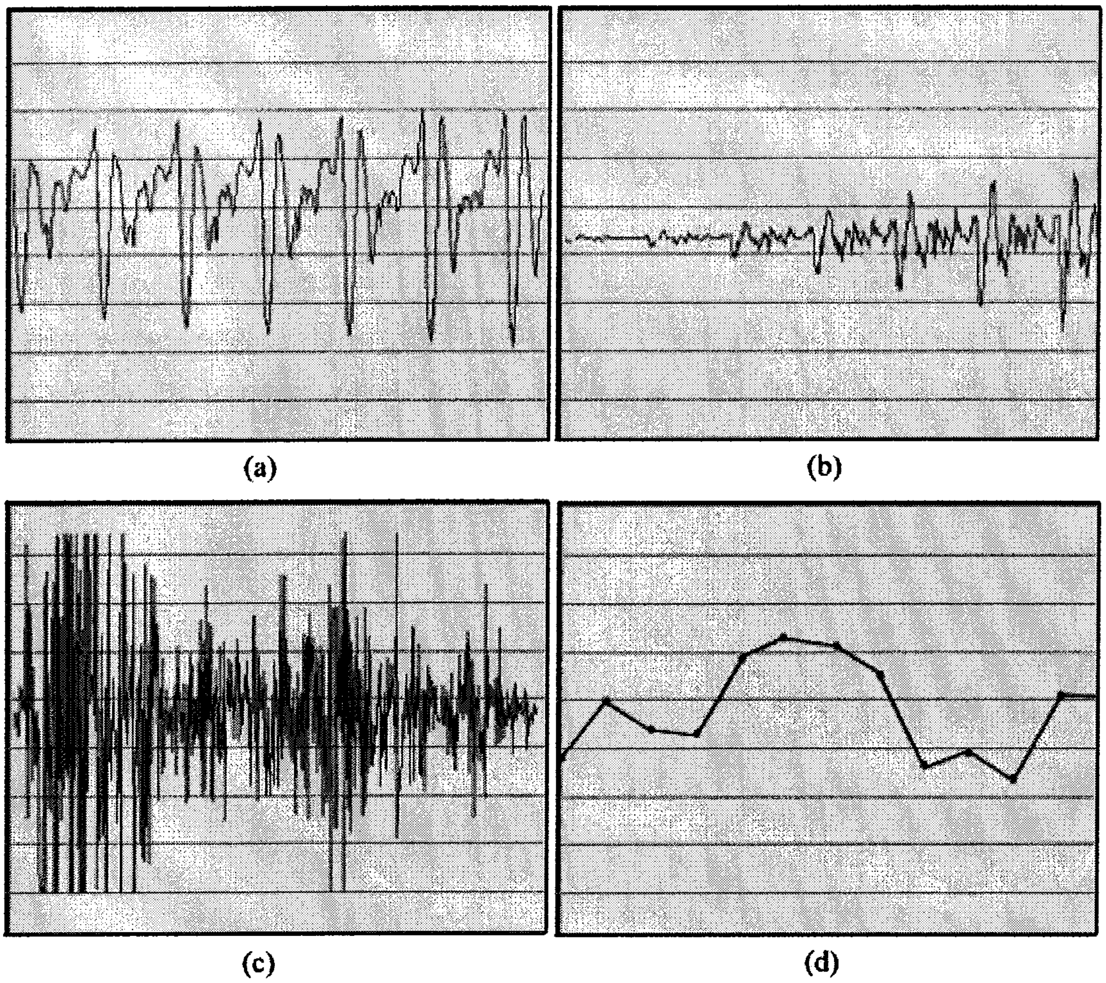


Figure 5.6: ***(a) a fragment of speech wave for the phoneme (kha), (b) after Preemphasis and Windowing, (c) FFT, (d) and MFCC***

5.1.4 Delta Coefficients

Dynamic spectral transitions or features are believed to play an important role in human perception. Based on this idea, the first- and second-delta coefficients are calculated. As shown in figure 5.4 above, in order to calculate the first and second order derivatives the MFCC coefficients and the energy are needed. Although we can compute the MFCC in real-time, the 1st order delta coefficients and the 2nd order delta coefficients (also known as acceleration coefficients) cannot be calculated directly. This is due to the fact that delta coefficients depend on previous and future coefficients. Figure 5.7 shows the dependency of the previous and future frames. As can be seen, the ith 1st order delta coefficient depends on the i-2, i-1, i+1, and i+2 MFCC. Similarly, the ith 2nd order delta coefficient depends on the i-2, i-1, i+1, and i+2 1st order delta coefficients. The 1st and 2nd order delta computation of MFCC coefficients are given by [5, 6, 19]:

$$\Delta c_k = c_{k+2} - c_{k-2}$$
$$\Delta\Delta c_k \text{ or } \Delta^2 c_k = \Delta c_{k+2} - \Delta c_{k-2}$$

where c_k stands for the MFCC coefficients, Δc_k for the 1st order delta coefficients, and $\Delta\Delta c_k$ the 2nd order delta coefficients

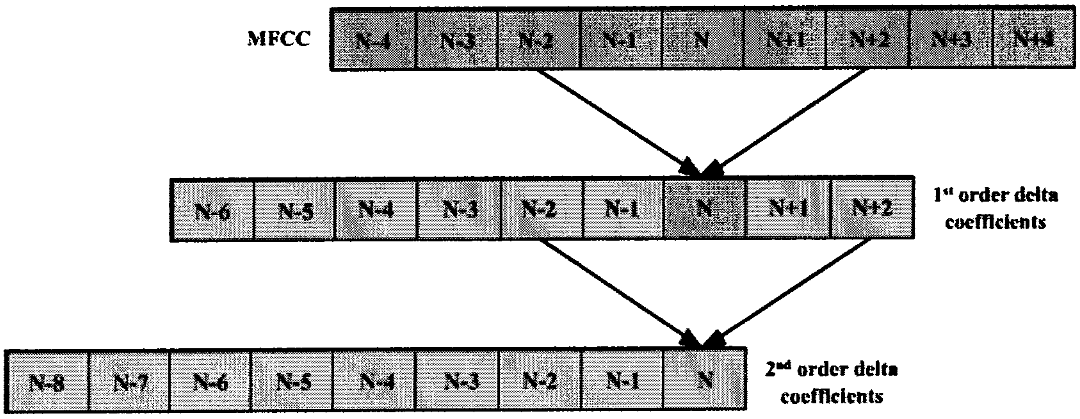


Figure 5.7: ***Delta coefficients computation***

The feature vector used for speech recognition is the combination of these features:

$$x_k = \begin{pmatrix} c_k \\ \Delta c_k \\ \Delta\Delta c_k \end{pmatrix}$$

5.1.5 Vector Quantization

Quantization is the process of assigning discrete values to continuous amplitude signals. While quantization of a single parameter is called scalar quantization, joint quantization of multiple parameter is called vector quantization. Vector Quantization is considered to be one of the most effective methods for reducing the amount of data needed to perform spectra analysis. The input vector x, which is assumed to be d-dimensional, is mapped into another k-dimensional vector y. Therefore, x is quantized into y as shown below [5, 6]:

$$y = q(x)$$

The vector y can be one of the vectors from the finite set of values:

$$Y = \{y_i\} \quad \text{where } 1 \leq i \leq K$$

The set Y is the dictionary of vectors or templates (called codewords) which is referred to as the codebook. K is the size of the codebook, and is also referred to as the number of levels. The quantization process starts by designing the codebook. The d-dimensional space of the vector x is partitioned into K regions or cells $\{C_i, 1 \leq i \leq K\}$, and each cell is associated with a codeword y_i. Then, the input vector is matched against each codeword using some distortion measure. The input vector is then mapped to the codeword y_i if x lies in C_i [5, 6, 12].

$$q(x) = y_i \qquad \text{iff } d(x, y_i) \leq d(x, y_j), \text{ where } i \neq j \text{ and } 1 \leq j \leq K$$

$d(x, y_i)$ is the distortion measure between the two vectors x and y_i. The most commonly used measure is the Euclidean distortion measure, which is defined as follows:

$$d(x,y) = \sum(x - y)^2$$

An example of vector quantization can be shown in figure 5.8. Here, the two-dimensional region is divided into 16 regions or cells. The shaded cell, is the cell that x is mapped or quantized to. That means that the distance between x and y_i has the minimum, or the smallest, distortion measure between all y_j where $1 \leq j \leq K$.

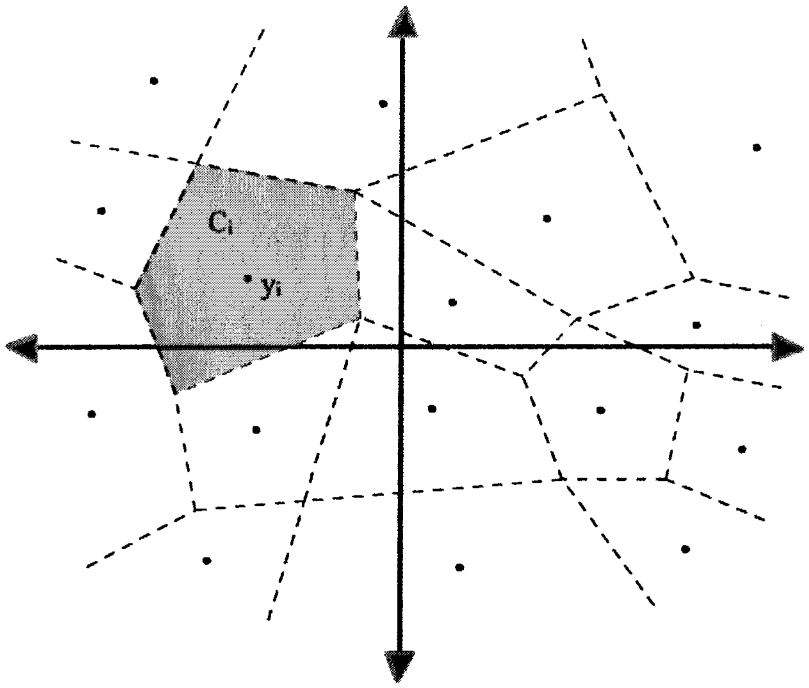


Figure 5.8: ***Vector Quantization-the portioning process (adapted from Huang, Spoken Language Processing, Copyright 2001, Prentice Hall)***

In this research, multiple codebooks is used instead of a single codebook. Three codebooks are used for c_k, Δc_k, and $\Delta\Delta c_k$, respectively. Each codebook represents a set of different speech parameters. In order to combine the multiple output observation of these codebooks, the observations are considered to be independent, and the result is the product of the probabilities of each codebook. The usage of multiple codebooks adds the following two advantages [6]:

- Multiple codebooks for a number of features, if used, can significantly improve performance.
- Multiple codebook can increase the representation power of the VQ codebook and can substantially improve speech recognition accuracy. In comparison to building a single codebook for x_k, the multiple codebook system can reduce the error rate by more than 10%.

5.1.6 Dynamic Time Warping

Using word as a unit of recognition adds more complication to the recognition process. The speech signal is a time dependent process, therefore several utterances of the same word are likely to have different durations. Even the same word with the same duration might differ in the middle. The difference in duration is due to the different rates used while uttering the different parts of a word. For example, the word 'كيفك' has a different duration when the 'ي' is emphasized in the utterance 'كيييفك', and is also different in the utterance 'كيفاااك' when 'فَ' is emphasized.

This problem can be solved through the use of Dynamic Time Warping (DTW). DTW nonlinearly expands or contracts the time axis to match the same phoneme position between the input speech, or the word uttered, and the reference template. The DTW process can be efficiently accomplished by using the Dynamic Programming (DP) technique. The DTW process can be represented mathematically as follows [6, 18]:

$$g(i,j) = \min \begin{pmatrix} g(i, j\text{-}1) + d(i,j) \\ g(i\text{-}1, j\text{-}1) + 2d(i,j) \\ g(i\text{-}1, j) + d(i, j) \end{pmatrix}$$

Here, d(i,j) is the local distance between i and j, and g(i,j) stands for the global distance. An appropriate distance measure is the Euclidean distance. An illustration of how DTW works is shown in figure 5.9. The x-axis represents the input word, while the y-axis represents the template reference. The uttered word ‘كييفاك’ is matched here with the template word ‘كيفك’.

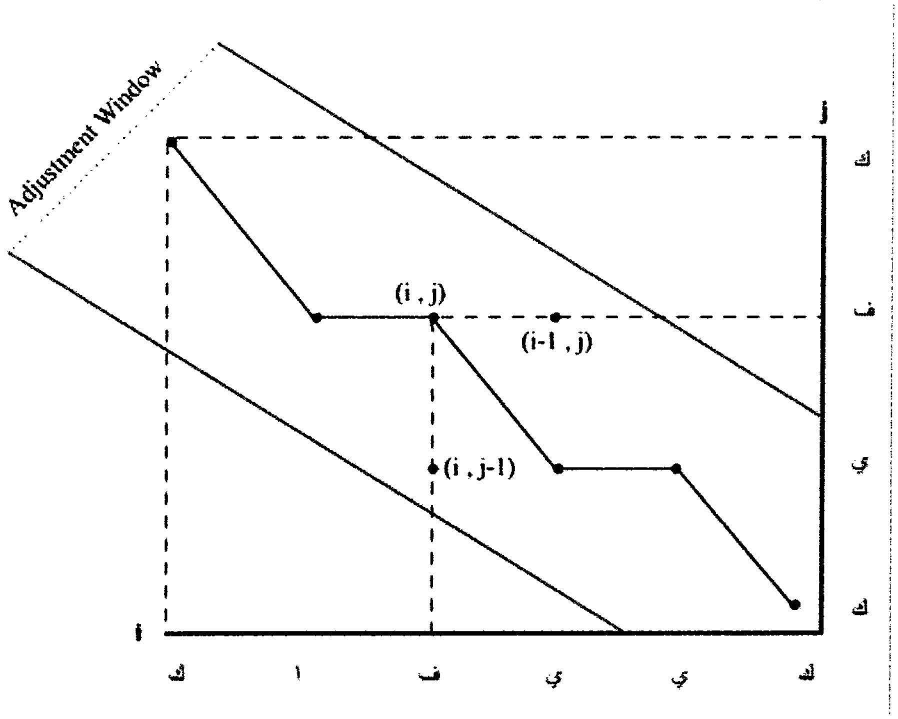


Figure 5.9: *Application of DTW*

5.2 The Tools and Environment

The speech recognition system is implemented using the C++ language with the support of MFC libraries. The compiler used is Visual Studio version 6. Therefore, the application only runs on a Microsoft Windows operating system. In order for the application to run, only an ini file is needed. Additional libraries are not needed for operating systems that do not have Visual Studio installed on them.

5.3 The Application

The application, as shown in figure 5.10, is a simple editor that can write Arabic or English language. The default direction of writing is the right. In addition to the tasks and functions found in a standard editor, the toolbar contains two buttons. The first button enables the start of speech recognition, the latter halts the recognition process. The second button can not be seen because initially it is disabled.

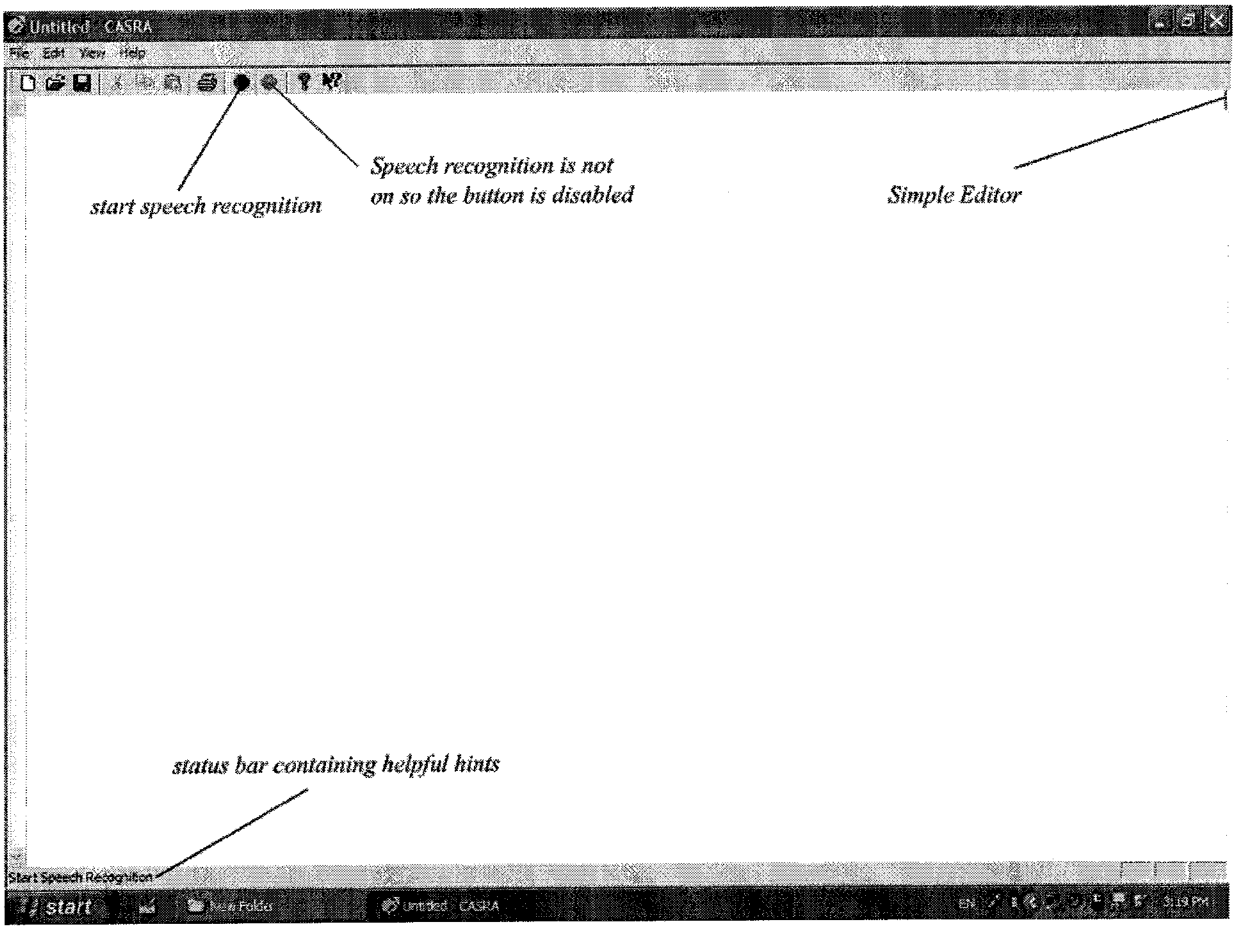
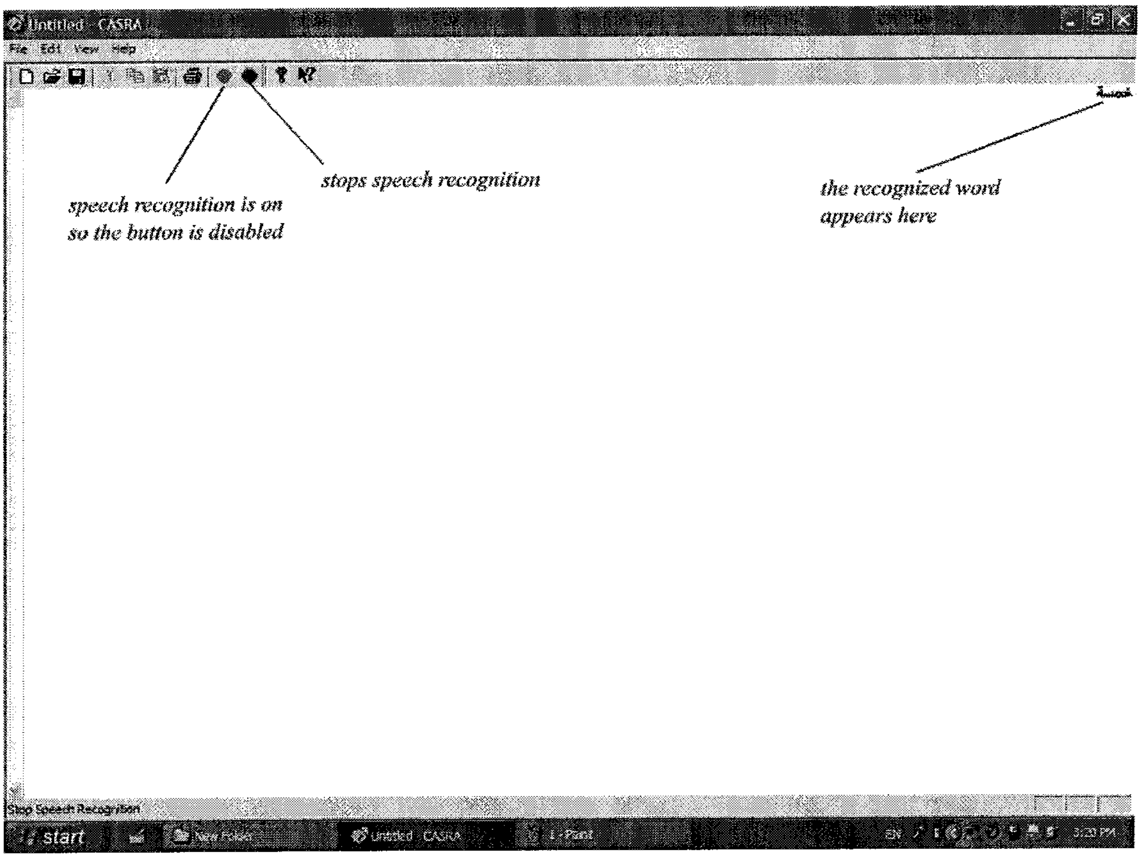


Figure 5.10: ***CASRA – the speech recognition is not started***

When the user starts the speech recognition process, as shown in figure 5.11, the first button gets disabled while the second button gets enabled. The recognition now accepts voice and send the recognized words immediately to the editor. In order to stop the recognition process, the second button must be pressed.



speech recognition is on so the button is disabled

stops speech recognition

the recognized word appears here

Figure 5.11: ***CASRA – the speech recognition is on***

Chapter 6

Testing and Performance Evaluation

This chapter is concerned with how a speech recognition system can be evaluated and tested. The chapter is divided into two sections. The first section gives the theory to how a speech recognition system can be evaluated. The next section describes the process of how the proposed speech recognition system is evaluated, and what kind of experimental tests are used.

6.1 Evaluating the Performance

Evaluating the performance of a speech recognition system is a vital step. It helps in determining the level of acceptance of the system, and the ability of the system to cope with its environment. The word recognition error rate, considered to be one of the most important measures, is used here. Recognition performance is evaluated by comparing the true sequence of units with the recognized sequence of units. It is unreasonable to simply compare two word sequences one by one in order to calculate the error rate. Therefore, as an example, the sentence 'و لقد غادرت من الساعة الثانية الى الساعة الخامسة.' recognized as 'لقد غادرت من الساعة الثامنة الى الساعة الخامس هي.' has an error rate of 100%. In fact, the error rate is only 45%. There are three basic classes of errors made by a recognition system [2, 5, 6]:

- Deletion: The speaker utters a word and the system omits it. For example, the system hears ' و لقد غادرت' when the speaker says ' ولقد غادرت'.
- Substitution: A spoken word is replaced with another, usually similar, word. For example, the system hears ' الثامنة' when the speaker says ' الثانية'.
- Insertion: The recognizer adds a word although the speaker didn't say it. For example, the system hears ' الى الساعة الخامس هي' when the speaker says ' الى الساعة

الخامسة".

The three types of errors are evaluated as follows:

% insertion = I/Nr

% substitution = S/Ns

% deletion = D/Ns

where:

I: Number of inserted words.

S: Number of substituted words.

D: Number of deleted words.

Nr: Number of words of the recognized speech that is for evaluation.

Ns: Number of words of the correct speech that is for evaluation.

The performance of the recognition system is usually measured according to % correct or % accuracy:

% correct = (Ns - S - D)/Ns * 100

% accuracy = (Ns - S - D - I)/Ns * 100

% error rate = 100 - % accuracy

6.2 Testing the System

The evaluation of a speech recognition system is done by testing the vocabulary of the system with a corpus of words or sentences, in this case words. The vocabulary of the system is composed of speech utterances of the words to be recognized by the system. The speech utterances could be collected from one speaker or many speakers. In this research, the vocabulary is only from one speaker, therefore the system is a speaker dependent one. The corpus chosen usually could be speech uttered by the speaker used in the vocabulary or speech uttered by a different speaker. The speech input and the vocabulary templates are from the same speaker, this is because, as mentioned above, the system proposed is a speaker dependent system. In order to evaluate the speech recognition system, two testing sets were applied. In the first test, the speaker uttered each digit ten times. In the second test, the speaker uttered each digit fifteen times. The vocabulary contained two utterances for every digit. Tables 6.1 and 6.2 show the results of the two testing sets respectively:

Table 6.1: ***Testing result for 10 utterances for every word***

Digit	Articulation	Correct	Substitution	Deletion	Insertion	%accuracy	%error rate
واحد	واحَدْ	8	2	0	0	80	20
إثنين	تِنِينْ	5	4	1	0	50	50
ثلاثة	تِلِيتِي	6	4	0	0	60	40
أربعة	أَرْبَعَ	7	3	0	0	70	30
خمسة	خَمْسِ	9	1	0	0	90	10
ستة	سِتٍّ	8	2	0	0	80	20
Total		**43**	**16**	**1**	**0**	**71.66**	**28.34**

Table 6.2: ***Testing result for 15 utterances for every word***

Digit	Articulation	Correct	Substitution	Deletion	Insertion	%accuracy	%error rate
واحد	واحَدْ	10	5	0	0	66.67	33.33
إثنين	تِنِينْ	9	5	1	0	60.00	40
ثلاثة	تِلِيتِي	8	7	0	0	53.33	46.67

أربعة	أَرْبَعَ	11	4	0	0	73.33	26.67
خمسة	خَمْسِ	14	1	0	0	93.33	6.67
ستة	سِتٍّ	13	2	0	0	86.67	13.33
Total		**65**	**24**	**1**	**0**	**72.22**	**27.78**

A second experiment was done to further test the system. In this time, the vocabulary contained the days of the week. The vocabulary contained at least two to three utterances of every day in the week. Due to the variations in the structure of the word 'جمعة', seven utterances were included in the vocabulary. In the first test, the speaker uttered each day fifteen times. In the second test, the speaker uttered each day twenty five times. Tables 6.3 and 6.4 show the results of the two testing sets respectively:

Table 6.3: ***Testing result for 15 utterances for every word***

Day	**Articulation**	**Correct**	**Substitution**	**Deletion**	**Insertion**	**%accuracy**	**%error rate**
إثَنين	تَتِينْ	13	2	0	0	86.67	13.33
ثُلاثاء	تَلِيتَا	11	4	0	0	73.33	26.67
أَربِعاء	أُرْبَعَ	10	5	0	0	66.67	33.33
خميس	خَمِيسْ	15	0	0	0	100	0
جمعة	جُمْعَ	12	3	0	0	80	20
سبت	سَبِتْ	13	2	0	0	86.67	13.33
أحَد	أَحَدْ	15	0	0	0	100	0
Total		**89**	**16**	**0**	**0**	**84.76**	**15.24**

Table 6.4: *Testing result for 25 utterances for every word*

Day	Articulation	Correct	Substitution	Deletion	Insertion	%accuracy	%error rate
إثنَين	تَتِينْ	20	5	0	0	80	20
ثُلاثاء	تَلِيتَا	17	8	0	0	68	32
أَربِعاء	أُرْبَعَ	15	10	0	0	60	40
خميس	خَمِيسْ	20	5	0	0	80	20
جمعة	جُمْعَ	20	5	0	0	80	20
سبت	سَبِتْ	20	5	0	0	80	20
أحَد	أَحَدْ	21	4	0	0	84	16
Total		**133**	**42**	**0**	**0**	**76**	**24**

Chapter 7

Conclusion and Future Work

This thesis covered the implementation of an Arabic speech recognition system concentrating on the Lebanese dialect. The system proposed here is:

- Speaker dependent model: the system must be trained by a specific user before the recognition process, and trying to recognize the speech of a different user with out changing the system vocabulary will yield unsatisfactory results.
- Isolated word model: the speaker should pose between every word uttered to enable the system to detect word boundaries.
- Word based model: the system recognizes speech utterance based on word model.
- Moderate noise toleration: the vocabulary system is implemented and tested in a quiet environment. The tolerance of the system to noise is moderate.

The restrictions applied here (small vocabulary, speaker dependency, and discontinuity of speech) during the implementation process is due to the broadness of the field of speech recognition. Therefore, in order to compress the time required for the implementation and testing processes in to a suitable one, those restrictions were applied.

Preliminary testing done on the system showed acceptable results. Two testing sets were applied; the first set was compromised of ten utterances of every digit, while the second set was fifteen utterances of every digit. The system vocabulary used contained two utterances of every digit from one to six. Another experimentation was done, this time the vocabulary contained the days of the week. Two testing sets also were applied; the first set was compromised of fifteen utterances of every day, while the second set was twenty five utterances of every day. The results showed further improvements on the preliminary results.

Future work will focus on making the system a speaker independent one. This will enable the system to be used by many users with no need to train the speech samples. Further work could be changing the model to accept continuous speech, and changing the word-based model into a phoneme model.

References

[1] Ancient Scripts, May 2005, http://www.ancientscripts.com/arabic.html.

[2] Becchetti, Claudio, and Lucio Prina Ricotti, *Speech Recognition: Theory and C++ Implementation*, 1999, Chichester, John Wiley & Sons.

[3] Cooley, J. W. and O. W. Tukey, *An Algorithm for the Machine Calculation of Complex Fourier Series,* 1965, Math. Comput. Vol. 19, pp. 297-301.

[4] El Choubassi, M. M., et al., *Arabic Speech Recognition Using Recurrent Neural Networks*, December 2003, IEEE International Symposium on Signal Processing and Information Technology ISSPIT, Germany.

[5] Furui, Sadaoki, *Digital Speech Processing, Synthesis, and Recognition*, 2001, New York, Marcel Dekker, Inc..

[6] Huang, Xuedong, Acero, Alec, and Hon Hsiao-Wuen, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 2001, Upper Saddle River, Prentice Hall.

[7] Hyun, Donghoon, and Chulhee Lee, *Optimization of Mel-Cepstrum for Speech Recognition*, 1999, IEEE SMC '99, pp. 500-503.

[8] Jurafsky, Daniel S., and James H. Martin, *Speech and Language Processing*, 2000, Upper Saddle River, Prentice Hall.

[9] Kirchhoff, Katrin, and Dimitra Vergyri, *Cross-Dialectal Acoustic Data Sharing for Arabic Speech Recognition*, 2004, Proceedings of ICASSP 2004, Montreal, Canada.

[10] Kirchhoff, Katrin, et al., *Novel Approaches to Arabic Speech Recognition: Final Report from the JHU summer workshop 2002*, 2002, Tech. Rep., John Hopkins University.

[11] Lazli, Lilia, and Mokhtar Sellami, *Speaker Independent Isolated Speech Recognition for Arabic Language Using Hybrid HMM-MLP-FCM System*, July 2003, AICCSA, Tunisia.

[12] Markowitz, Judith A., *Using Speech Recognition*, 1996, Upper Saddle River,

Prentice Hall.

[13] Quatieri, Thomas F., *Discrete Time Speech Signal Processing*, 2002, Upper Saddle River, Prentice Hall.

[14] Rabiner, Lawrence R., and Bernard Gold, *Theory and Application of Digital Signal Processing*, 1975, Englewood Cliffs, Prentice Hall.

[15] Rego, Jocelynn, and Jackie O'Neill, *Voice Recognition: Software for the Future,* May 2005, http://www.uri.edu/personal/jreg9435/Index.html.

[16] Sakhr ASR, May 2005, http://www.sakhr.com/Sakhr_e/Products/ASR.htm?Index=2&Main=Products&Sub=ASR.

[17] Tarazy, Fouad Hanna, *Al Aswat wa Makharej Al Hrouf Al Arabiet*, 1962, Beirut, Matbaet Dar Al Kotob.

[18] Wrigley, Stuart N., *Speech Recognition by Dynamic Time Warping: Symmetrical DTW Algorithm,* May 2005, http://www.dcs.shef.ac.uk/~stu/com326/sym.html.

[19] You, Kisun, Kim, Hoyoun, and Wonyong Sung, *Implementation of an Intonational Quality Assessment System for a Handheld Device*, October 2004, INTERSPEECH-ICSLP-2004, pp. 1857-1860.

Secondary References

Bahi, Halima, and Mokhtar Sellami, *A Connectionist Expert Approach for Speech Recognition,* April 2004, The International Arab Journal of Information Technology.

Berthommier, Frederic, and Herve Glotin, *A New SNR-Feature Mapping for Robust Multistream Speech Recognition,* 1999, Proc. Internat. Congress on Phonetic Sci. (ICPhS), XIV, San Francisco.

Chen, J., Huang, Y., Li , Q., and K.K. Paliwal, *Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids*, Feb. 2004, IEEE Signal Processing Letters, Vol. 11, No. 2, pp. 258-261.

Davis, Gillian M., Ed., *Noise Reduction in Speech Application,* 2002, Bocca Ranton, CRC Press.

Hamaker, Jonathan, *Homework #4: Signal-to-Noise Ratio Estimation*, May 1998, http://www.isip.msstate.edu/projects/speech/software/legacy/signal_to_noise/doc/tutorial.pdf.

Jelinek, Frederick, *Statistical Methods for Speech Recognition*, 2001, Cambridge, The MIT press.

Kondo, K., Kamata, H., and Y. Ishida, *Speaker-Independent Spoken Digits Recognition Using LVQ*, June 1994, IEEE Proc. ICNN'94, pp. 4448-4451.

Lee, Chulhee, et al., *Optimizing Feature Extraction for Speech Recognition*, January 2003, IEEE Transactions on Speech and Audio Processing vol. 11, no. 1, pp. 80 -87.

McCowan, Iain A., Morris, Andrew, and Herve Bourlard, *Improving Speech Recognition Performance of Small Microphone Arrays Using Missing Data Techniques*, September 2002, ICSLP, pp.2181-2184.

Navratil, Jiri, Jin, Qin, Andrews, Walter D., and Joseph P. Campbell, *Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models*, April 2003, Invited paper, ICASSP, Hong Kong.

Quatieri, Thomas F., and Robert J. McAulay, *Noise Reduction Using a Soft-Decision Sine-Wave Vector Quantizer*, April 1990, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Albuquerque, New Mexico.

Rutkowski, Tomasz, Cichocki, Andrzej, and Allan Kardec Barros, *Speech Extraction from Interferences in Real Environment Using Bank of Filters and Blind Source Separation*, 2000, WOSPAA.

Schuller, Gerald, Edler, Bernd, and Adele Doser, *A Method for Alias Reduction in Cascaded Filter Banks*, October 2000, 9th IEEE DSP Workshop, Hunt, TX.

Schuller, Gerald, Yu, Bin, and Dawei Huang, *Lossless Coding of Audio Signals Using Cascaded Prediction*, 2001, IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, May 7-11.

Vergyri, Dimitra, Kirchhoff, Katrin, Duh, Kevin, and Andreas Stolcke, *Morphology-Based Language Modeling for Arabic Speech Recognition*, 2004, Proceedings of ICSLP, Jeju, South Korea.

Appendix A

Glossary of Technical Terms

- ***Acoustics***: a science that deals with the production, control, transmission, reception, and effects of sound.
- ***Articulation***: adjusting the vocal tract shape to produce various linguistic sounds.
- ***Articulators***: the parts of the vocal tract that can actually move, such as the tongue, lips, and velum.
- ***Diacritics***: are short strokes placed either above or below a phonetic element (in the case of Arabic a letter)indicating a phonetic value different from that given the unmarked or otherwise marked element.
- ***Digital Signal***: is defined as a discrete-time signal whose values are represented by integers within a range, whereas a general discrete-signal would be represented by real numbers.
- ***Digital Signal Processing (DSP)***: refers to methods for manipulating the sequence of numbers x[n], the speech signal, in a digital computer to obtain a new signal with some desired properties. The acronym DSP is also used to refer to a Digital Signal Processor, that is a microprocessor specialized to perform DSP operations.
- ***Discourse conventions***: the study of linguistic units larger than a single utterance. In other words, it is the knowledge needed to correctly structure conversations.
- ***Discrete Time System***: is essentially an algorithm for converting one sequence (the input) into another sequence (the output).
- ***Formant***: any of several resonance bands held to determine the phonetic quality of a vowel.
- ***Fundamental Frequency***: is the rate at which the vocal cords flap against each other when producing a voiced phoneme.
- ***Linguistics***: the study of human speech including the units, nature, structure, and modification of language.

- ***Morphology***: captures information about the shape and behavior of words in context. Example of that recognizing ‘doors’ as plural.
- ***Phonetics***: the study of speech sounds and their production, classification and transcription.
- ***Phonology***: the science of speech sounds including especially the history and theory of sound changes in a language or in two or more related languages.
- ***Phoneme***: in speech science the term is used to denote any of the minimal units of speech sound in a language that can serve to distinguish one word from another.
- ***Pitch***: the semi-musical rising and falling of voice tones.
- ***Pragmatics***: the study of how language is used to accomplish goals. An example is the usage of polite and indirect language.
- ***Resonance***: is the quality imparted to voiced sounds by vibration in anatomical resonating chambers or cavities.
- ***Resonating Chamber***: are chambers or cavities that exhibit resonance. The throat, mouth, and nose are all resonating chambers.
- ***Semantics***: the study of meaning.
    - ***Lexical Semantics***: the knowledge of the meanings of the component words.
    - ***Compositional Semantics***: knowledge of how the components of words combine to form a larger meaning.
- ***Syntax***: the knowledge needed to order and group words together.