Sudan University of Science and Technology

College of Graduate Studies

Voice Recognition by using Machine Learning

A Case Study of some Rules of Tajweed

التعرف على الأصوات باستخدام تعلم الآلة .

دراسة حالة : بعض أحكام التجويد

BY

Safaa Omer Mohammed Nssr.

Research Summited In Fulfillment of the Requirements

For Master Degree in Computer Science

Supervisor: Dr.Howida Ali Abdel Gader

October 2016

DEDICATION

.

To human prophets and teachers . Prophet Muhammad God Pray upon him.

To the spirit of my father may God grant him rest in peace...

To dear mother ... God give her health and wellness

To my husband and my children ...

To my sisters ...

And to all of the lights a candle to illuminate the darkness

ACKNOWLEDGEMENTS

My gratitude and thanks to my God. I am grateful to my supervisor Dr.Howida Ali Abdelgader, for her invaluable, support and guidance during implementation of this research.

Special thanks are posed to Sudan University of Science and Technology for its funding in carrying out this research special department of Computer Science .I extend my sincere thanks and gratitude and appreciation to all those who contributed effort, or gave me help to accomplish this search. I especially thank to my mother, husband; my children and my sisters. Thanks are due to several of my friends for supporting me, special thanks to my family for their patients and cooperation.

Gratitude to almighty Allah before, after and ever .

Table of Contents

Subjects

CHAPTER ONE: INTRODUCTION

CHAPTER TWO: LITERATURE REVIEW

CHAPTER THREE: IMPLEMENTATION AND RESULTS

CHAPTER FOUR: CONCULTION AND FUTURE WORK

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Voice recognition is considered as one of the most important aspects of machine learning domain. But it still has limited and modest applications in Arabic language. The Holy Quran is the largest container of Arabic language grammar in terms of speaking and utterance as it is considered as a message for all humanity. However, we present within this study a classification model for four different altajweed rules like the Allah name(mofakham, morakaq) and moon and sun L(لام),as we depended on two different kinds of voice features LPC(liner predictive coding),MFCC(Mel-frequency cepstrum), where these two types of features are the most used within the domain of processing voice signaldomain.as we depended on two classifying mechanisms (neural networks and hidden Markov model(HMM)) in order to study all possible cases of those studied rules, then we extracted those features of three different readers(males) and two different readers (female).each of Markov hidden model and neural networks have been trained by using two different types of extracted features and then we tested those trained models in order to obtain final results as to evaluate them. And resulted results in the training of Hidden Markov Models accuracy amount 90% with Allah (moufakhum), 92% with Allah (mourqeq), 83.3% with sunny لام and 80% with moony لام .

The study resulted that neural network able to distinguish between four rules of Tajweed and training samples processing with high accuracy reached 95% with Allah (moufakhum) , 94% with Allah (mourqeq) , 93% with moony لام and 92.3% with sunny لام .

المستخلص

يعتبر علم التعرف على الأصوات من الجوانب الهامة ضمن تعليم الألة. ولكن التطبيقات العملية له ضمن مجال اللغة العربية محصورة و متواضعة . إن القرآن الكريم يمثل أكبر حافظ لقواعد اللغة العربية من حيث اللفظ و النطق كما أنه رسالة للبشرية جمعاء. نقدم في دراستنا نموذجاً للتصنف بين أربع أحكام رئيسية من تجويد القرآن الكريم لفظ الجلالة " مفخم ، مرقق" و اللام الشمسية والقمرية "ال " . لقد اعتمدنا في دراستنا على نوعيين مختلفين LPC,MFCC, **لاستخلاص السمات الصوتية. ويعتبر هذان النوعان من المميزات الأكثر استخداماً ضمن مجال معالجة الإشارة الصوتية . كما تم الاعتماد على نوعين من أليات التصنيف , الشبكات العصبونية** (NN) **و نموذج ماركوف المخفي** (HMM) **لاجل دراسة أغلب الحالات الممكنة للأحكام المدروسة قمنا بعملية استخراج الممميزات الصوتية لثلاثة أنواع منالقرّاء الذكور ونوعين من القراء الإناث. و تمت عملية تدريب كل من الشبكة العصبونية و نموذج ماركوف المخفي بالاعتماد على النوعين للمميزات المستخرجة و من ثم قمنا باختبار النماذج المدربة للحصول على النتائج وتقيمها .واسفرت نتائج تدريب النماذج المخفية على دقة تعرف بلغت 90% مع اسم الجلالة (مفخم) 92% , مع اسم الجلالة (مرقق) , 83.3% مع اللام الشمسية و 80% مع اللام القمرية .**

واسفرت الدراسة بأن الشبكات العصبونية قادرةعلى تمييز الاحكام الأربعة ومعالجة عينات التدريب بدقة أعلى بلغت95% مع اسم الجلالة (مفخم), 94% مع اسم الجلالة(مرقق) , 93% مع اللام القمرية و 92.3% مع اللام الشمسية .

CHAPTER 1

CHAPTER 1

INTRODUCTION

1.1 Overview

Talking is the human language and it is a natural communication phenomenon, as it is god's gift to human kind to distinguish them from other creatures, talking is some kind of generating sound waves by the involuntary movement of the anatomy within the human system of generating the talk and which consists of many channels which aim to form the speaking device, it’s shape differs from one person to another and that is why each one has a unique method of producing words.

Computer scientists have shown interest in this matter since more than four decades in order to make humans able to communicate with computers, by the development of computers and electronics a new need have emerged which lies in communicating with computers because it is considered as one of the best standards to save more time and efforts, manned is still looking to find more simple methods to communicate, those methods include talking recognition applications. Talking recognition was the goal of those researches that have been made for more than four decades and along with the emerging of computing and digital signals, this technique has become more popular because those techniques that are equipped with this method are more usable and they have developed a lot lately and entered most of our daily life fields.

since the Holy Quran is the most important science of them all thus there have been more attempt to preserve the process of Tajweed the holy Quran as to reading it properly, where most Sahaba have taken on their shoulders the task of memorizing, writing and teaching it to the followers (may Allah be pleased with them), besides more scientists have studied the Holy Quran where some of them studied the phonetics and sat the rules of Altajweed [1]

Talking recognition is the process in which the computer will be able to specify the pronouncing word. This means to speak with your computer and then to recognize what you have said properly, there are many methods to perform this procedure, but the key principle is to extract some of the main features of the uttered talk then to process those features in order to recognize this word when it is uttered again. .

Since the process of reading the Holy Quran properly within the same pattern that has been passed to us by god's prophet (peace be upon him) represented a good job as a responding to god's will, besides reading the Holy Quran properly will give a higher profit because the Holy Quran will intercedes for his/her reader The process of listening, correction and repetition of the correct Al-Quran recitation took place in real-time condition. However, this method is believed to become less effective and unattractive to be implemented, especially towards the young Muslim generation who are more attracted to the latest technology.

This chapter will give a brief about the project background: it solve problems and shows objectives and methodology as it shows the study plan which will display chapters and its description as a background to recognize the utterance of Quran, as we discussed how produce and identify the utterances of Quran and recognizing moony and sunny L and maximizing and softening the holiness name as to determine the methodologies that are used to recognize the utterances of Quran as to view the basics of recognizing Tajweed Rules and the various methods that are used within this field.

1.2 Statement of the Problem

Many people entrants new in Islam they have no time to come to memorization rings.

Learning AlTajweed must be done through listening to a good sheikh that has knowledge of AlTajweed as to correcting the recitation and repeating it and this will be hard for women, children and elderly where they may not be able to visit the memorization rings.

AlTajweed rules has to be reviewed periodically may be forgotten.

Difficult to differentiate between sunny ,لام moony لامand Allah moufakhum , Allah mouraqeq when recitation the Holly Quran .

1.3 Research Importance

This research helps to teach some rules of Tajweed for people who cannot attend themselves to memorization ring. Thus we minimize time and effort to learn.

Learning the Holy Quran properly will give a higher profit because the Holy Quran will intercede for his/her reader.

1.4 Hypothesis

1. Use two methods Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) it helps to extract the unique features which improve the identification process.
2. Trained Neural networks give high accuracy in the recognizing process.
3. Build Markov Models for the classification rules have significant impact on the accuracy.

1.5 Research Objective

1. Design System to identify the Arab Speech based on some rules of Tajweed .
2. Focus on the best methods of Feature extraction Arab words.

3. Training two types of kinds of ways of recognition neural network and Hidden Markov Model.
4. To obtain satisfactory results for accuracy.

1.6 Methodology

Study Boundary of word, Allah moufakhum , AlLah moureqeq and sunny لام , moony لام , First feature extraction for the word with (MFCC, LPC) and classification (NN,HMM) .

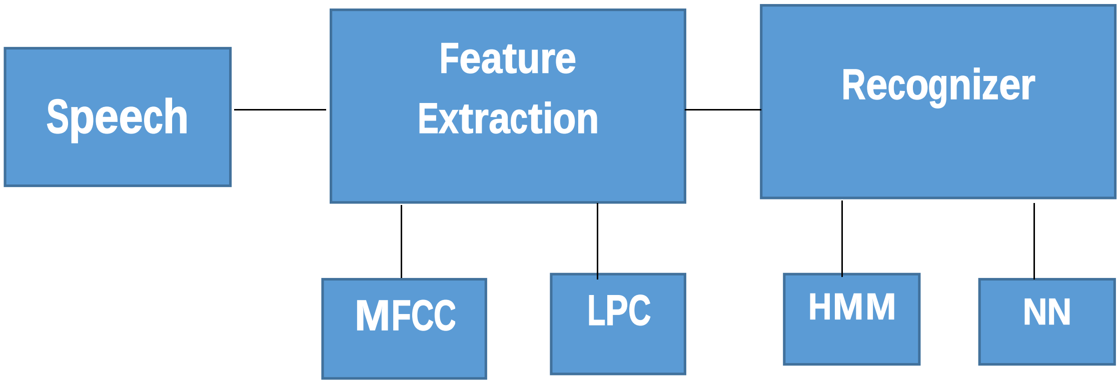


Figure (1.1) Block Diagram of Speech Recognition

1.7 Previous Studies

Reference No	Name Of Paper	Author Name	Technologies	Result
[16]	Automated tajweed checking rules engine for Quranic learning	Jamaliah Ibrahim Noor , Yamani Idna Idris Mohd , Razak Zaidi , Naemah Abdul Rahman Noor	MFCC &Hidden Markov(HMM)	91.95% (ayates) and 86.41 % (phonemes),

[18]	Automatic Extraction Phonetically Rich and Balanced Verses for Speaker–Dependent Quranic Speech Recognition System	Rahmi Yuwan,Dessi Puji Lestari	MFCC & HMM	97.47 %
[19]	MFCC-VQ approach for Qalqalah Tajweed rule checking	Ahsiah Ismail , Mohd Yamani Idna Idris , Noorzaily Mohamed Noor , Zaidi Razak , ZulkifliMohd Yusoff	Mel-Frequency Cepstral Coefficient andVector Quantization (MFCC-VQ) hybridalgorithm	The speed performance of a hybrid MFCC-VQis86.928%, 94.495% and 64.683% faster than theMel-FrequencyCepstral Coefficient for male, female and children respectively
[7]	Comparative Analysis of MFCC, DTW&ANN for Arabic Speech Recognition	Bidoor Noori Ishaq , Bharti W. Gawali ,	MFCC, DTW, Neural Network	The better recognition accuracy of about 90% was obtained with MFCC-based system

1.8 Thesis layout

In this section, we present brief information about the rest of this thesis. The remainder part of this thesis is:

Chapter 2: Literature Review and: Mothodology:

this study a classification model for four different altajweed rules like the Allah name (mofakham , morakaq) and moon and sun L(لام) , as we depended on three different kinds of voice features LPC (liner predictive coding) , MFCC(Mel-frequency cepstrum) , where those two types of features are the most used within the domain of processing voice signal domain.as we depended on two classifying mechanisms (neural networks (NN) and hidden Markov model(HMM)) in order to study all possible cases of those studied rules , then we extracted those features of two different types of readers(males, females) and two famous recitation Quran readers.(Alsudis and Alhuzafi) Each of Markov hidden model and neural networks have been trained by using two different types of extracted features and then we tested those trained models in order to obtain final results as to evaluate them.

Chapter 3: Implementation and Results:

This chapter shows all the work carried out in the project in software and hardware levels and the results of the implementation. The results discussion will also be briefed in this chapter.

Chapter 4: Conclusion and Future work:

This chapter shows a conclusion for the results obtained, features and limitations of the different methods of implementation. Also the possible upgrade and bug removal will be shown in the chapter.

References:

Appendix A:

CHAPTER 2

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

As relevant background to the field of speech recognition, this chapter intends to discuss how the speech signal is produced and perceived by human beings. This is an essential subject that has to be considered before one can pursue and decide which approach to use for speech recognition. Also the basics of speech recognition and different methods used in this field are shown.

We will show also how to design the Tajweed Automation system using two different methods (algorithms) : Mel Frequency Cestrum Coefficients (MFCC) & Linear Predictive Coding (LPC) for feature extraction and for classification we used Neuralnetwork & hidden markov model techniques will be shown and discussed.

2.2 Discrete Time Speech Signal Processing

Speech has evolved as a primary form of communication between humans. We converted speech signal to another form, one early case of this converted is the transduction by a telephone handset of continuously – varying speech pressure signal at lips output to a continuously – varying (analog) electric signal.

The resulting signal can be transmitted and processed electrically with analog circuitry and then transduced back by the receiving handset to speech pressure signal. Digital (A|D) converter has entered as further transduction that samples the electrical speech samples e.g. 8000 samples per second for telephone speech

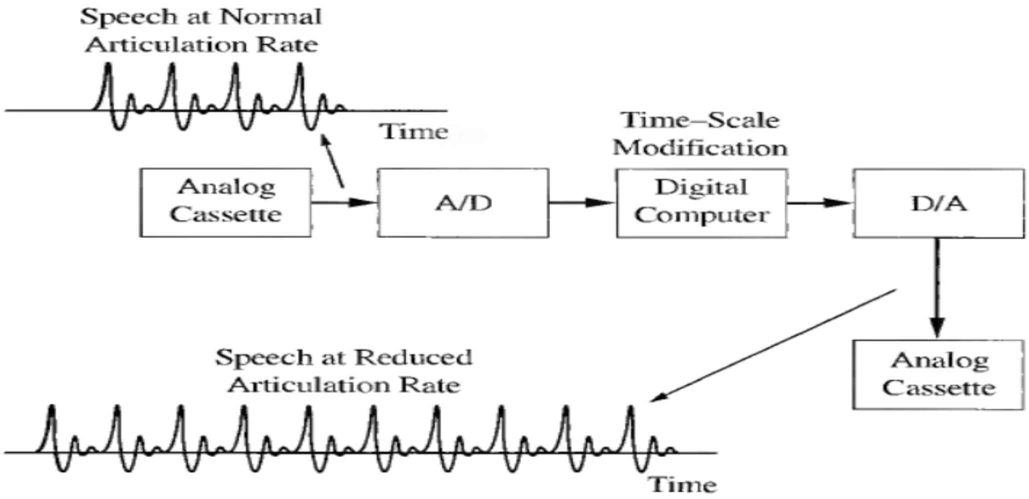


Figure (2.1) Time scale modification as an example of discrete time speech signal processing

2.2.1 Speech Communication Pathway

In the processing of speech signals it’s important to understand the pathway of communication from speaker to listener .At the linguistic level of communication, an idea is first formed in the mind of the speaker the idea is then transformed to word, phrases, and sentences. According to the grammatical rules of the language, at the physiological level of communication the brain creates electric signals activate muscles in in the vocal cords. This vocal tract and vocal cord movement result in pressure changes within the vocal tract, and in particular, at the lips initiating a sound wave that propagates in space.[5]

The sound wave propagates in space as a chain reaction among the air particles resulting in a pressure change at the air canal and thus vibrating the ear drum.

The pressure change at the lips , the sound change at the ear drum of the listener are considered the acoustic level in the speech communication pathway .

The vibration at the ear drum induces electric signals that move along the sensory nerves to the brain, we are back to the physiological level. Finally at the linguistic level of the listener the brain performs speech recognition and understanding.

Figure (2.2) which show a model of vowel production .Air is forced from the lungs by contraction of muscles around the lungs cavity, Air then flows past the vocal cords, which are two masses of flesh, causing periodic vibration of the cords whose rates gives the pitch of the sound. The resulting periodic puffs of air acts as an excitation inputs, or source to the vocal cords and the lips , and acts as a resonator that spectrally shapes the periodic input like the cavity of a musical wind instrument .[5]

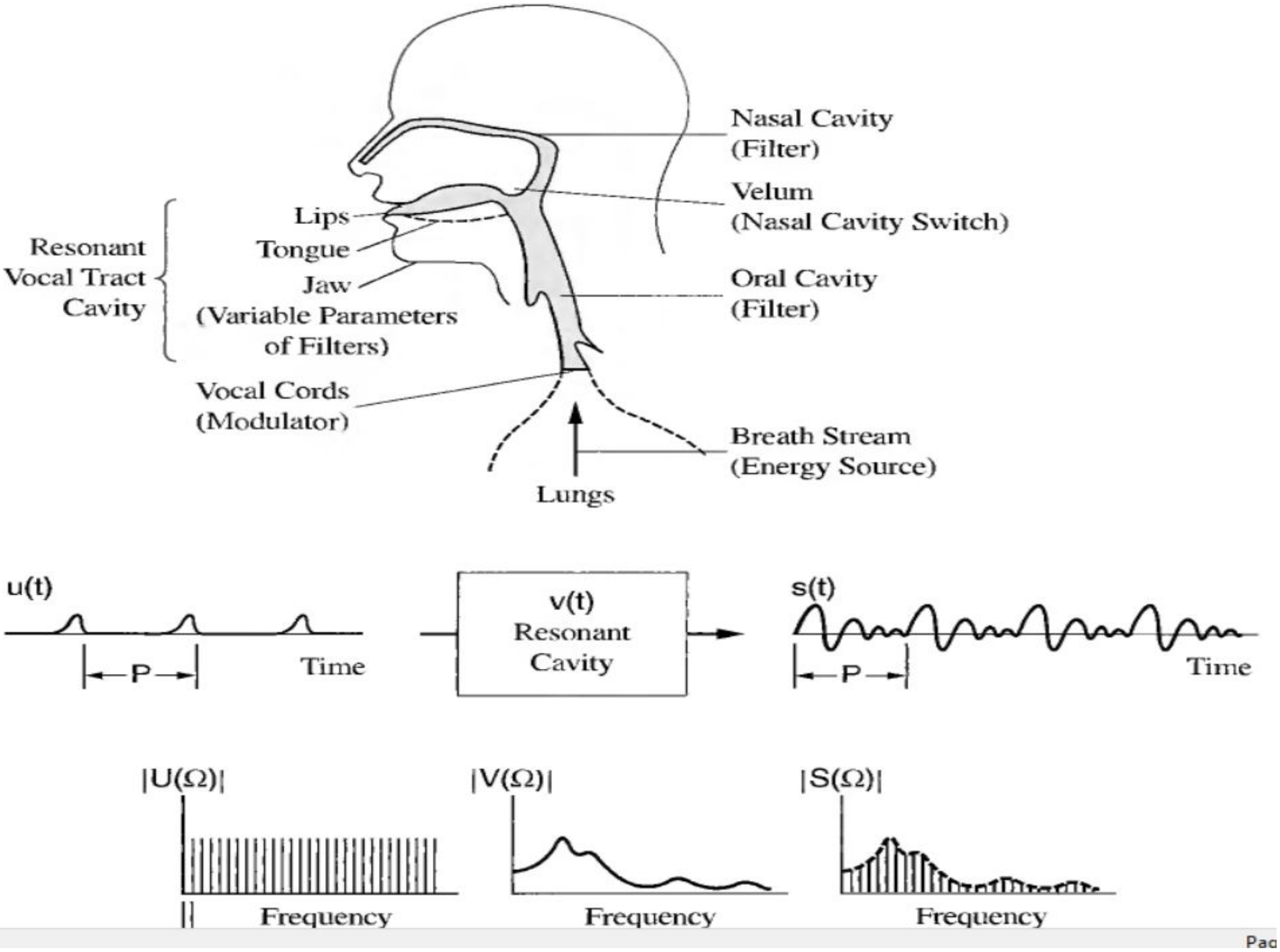


Figure (2.2) simple view of speech production mechanism and model of steady state vowel, the acoustic wave form is modeled as the output with periodic impulse like input, in the frequency domain the vocal tract system function spectrally shapes the harmonic input

2.2.2 Production and Classification of Speech

A simplified view of speech production is given in Figure 2.3, where the speech organs are divided into three main groups: the lungs, larynx and vocal tract.

The lungs act as a power supply and provide airflow to the larynx stage of the speech production mechanism. The larynx modulates airflow from the lungs and provides either periodic puff like or a noisy airflow source to the third organ group, the vocal tract. The vocal tract consist s of oral, nasal and pharynx cavities giving modulated airflow its “Color” by spectrally shaping the source . Sound sources can also be generated by constrictions and boundaries, the variation of air pressure at the lips results in a traveling sound wave that the listener perceives as speech.

The speech sounds consist of three categories:

- Periodic
- Noisy
- impulsive

Although combinations of these sources are often present. Examples of speech

Sounds generated with each of these source categories are seen in the word “shop,” where the “sh,” “o,” and “p” are generated from a noisy, periodic, and impulsive source, respectively. The reader should speak the word “shop” slowly and determine where each sound source is occurring, i.e., at the larynx or at a constriction within the vocal tract. [5]

Such distinguishable speech sounds are determined not only by the source, but by different vocal tract configurations, and how these shapes combine with periodic, noisy, and impulsive sources. These more refined speech sound classes are referred to as phonemes, the study of which is called phonemics. A specific phoneme class provides a certain meaning in a word, but within a phoneme class, , there exist many sound variations that provide the same meaning. The study of these sound variations is called phonetics. Phoneme, the basic building blocks of a language, is concatenated, more or less, as discrete elements into words, according to certain phonemic and grammatical rules.

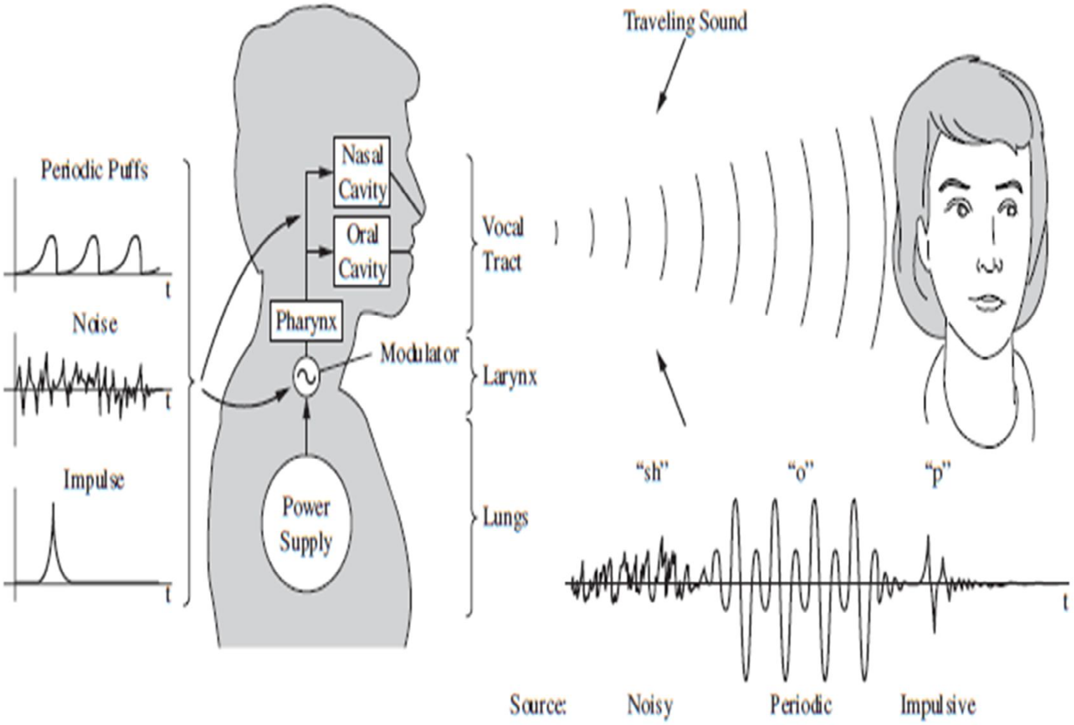


Figure (2.3) simple view of speech production .The sound wave are identified as noise , periodic or impulsive and can occur in the larynx or vocal tract

2.2.3 Anatomy and Physiology of Speech Production

We now look in detail at this anatomy, as well as at the associated physiology and its importance in speech production. [5]

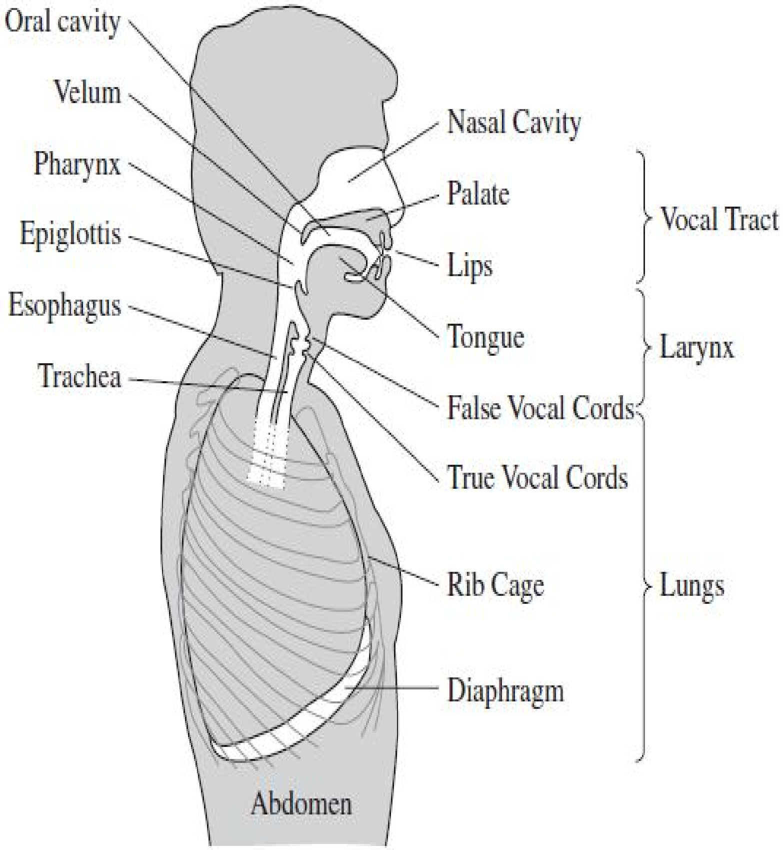


Figure (2.4) Cross-sectional view of the anatomy of speech production

2.2.3.1 Lungs

One purpose of the lungs is the inhalation (inspiration) and exhalation (expiration) of air. When we inhale, we enlarge the chest cavity by expanding the rib cage surrounding the lungs and by lowering the diaphragm that sits at the bottom of the lungs and separates the lungs from the abdomen; this action lowers the air pressure in the lungs, thus causing air to rush in through the vocal tract and down the trachea

Into the lungs the trachea, sometimes referred to as the “windpipe,” is about a 12-cm-long and 1.5–2-cm-diameter pipe which goes from the lungs to the epiglottis. The epiglottis is a small mass, or “switch,” which, during swallowing and eating, deflects food away from entering the trachea. When we eat, the epiglottis falls, allowing food to pass through a tube called the esophagus and into the stomach. When we exhale, we reduce the volume of the chest cavity by contracting the muscles in the rib cage, thus increasing the lung air pressure. This increase in pressure then causes air to flow through the trachea into the larynx. In breathing, we rhythmically inhale to take in oxygen, and exhale to release carbon dioxide.

During speaking, on the other hand, we take in short spurts of air and release them steadily by controlling the muscles around the rib cage. We override our rhythmic breathing by making the duration of exhaling roughly equal to the length of a sentence or phrase .

2.2.3.2 Larynx

The larynx is a complicated system of cartilages, muscles, and ligaments whose primary Purpose, in the context of speech production, is to control the vocal cords or vocal folds [5]. The vocal folds are two masses of flesh, ligament, and muscle, which stretch between the front and back of the larynx, as illustrated in (Figure 2.5). The folds are about 15 mm long in men and 13 mm long in women. The glottis is the slit-like orifice between the two folds. The folds are fixed at the front of the larynx where they are attached to the stationary thyroid cartilage. The thyroid cartilage is located at the front (or Adam's apple) and sides of the larynx. The folds are free to move at the back and sides of the larynx; they are attached to the two arytenoid cartilages that move in a sliding motion at the back of the larynx along with the cricoid cartilage. folds (Figure 2.6).

There are three primary states of the vocal folds: breathing, voiced, and unvoiced. In the breathing state, the arytenoid cartilages are held outward (Figure 2.5 b), maintaining a wide glottis, and the muscles within the vocal folds are relaxed. In this state, the air from the lungs flows freely through the glottis with negligible hindrance by the vocal folds. In speech production, on the other hand, an obstruction of airflow is provided by the folds. In the voicing state, as, for example, during a vowel, the arytenoid cartilages move toward one another (Figure 2.5a). The vocal folds tense up and are brought close together. This partial closing of the glottis and increased fold tension cause self-sustained oscillations of the folds. [5]

Sounds coming out of it, such as hamza and alhaa (الهاء , الهمزة) [1] .

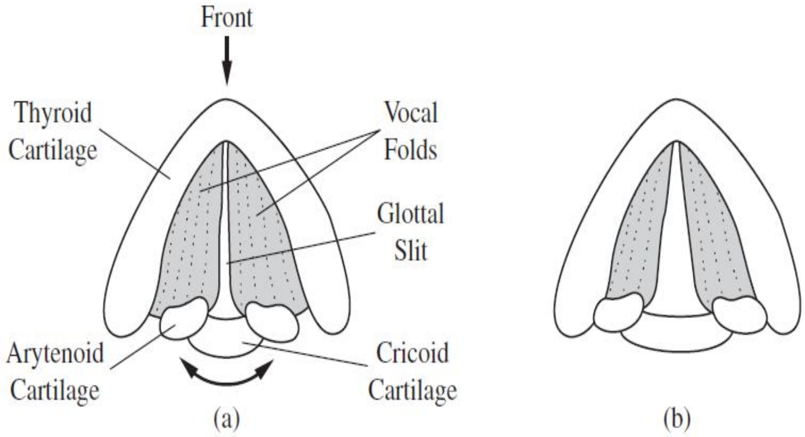


Figure (2.5) Sketches of the human Larynx (a) voicing, (b) breathing

According to our description of the air flow velocity in the glottis, if we were to measure the airflow velocity at the glottis as a function of time, we would obtain a waveform approximately similar to that illustrated in (Figure 2.6) that roughly follows the time-varying area of the glottis.

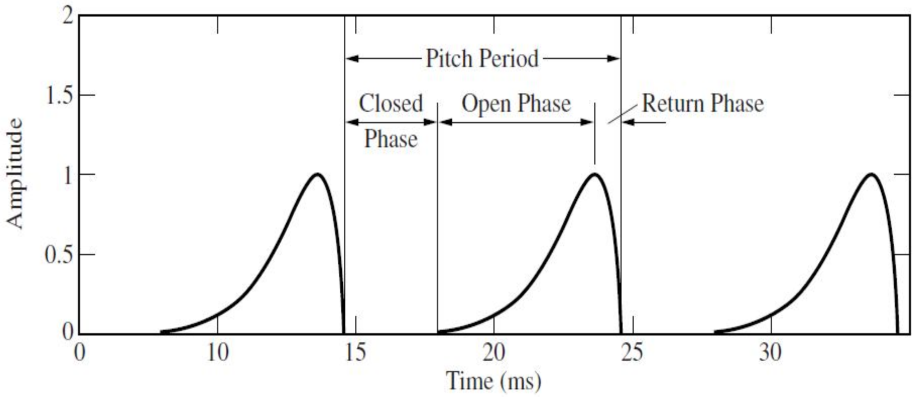


Figure (2.6) Illustration of periodic glottal airflow velocity

The time duration of one glottal cycle is referred to as the pitch period, also referred to as the fundamental frequency .The number of pitch periods changes with numerous factors such as stress and speaking rate. The rate at which the vocal folds oscillate through a closed, open, and return cycle is influenced by many factors. The pitch range is about 60 Hz to 400 Hz. typically; males have lower pitch than females because their vocal folds are longer and more massive. These include vocal fold muscle tension (as the tension increases, so does the pitch), the vocal fold mass (as the mass increases, the pitch decreases because the folds are more sluggish), and the air pressure behind the glottis in the lungs and trachea, which might increase in a stressed sound or in a more excited state of speaking (as the pressure below the glottis increases, so does the pitch). A simple mathematical model of the glottal flow is given by the convolution of a periodic impulse train with the glottal flow over one cycle. [5]

Consider a glottal flow waveform model of the form

$$u[n] = g[n] * p[n] \text{……………………………………….}(2.1)$$

Where g[n] is the glottal flow waveform over a single cycle and

$$p[n] = \sum_{k=-\infty}^{\infty} \delta[n - kP] \quad \text{………………………………………} (2.2)$$

Is an impulse train with spacing P. Because the waveform is infinitely long, we extract a segment by multiplying x[n] by a short sequence called an analysis window or simply a window. The window, denoted by w[n, τ], is centered at time τ , as illustrated in Figure 2.7, and the resulting waveform segment is written as . [5]

$$u[n, \tau] = w[n, \tau](g[n] * p[n]). \text{……………………………………….. } (2.3)$$

Using the Multiplication and Convolution Theorems, we obtain in the frequency domain

$$U(\omega,\tau) = \frac{1}{P}W(\omega,\tau) \circledast \left[\sum_{k=-\infty}^{\infty} G(\omega)\delta(\omega-\omega_k)\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4)$$

Where W(ω, τ) is the Fourier transform of w[n, τ], where G(ω) is the Fourier transform of g[n], where $\omega_k = \frac{2\pi}{P}k$, and where $\frac{2\pi}{P}$ is the fundamental frequency or pitch. As illustrated in Figure 2.7, the Fourier transform of the window sequence is characterized by a narrow main lobe centered at ω = 0 with lower surrounding side lobes. The window is typically selected to trade off the width of the main lobe and attenuation of the side lobes. Figure 2.8 illustrates how the Fourier transform magnitude of the waveform segment changes with pitch and with characteristics of the glottal flow. As the pitch period decreases, the spacing between the frequencies , which are $\omega_k = \frac{2\pi}{P}k$, referred to as the harmonics of the glottal waveform, increases, as can be seen by comparing Figures 2.7 a and 2.7 b. The first harmonic is also the fundamental frequency. [5]

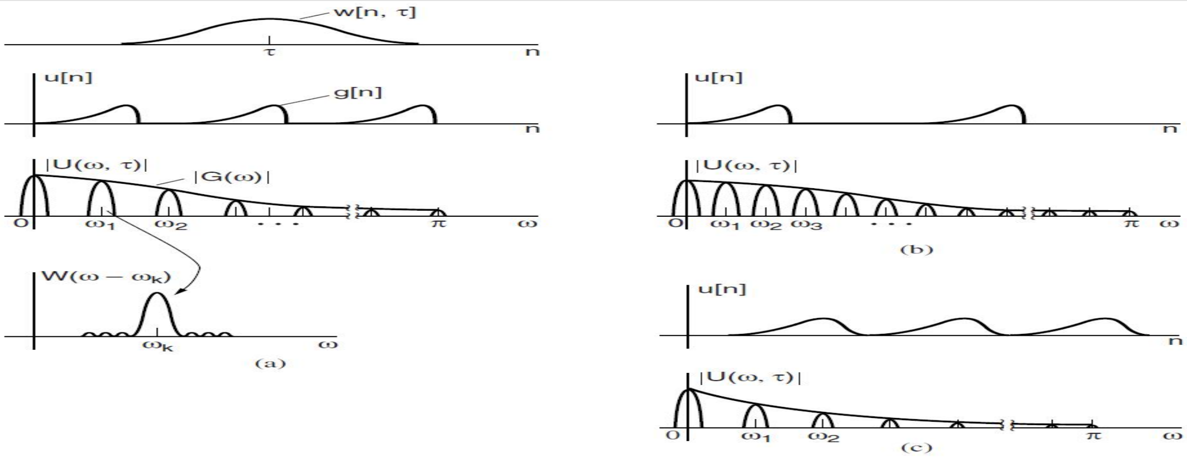


Figure (2.7) Illustration of periodic glottal flow: (a) typical glottal flow and its spectrum; (b) same as (a) with lower pitch; and (c) same as (a) with “softer” or more “relaxed” glottal flow.

2.2.3.2 Vocal Tract

The vocal tract is comprised of the oral cavity from the larynx to the lips and the nasal passage that is coupled to the oral tract by way of the velum. The oral tract takes on many different lengths and cross-sections by moving the tongue, teeth, lips, and jaw and has an average length of 17 cm in a typical adult male and shorter for females, and a spatially-varying cross section of up to 20 cm^2 . If we were to listen to the pressure wave at the output of the vocal folds during voicing, we would hear simply a time-varying buzz-like sound which is not very interesting.. [5].

Spectral Shaping

Under certain conditions, the relation between a glottal airflow velocity input and vocal tract airflow velocity output can be approximated by a linear filter with resonances, much like resonances of organ pipes and wind instruments. The resonance frequencies of the vocal tract are, in a speech science context, called formant frequencies or simply formants.

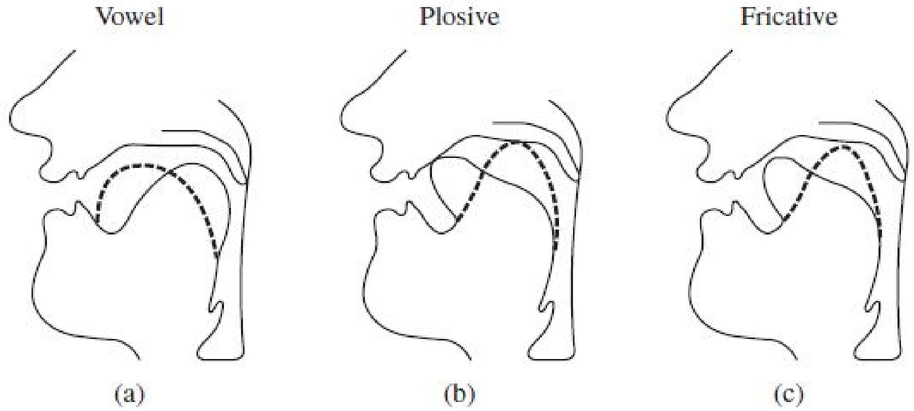


Figure (2.8) Illustration of changing vocal tract shapes for (a) vowels (having a periodic source), (b) plosives (having an impulsive source), and (c) fricatives (having a noise source).

The formants of the vocal tract are numbered from the low to high formants according to their location; the first formant is denoted by F1 , the second formant by F2 , and so on up to the highest formant. Generally, the frequencies of the formants decrease as the vocal tract length increases; as a consequence, a male speaker tends to have lower formants than a female, and a female has lower formants than a child. Under a vocal tract linearity and time-invariance assumption, and when the sound source occurs at the glottis, the speech waveform, i.e., the airflow velocity at the vocal tract output can be expressed as the convolution of the glottal flow input and vocal tract impulse response, as illustrated in the following formula:

Consider a periodic glottal flow source of the form [5].

$$u[n] = g[n] * p[n] \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.5)}$$

Where g[n] is the airflow over one glottal cycle and p[n] is the unit sample train with spacing P. When the sequence u[n] is passed through a linear time-invariant vocal tract with impulse response h[n], the vocal tract output is given by

$$x[n] = h[n] * (g[n] * p[n]). \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.6)}$$

A window centered at time τ, w [n, τ], is applied to the vocal tract output to obtain the speech segment

$$x[n, \tau] = w[n, \tau]\{h[n] * (g[n] * p[n])\}. \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.7)}$$

Using the Multiplication and Convolution Theorems, we obtain in the frequency domain the Fourier transform of the speech segment

$$\begin{aligned} X(\omega, \tau) &= \frac{1}{P} W(\omega, \tau) \circledast \left[H(\omega) G(\omega) \sum_{k=-\infty}^{\infty} \delta(\omega - \omega_k) \right] \qquad \ldots\ldots\ldots \text{(2.8)} \\ &= \frac{1}{P} \sum^{\infty} H(\omega_k) G(\omega_k) W(\omega - \omega_k, \tau) \end{aligned}$$

Where W (ω, τ) is the Fourier transform of w [n, τ], where , and where 2π P is the fundamental frequency or pitch.

$(1 - c_k z^{-1})$ and $(1 - c_k^* z^{-1})$;

Figure 2.9 illustrates that the spectral shaping of the window transforms at the harmonics ω1 , ω2 , . . . ωN is determined by the spectral envelope |H(ω)G(ω)| consisting of a glottal and vocal tract contribution, where only the glottal contribution occurred. The peaks in the spectral envelope correspond to vocal-tract formant frequencies, F1, F2, . . . FM. The general upward or downward slope of the spectral envelope, sometimes called the spectral tilt, is influenced by the nature of the glottal flow waveform over a cycle, e.g., a gradual or abrupt closing, and by the manner in which formant tails add. We also see in Figure 2.9 that the formant locations are not always clear from the short-time Fourier transform magnitude |X(ω, τ)|.[5] .

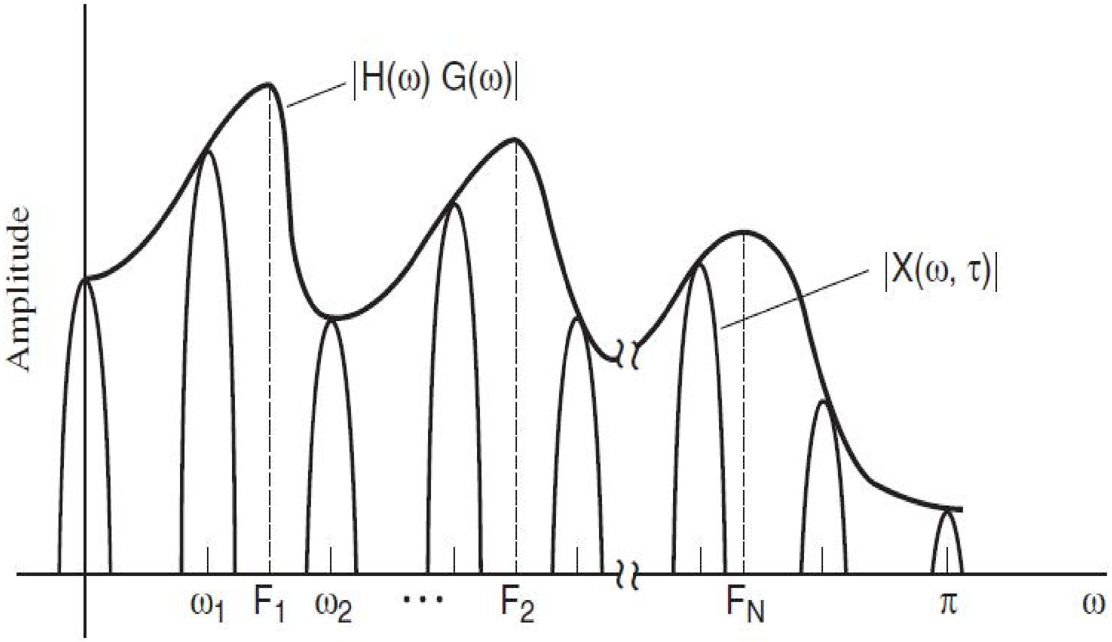


Figure (2.9) of Illustration of relation of glottal source harmonics ω1 , ω2 , . . . ωN , vocal Tract formants F1, F2 , . . . FM, and the spectral envelope |H(ω)G(ω)|.

This example illustrates the important difference between a formant, or resonance, frequency and a harmonic frequency. A formant corresponds to the vocal tract poles, while the harmonics arise from the periodicity of the glottal source.

2.3 Categorization of Sound by Source

There are various ways to categorize speech sounds. For example, we can categorize speech sounds based on different sources to the vocal tract; we have seen that different sources are due to the vocal fold state, but are also formed at various constrictions in the oral tract. Speech sounds generated with a periodic glottal source are termed voiced; likewise, sounds not so generated are called unvoiced. There are a variety of unvoiced sounds, including those created with a noise source at an oral tract constriction. Because the noise of such sounds comes from the friction of the moving air against the constriction, these sounds are sometimes referred to as fricatives. An example of frication is in the sound “th” in the word “thin” where turbulence is generated between the tongue and the upper teeth. The reader should hold the “th” sound and feel the turbulence.

A second unvoiced sound class is plosives created with an impulsive source within the oral tract. An example of a plosive is the “t” in the word “top.” The location of the closed or partial constriction corresponds to different plosive or fricative sounds, respectively. We noted earlier that a barrier can also be made at the vocal folds by partially closing the vocal folds, but without oscillation, as in the sound “h” in “he.” These are whispered unvoiced speech sounds. These voiced and unvoiced sound categories, however, do not relate exclusively to the source state because a combination of these states can also be made whereby vocal fold vibration occurs simultaneously with impulsive or noisy sources. For example, with “z” in the word “zebra,” the vocal folds are vibrating and, at the same time, noise is created at a vocal tract constriction behind the teeth against the palate. Such sounds are referred to as voiced fricatives in contrast to unvoiced fricatives where the vocal folds do not vibrate simultaneously with frication.

Voice cords have different conditions and the most important of these conditions:-[5]

2.3.1 Placing cords in breathing state

cords open remarkably allowing the passage of air through them without any objection and offset this case the so-called whisper and external voices are called Voice Less Sounds like Alta, Althae, ha and kha .. etc.(التاء ، الثاء ، الحاء
والخاء ...الخ)

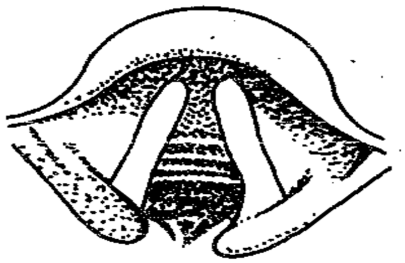


Figure (2.10): cords position in case of breathing

2.3.2 Placing cords in state of issuing a musical tone

cords combine in whole or in part, so that the air rushing through the opens and closing fast and erratic and result in the so-called frequency vocal cords and this oscillation occur musical tone is different in terms of intensity and strength and know this tune sounds voiced Sounds like Alaba, algym, aldal and alzal … etc. Arab and movements are all long (Alalef and alyaa and Alaba) and short (alfatha and alKasra and aldama [1]
الباء ، الجيم ، الدال والذال ...الخ والحركات العربية جميعها الطويلة (الألف والياء والباء) و
القصيرة (الفتحة والكسرة والضمة)

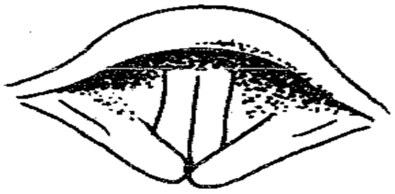


figure (2.11):cords position in case of releasing a musical tone

2.3.3 Placing cords in whispering state

Cords will have a similar state of roughness but it differs that but will be rough and hard and cannot vibrate

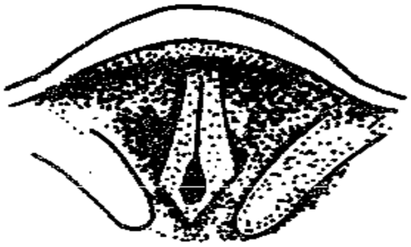


Figure (2.12): cords position in case of whispering

2.3.4 Placing Cord at Composing Hamza katta

It is utterance decline where the cord will be closed completely for a while no air will pass from or to lungs and the explosion voice will happen due to air and this voice is hamzat kattaa[1]

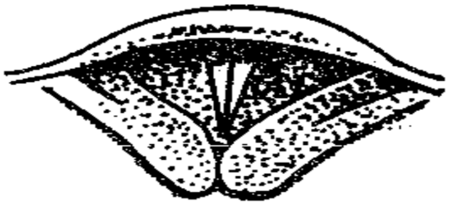


Figure (2.13): cords position in case of creating hamzat kata

2.3 Spectrographic Analysis of Speech

We have seen that a speech waveform consists of a sequence of different events. This time variation corresponds to highly fluctuating spectral characteristics over time. There are two kinds of spectrograms: narrowband, which gives good spectral resolution, e.g., a good view of the frequency content of sine waves with closely spaced frequencies, and wideband, which gives good temporal resolution, e.g., a good view of the temporal content of impulses closely spaced in time, the difference between the narrowband and wideband spectrogram is the length of the window w[n, τ]..For voiced speech [1]

2.4.1 Narrowband Spectrogram

The difference between the narrowband and wideband spectrogram is the length of the window w[n, τ]. For the narrowband spectrogram, we use a “long” window with a duration of typically at least two pitch periods., consists of a set of narrow “harmonic lines,” whose width is determined by the Fourier transform of the window, which are shaped by the magnitude of the product of the glottal flow Fourier transform and vocal tract transfer function. The narrowband spectrogram gives good frequency resolution because the harmonic lines are “resolved”; these harmonic lines are seen as horizontal striations in the time-frequency plane of the spectrogram

2.4.2 Wideband Spectrogram

For the wideband spectrogram, we choose a “short” window with a duration of less than a single pitch period; since the window length is less than a pitch period, as the window slides in time it “sees” essentially pieces of the periodically occurring sequence For the steady-state voiced sound, we can therefore express the wideband spectrogram (very) roughly , also gives vertical striations in time every pitch period, rather than the harmonic horizontal striations as in the narrowband spectrogram.,. With plosive sounds, the wideband spectrogram is often preferred because it gives better temporal resolution of the sound’s components, especially when the plosive is closely surrounded by vowels. [5].

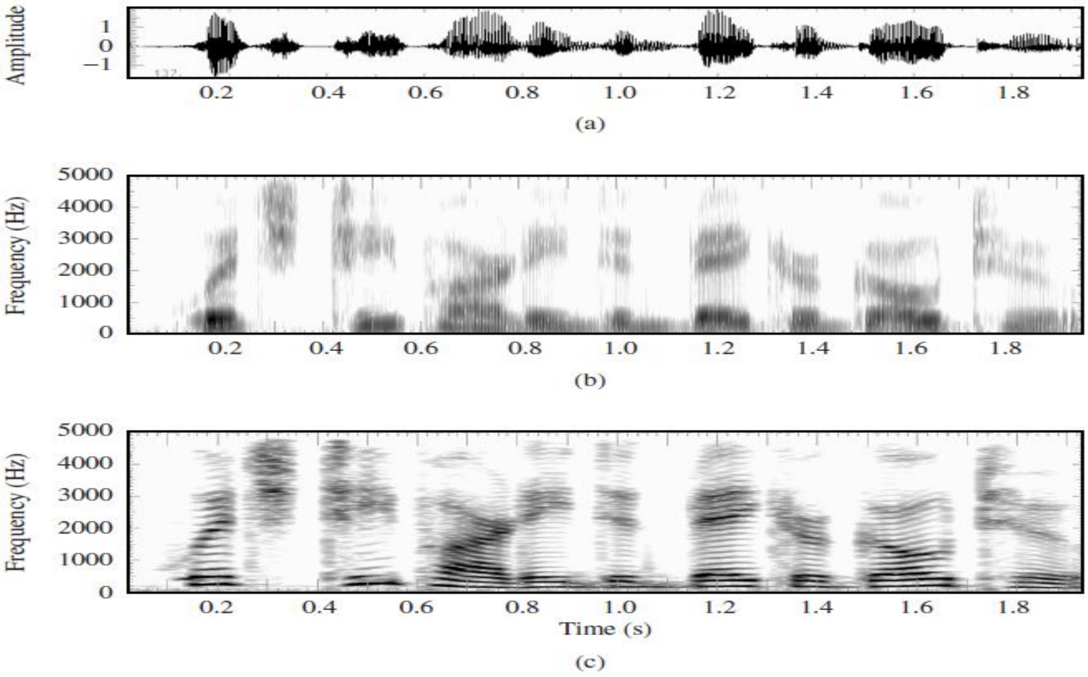


Figure (2.14) Comparison of measured spectrograms

(a) Speech waveform; (b) wideband spectrogram; (c) narrowband spectrogram

2.4 Categorization of Speech Sounds

Sound source can be created with either the vocal folds or with a constriction in the vocal tract, and, based on the various sound sources; we proposed a general categorization of speech sounds. .

Speech sounds are studied and classified from the following perspectives:

1) The nature of the source: periodic, noisy, or impulsive, and combinations of the three.

2) The shape of the vocal tract.
3) The time-domain waveform, which gives the pressure change with time at the lips output.

4) The time-varying spectral characteristics revealed through the spectrogram.

2.5 Elements of Language

A fundamental distinctive unit of a language is the phoneme; the phoneme is distinctive in the sense that it is a speech sound class that differentiates words of a language For example, the words “cat,” “bat,” and “hat” consist of three speech sounds, the first of which gives each word its distinctive meaning, being from different phoneme classes. Different languages contain different phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables, concatenated to form phrases and sentences. Linguistics is the study of the arrangement of speech sounds.

One broad phoneme classification for English is in terms of vowels, consonants, diphthongs, affricates, and semi-vowels. Figure 2.15 Phonemes arise from a combination of vocal fold and vocal tract articulatory features.

The variants of sounds or phones that convey the same phoneme are called the allophones of the phoneme Consider, for example, the words “butter,” “but,” and “to,” where the /t/ in each word is somewhat different with respect to articulation, being influenced by its position within the word. [5].

2.6.1 Vowels

The largest phoneme group is that of vowels. Vowels contain three subgroups defined by the tongue hump being along the front, central, or back part of the palate.

It is the second major section of the votes of the language, in the language movements differ from environment to another is because of this difference to the habits pronunciation local dialect, put lips is the general standard, which is classified Ken types of movements or patterns, movements term that was called in the old on the fatha and damma and Kasra or what is known as short vowels, where long vowels are الالف والواو والياء sleepless movements a: (ـــُـــِـــَـ) [5]

2.6.1.1 Long Vowels

Almad letters are known by the stretching of the letter exits so the air does not break from its extend and the letters that has wide exits are والواو والياء الالف). [5]

2.7.1.2. Short Vowels

They are alfatha which is half of the الالف and alkasra is half of الياء and aldama which is half of الواو and lips place is one of the features that distinguish this vowels from others where alfatha opens lips and alksra will break and widen the lips and aldamma will close lips [5]

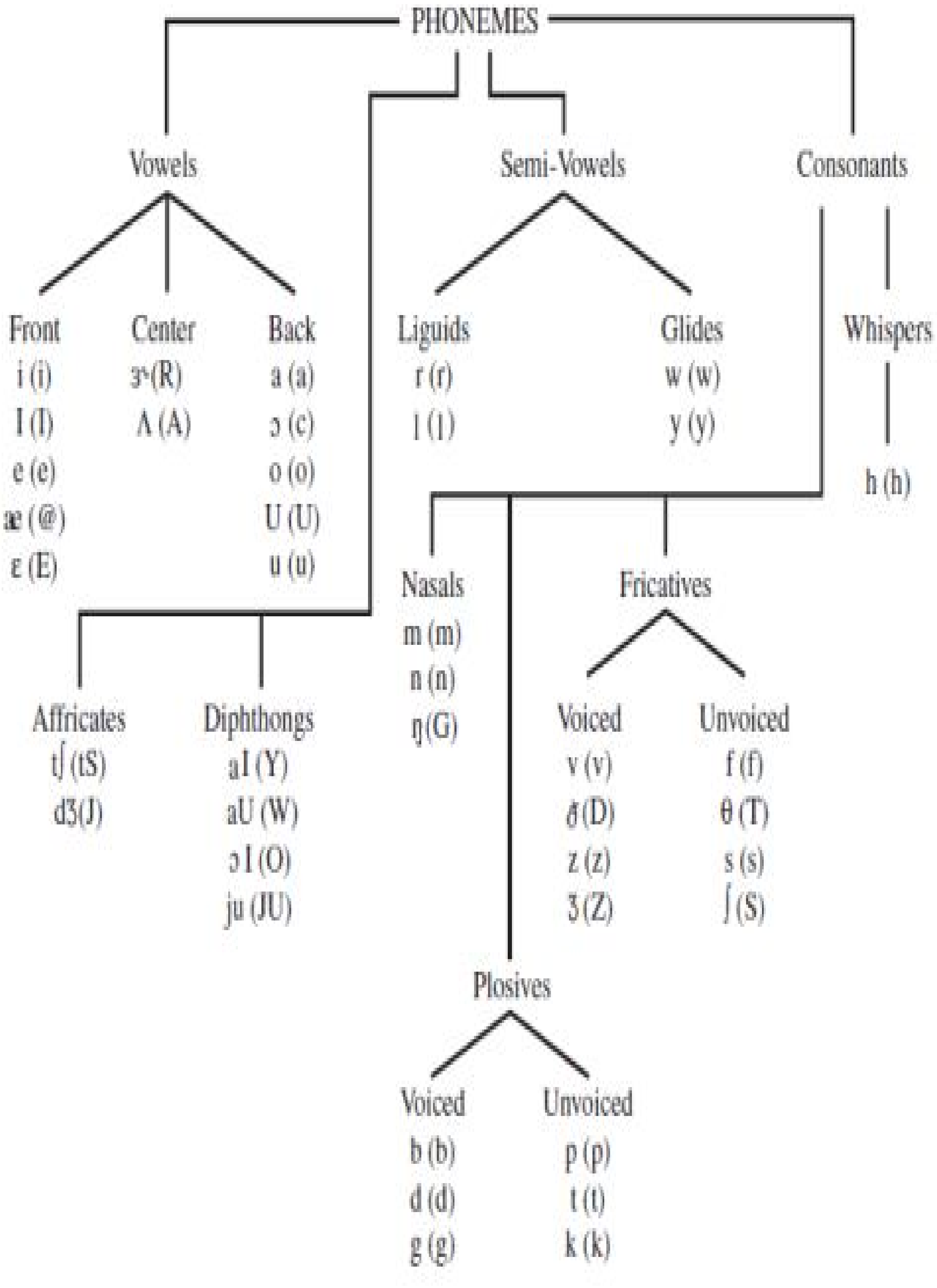


Figure (2.15) Phonemes in American English

2.6.2 Consonants

Consonant identification depends on a number of factors including the formants of the consonant, formant transitions into the formants of the following vowel, the voicing (or unvoicing) of the vocal folds during or near the consonant production.

2.6.2.1 Nasals

The second large phoneme grouping is that of consonants. The consonants contain a number of subgroups: nasals, fricatives, plosives, whispers, and affricates. We begin with the nasals since they are closest to the vowels. In Arabic language ن, م

Source: As with vowels, the source is quasi-periodic airflow puffs from the vibrating vocal folds.

System: The velum is lowered and the air flows mainly through the nasal cavity, the oral tract being constricted

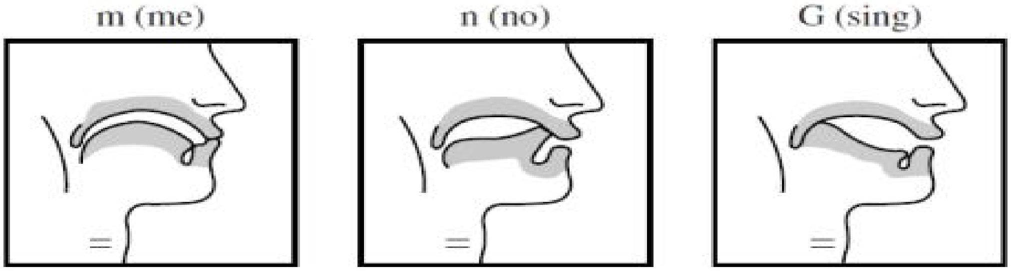


Figure (2.16) Vocal tract configurations for nasal consonants. Oral tract constrictions occur at the lips for /m/, with the tongue tip to the gum ridge for /n/, and with the tongue body against the palate near the velum for /ng/. Horizontal lines denote voicing.

Spectrogram: For the /m/ in Figure 2.16b, there is a low F1 at about 250 Hz with little energy above this frequency. A similar pattern is seen for the /n/ in Figure 2.16a.

2.6.2.2 Fricatives

Fricative consonants are specified in two classes: voiced and unvoiced fricatives.

Source: In unvoiced fricatives, the vocal folds are relaxed and not vibrating.

System: The location of the constriction by the tongue at the back, center, or front of the oral tract, as well as at the teeth or lips, influences which fricative sound is produced. e.g. in Arabic الظاء ، الزاى ، السين ، الصاد ، الشين ، الخاء ، الغين ، الحاء ، العين ، الهاء والفاء الثاء ، الذال ، [5] .

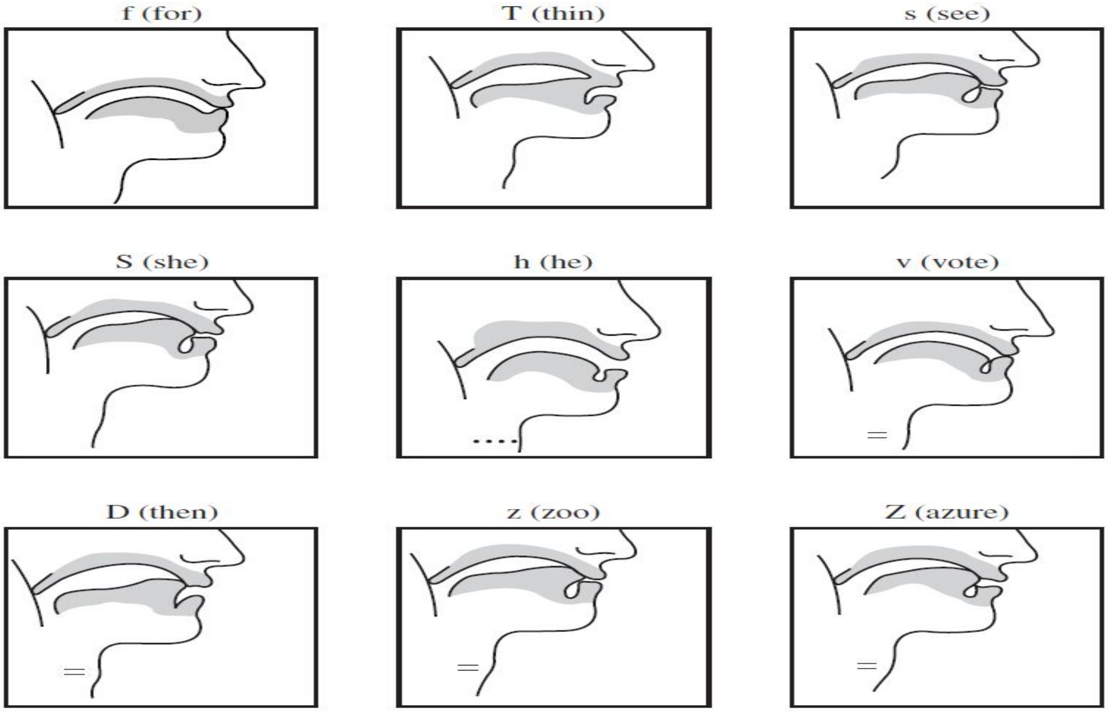


Figure (2.17) Vocal tract configurations for pairs of voiced and unvoiced fricatives. Horizontal lines denote voicing and dots denote aspiration.

2.6.2.3 Whisper

The whisper is a consonant similar in formation to the unvoiced fricative; we place the whisper in its own consonantal class. We saw earlier that with a whisper the glottis is open and there is no vocal fold vibration. An example is /h/. [5].

2.6.2.4 Plosives

As with fricatives, plosives are both unvoiced and voiced.

Source and System: With unvoiced plosives, a “burst” is generated at the release of the buildup of pressure behind a total constriction in the oral tract. E.g. in Arabic الطاء ، التاء ، الضاد ، الدال ، الكاف ، القاف ، همزة

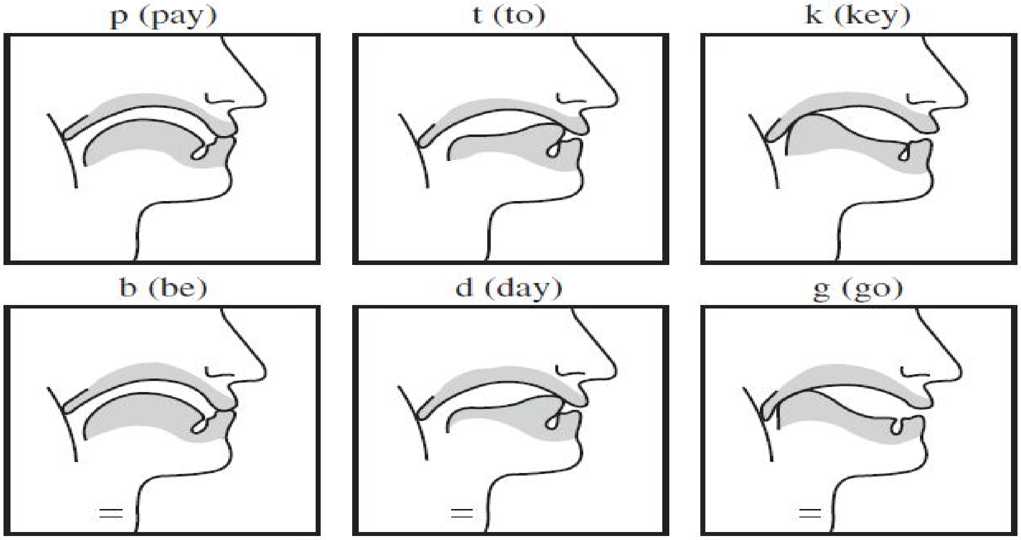


Figure (2.18) Vocal tract configurations for unvoiced and voiced plosive pairs. Horizontal lines denote voicing.

2.6.3 Semi-Vowels:

This class is also vowel-like in nature with vibrating folds. There are two categories of semi-vowels: glides (/w/ as in "we" and /y/ as in "you") and liquids (/r/ as in "read" and /l/ as in "let"). [11]. In Arabic is و, ي [1]

2.7 Transitional Speech Sounds

Speech sounds are "nonstationary" and in some cases the rapid transition across two articulatory states defines the sound. nonstationarity is further imparted through a phenomenon known as coarticulation . Such nonstationarity poses interesting challenges to speech signal processing algorithms that typically assume (and often require) stationarity over intervals of 10–20 ms. [5]

2.8 Speech Perception

The acoustic properties of speech sounds that are essential for phoneme discrimination by the auditory system .

2.8.1 Acoustic Cues

We are interested in acoustic components of a speech sound used by the listener to correctly perceive the underlying phoneme, Formant frequencies have been determined to be a primary factor in identifying a vowel. The first two formants (F1 and F2) to vowel identification. Higher formants also have a role in vowel identity. We would expect the listener to normalize formant location in doing phoneme recognition.

Another factor in vowel perception is nasalization, which is cued primarily by the bandwidth increase of the first formant (F1) and the introduction of zeros[5].

2.9 Feature Extraction

Feature extraction stage is the most important one in the entire process, since it is responsible for extracting relevant information from the speech frames, as feature parameters or vectors. Common parameters used in speech recognition are Linear Predictive Coding (LPC) coefficients, and Mel Frequency Cepstral Coefficients (MFCC). These parameters have been widely used in recognition system [7]

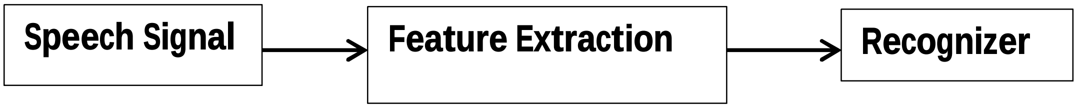


Fig. 2.19: Frame Work for Speech Recognition

2.9.1 Spectral Shaping

In this process the audio signal is converted from the wave sound to a digital signal and then converts the signal from analog to digital and determining sampling frequency (sample frequency) and the value of clarity (resolution).

This digital signal Subject to the factorization process to focus on the most important frequency components of the signal and exclude the high frequencies that are not included in the audio human itarian field . audio includes a range of features such as:

- Sample frequency is 44100HZ
- Bit rate in the sample 16 bits per sample
- The number of voice channels stereo [7]

2.9.2 Spectral Analysis

Any audio signal has a special spectral analysis distinguishes it from the rest of the signals. Output of this process is the numerical vectors (Numerical vectors) subject to other processors, so that it can distinguish these audio signal, in this stage the speech division into group of sections (Frames) each one represents a 20 ml seconds, which corresponds to 256 part of the sample, and then the sound cutting to turn it into parts by windows Technology (Windowing) from the beginning to the end of each section And these two phases is a sound processing to extract the distinctive characteristics, using several techniques used, including Mel frequency cepstral coefficients (MFCC) and linear predictive coding(LPC) [7]

2.9.3 Mel Frequency Cepstral Coefficients Processor(MFCC):

MFCC's is a type of algorithm i.e. basically used to define relationship between human ear's critical bandwidths with frequency. This method is basically used for analyzing and extraction of pitch vectors. [7]

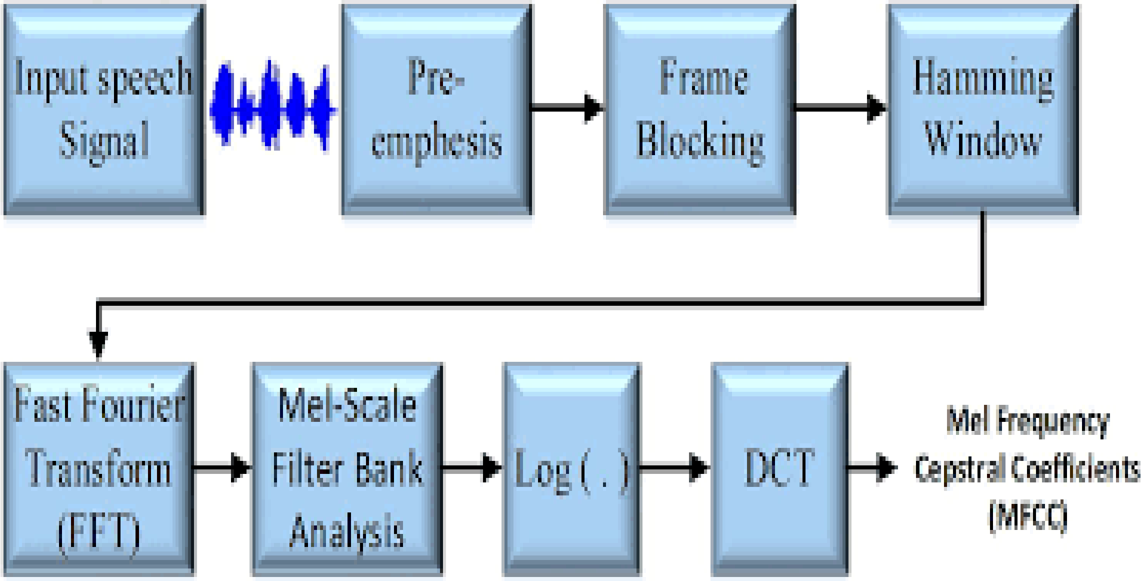

Input speech Signal
Pre-emphesis
Frame Blocking
Hamming Window
Fast Fourier Transform (FFT)
Mel-Scale Filter Bank Analysis
Log (.)
DCT
Mel Frequency Cepstral Coefficients (MFCC)

Fig. 2.20: MFCC Block Diagram

As shown in Figure 2.21 MFCC consists of seven computational steps. Each step has its function and mathematical. Figure 2.22 represents Extraction of MFCC Feature for a Frame. [7]

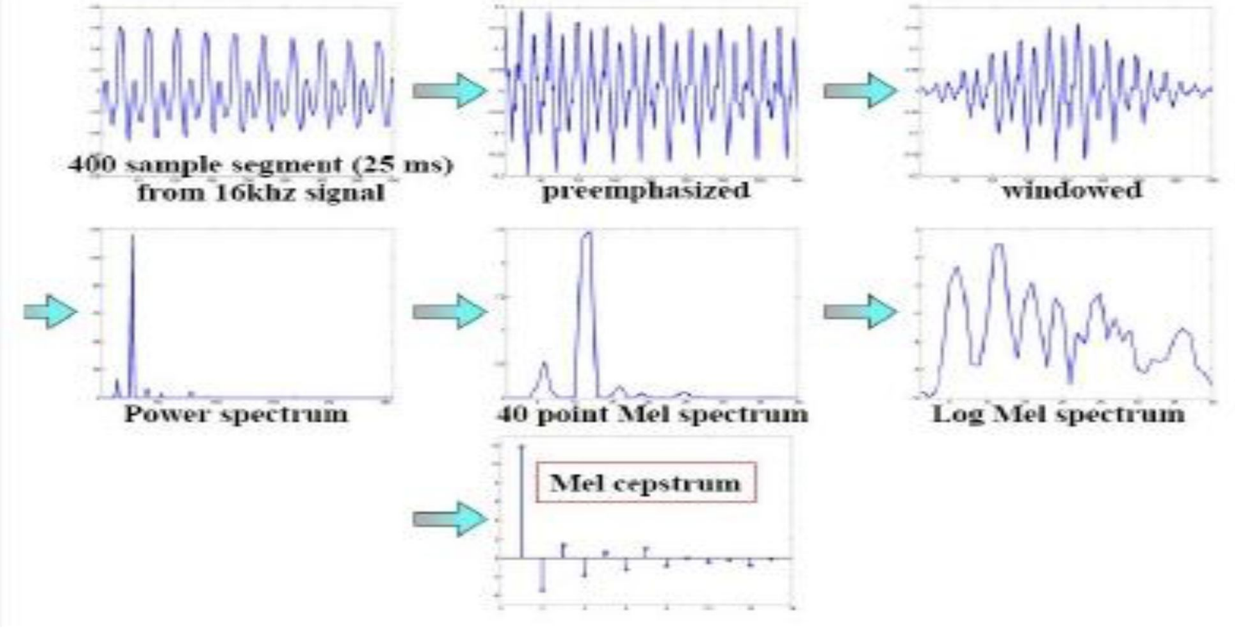


Fig 2.21. Extraction of MFCC Feature for a Frame

(a) Pre-emphasis

The speech signal x(n) is sent to a high-pass filter

Formulae:

y(n) = x(n) - a*x(n-1)…………………………………………………..(2.9)

Where s_2(n) is the output signal and the value of a is typically between 0.9 and 0.99.

. (b) Frame Blocking

The input speech signal is segmented into frames of 20~30 ms. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. [7]

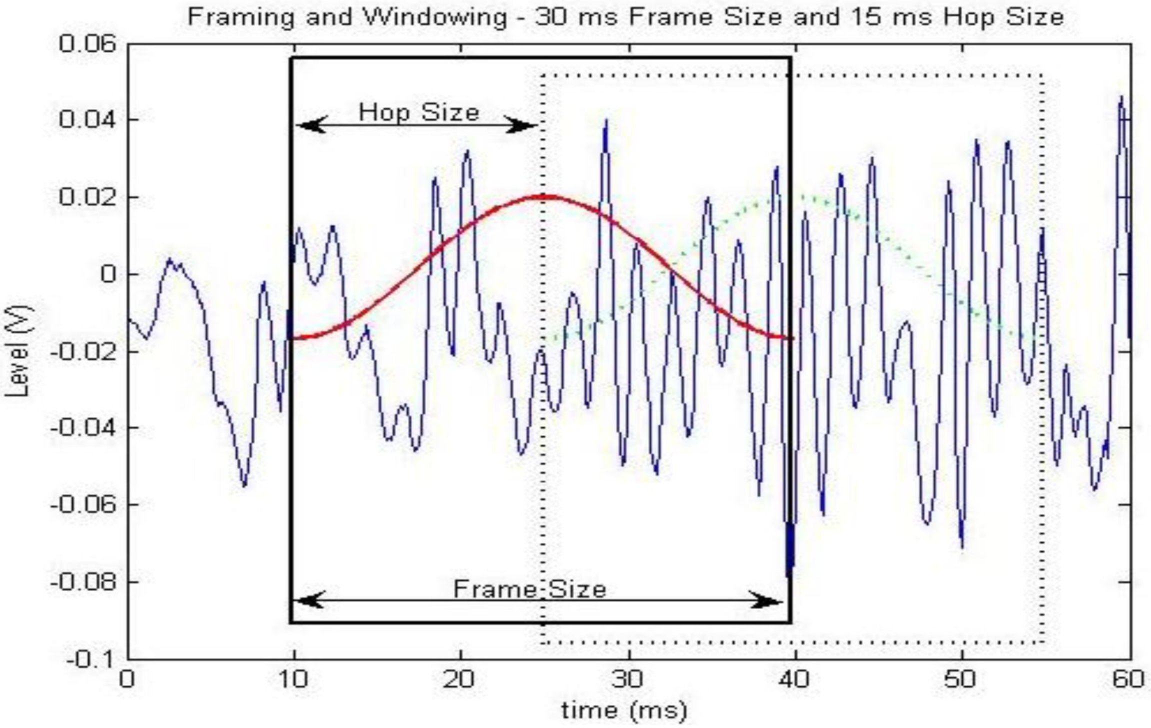


Fig 2.22:. Framing and overlapping

. (c) Hamming windowing

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by x(n), n = 0,…N-1, then the signal after Hamming windowing is

Formulae:

x(n)*w(n), where w(n) is the Hamming window defined by:

$w(n, a)=(1- a)- a \cos(2pn/(N-1)),\quad 0\leqq n\leqq N-1$…………………….. (2.10) [15].

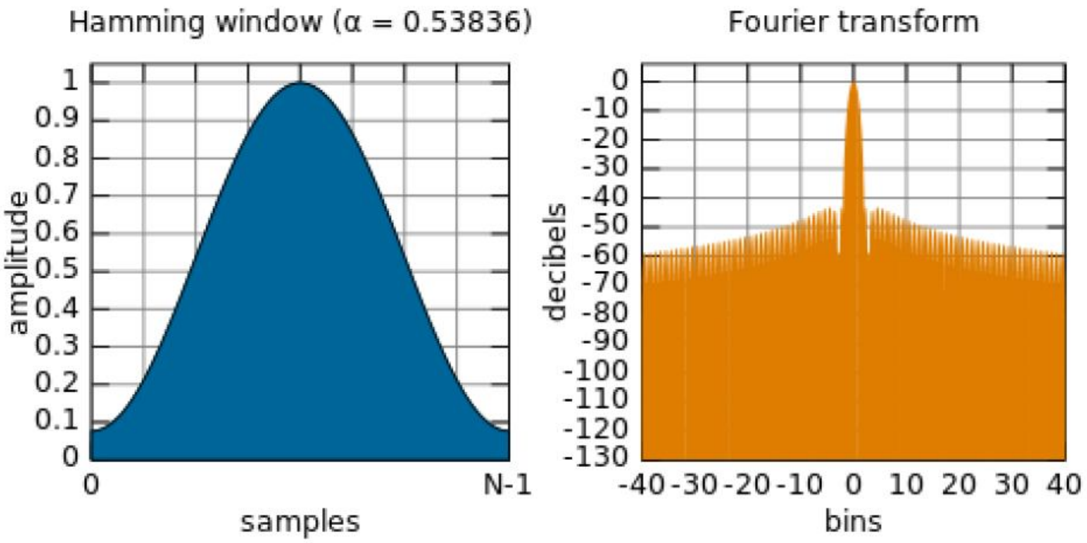


Fig 2.23: Hamming window and its frequency spectrum

(d) Fast Fourier Transform or FFT

Spectral analysis of different pitches in speech signals corresponds to different energy distribution on frequency scale.

Therefore FFT is used to obtain the magnitude frequency response of each frame.

(e) Mel-frequency wrapping

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Therefore for each tone with an actual frequency is measured in Hz,. We can use the following approximate formula to compute the mels for a given frequency f in Hz.

Formulae:

$Mel(f) = 2595*\log10(1 + f/700)$………………………………………… (2.11)

The mel scale filter bank is a series of triangular band pass filters that have been designed to simulate the band pass Filtering believed to occur in the audible system.. [8].

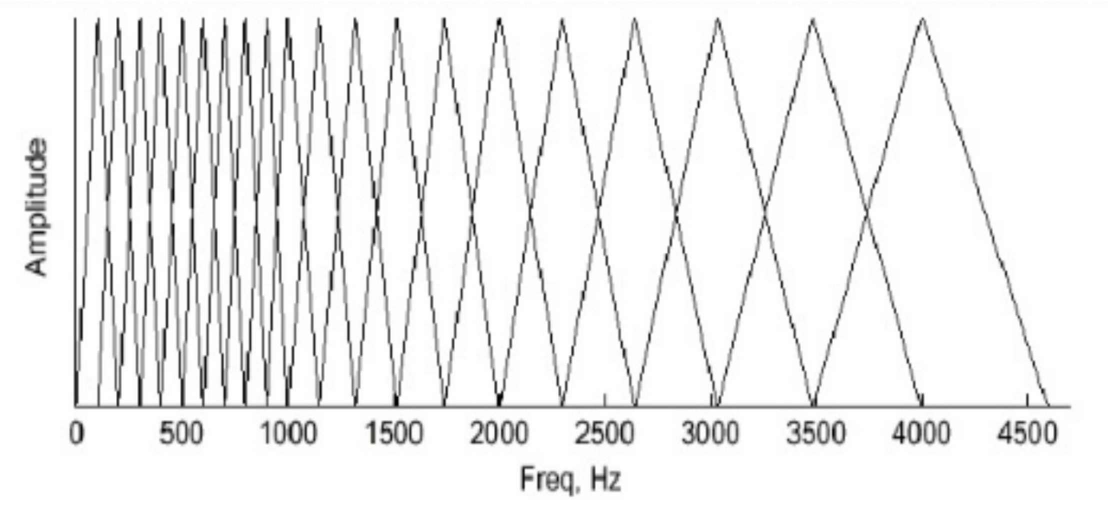


Fig 2.24: Filter bank in Mel frequency scale

(f) Discrete cosine transform or DCT

In this step apply DCT on the 20 log energy E_k obtained from the triangular band pass filters to have L mel-scale

Cepstral coefficients.

Formulae:

Cm = Sk=1

N cos [m*(k-0.5)*p/N]* E_k, m=1,2, ..., L……………………………… (2.12)

Where N is the number of triangular band pass filters, L is the number of Mel Scale Cepstral Coefficients. Set N=20 .

And L=12 performed FFT. DCT converts the frequency domain into a time domain called quefrency domain. The obtained features are same as cepstrum, thus it is referred to as the mel-scale Cepstral coefficients (MFCC). [11]

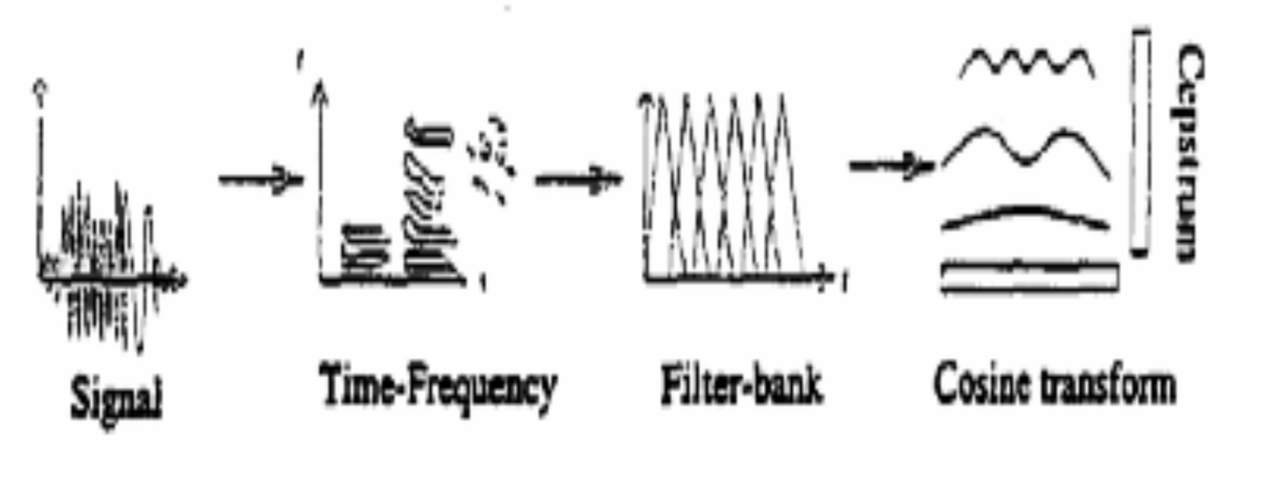


Fig 2.25: Pictorial representation of mel-frequency cepstrum (MFCC) calculation

Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is Shifted by 10 ms. each frame is converted to 12 MFCCs plus a normalized energy parameter.

In my opinion there are two ways of looking to the MFCCs:

(a) As a filter-bank processing adapted to speech specificities.

(b) As a modification of the conventional cepstrum .

A well-known deconvolution technique based on homomorphic processing we used the mel scale to organize the filter bank used in MFCC calculation., Cepstrum is the most popular homomorphic processing because it is useful for deconvolution. The basic human speech production model adopted is a source-filter model.[6]

Source: is related to the air expelled from the lungs. If the sound is unvoiced, like "s" and "f", the glottis is open and the vocal cords are relaxed. If the sound is voiced, "a", "e", for example, the vocal cords vibrate and the frequency of this vibration is related to the pitch.

Filter: is responsible for giving a shape to the spectrum of the signal in order to produce different sounds. It is related to the vocal tract organs.

The problem is: source e(n) and filter impulse response h(n) are convoluted. Then we need deconvolution in speech recognition applications. Mathematically:

The problem is: source e(n) and filter impulse response h(n) are convoluted. Then we need deconvolution in speech recognition applications. Mathematically:

In the time domain, convolution: source * filter = speech,

e(n) * h(n) = x(n) ………………………………………………………..… (2.13)

In the frequency domain, multiplication: source x filter = speech,

E(z) H(z) = X(z)…………………………………………………………..(2.14)

We can make the deconvolution ? Cepstral analysis

Use the logarithm to transform the multiplication in (2.14) into a summation (log ab = log a + log b). It is not easy to separate (to filter but it is easy to design filters to separate things that are parcels of a Sum as below: [6][8]

C(z) = log X(z) = log E(z) + log H(z)………………………………………..(2.15)

If we were dealing with E(z) + H(z). In fact, we have, the following equation:

Co(z) = E(z) + H(z)………………………………………………………… (2.16)

2.9.4 Linear Predictive Coding (LPC)

A very powerful method for speech analysis is based on linear predictive coding (LPC), also known as LPC analysis or auto-regressive (AR) modeling. This method is widely used because it is fast and simple, yet an effective way of estimating the main parameters of speech signals. All-pole filter with a sufficient number of poles is a good approximation for speech signals. Thus, we could model the filter H(z) in Figure 2.28 as

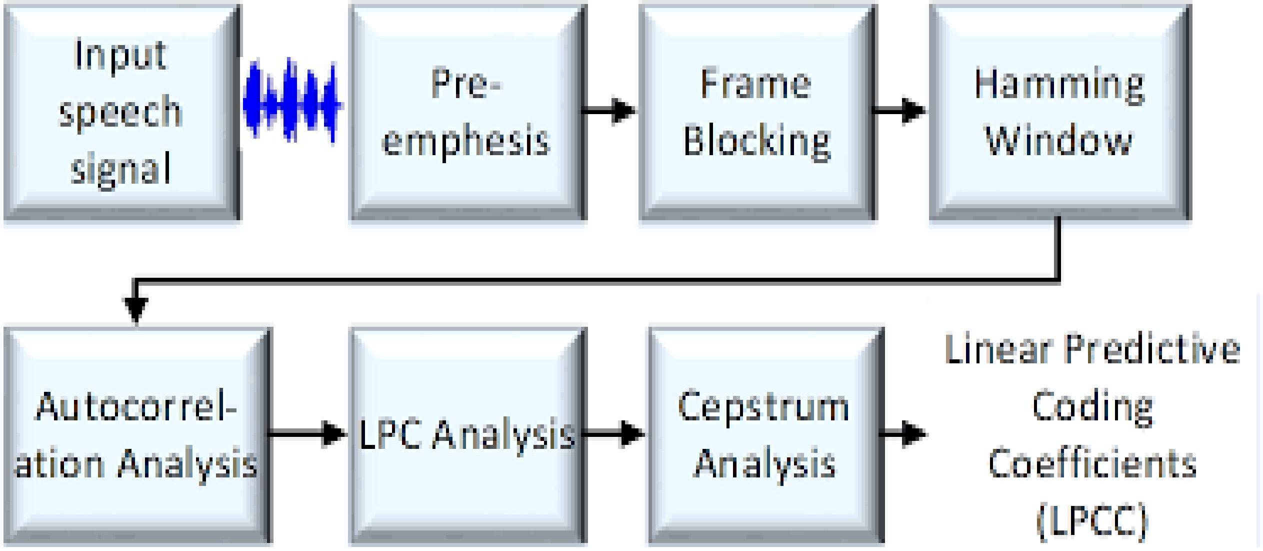


Fig 2.26:. Block diagram for LPC processor for Speech Recognition

(a)Pre – emphasis

From the speech production model it is known that the speech undergoes a spectral tilt of -6dB/Oct. To counteract this fact a pre-emphasis filter is used. The main goal of the pre-emphasis filter is to boost the higher frequencies in order to flatten the spectrum. Pre emphasis follows a 6 dB per octave rate. This means that as the frequency doubles, the amplitude increases 6 dB. This is usually done between 300 - 3000 cycles. Pre emphasis is needed in FM to maintain good signal to noise ratio. Perhaps the most widely used pre emphasis network is the fixed first-order system: [8].

Formulae:

$$H(z) = 1 - az^{-1}, \qquad 0.9 \leq a \leq .0 \qquad \text{……………………………………………} \quad (2.13)$$

(b) Frame – Blocking

In this step the pre-emphasized speech signal is blocked into frames of N samples, with adjacent frames being separated by M samples .Thus frame blocking is done to reduce the mean squared predication error over a short segment of the speech wave form. In this step the pre emphasized speech signal, S(n) is blocked into frames of N samples, with adjacent frames being separated by M samples[8].

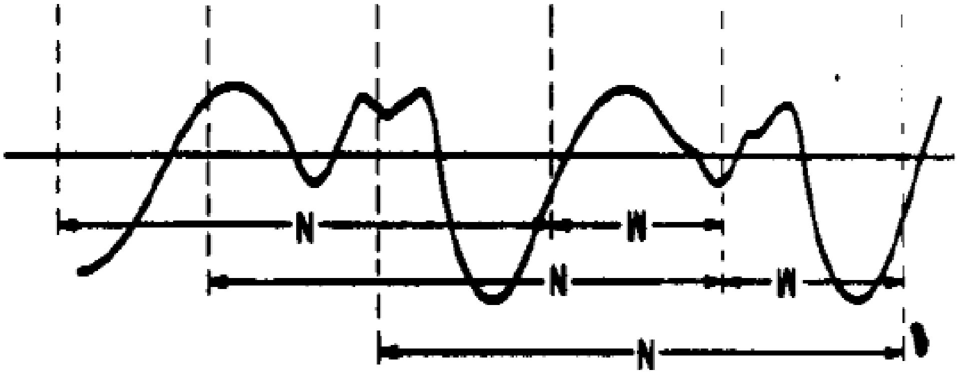


Fig 2.27. Blocking of speech into overlapping frames

Typical values for N and M are 256 and 128 when the sampling rate of the speech is 6.67 kHz. These correspond to 45-msec frames, separated by 15-msec, or a 66.7-Hz frame rate. [8].

. (c) Windowing

Here we want to extract spectral features of entire utterance or conversation, but the spectrum changes very quickly. Technically, we say that speech is a non-stationary signal, meaning that its statistical properties are not constant across time. Instead, we want or extract spectral features from a small window of speech. [1].

$$\begin{array}{ll} w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \ (2.2) \\ 0 \quad , & Else\ where \end{array} \qquad \text{........... .... (2.14)}$$

(d) Autocorrelation analysis

Each frame of windowing signal is next auto correlated to give

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), m = 0,1,2,\dots\dots\dots, P, \quad \dots\dots\dots\dots \dots\dots\dots\dots \dots\dots\dots\dots \quad (2.15)$$

Where the highest autocorrelation value, p, is the order of the LPC analysis. Typically $R_l(0)$, , values of p from 8 to 16 have been used, with p = 10 being the value used for this systems. A side benefit of the autocorrelation analysis is that the zeros autocorrelation, $R_l(0)$, is the energy of the l^{th} frame.

(e) LPC Analysis

The next processing step is the LPC analysis, which converts each frame of P+1 autocorrelations into an LPC parameter set in which the set might be the LPC coefficients, [6]

(g) LPC parameter conversion to Cepstral coefficients

A very important LPC parameter set, which can be derived directly from the LPC coefficients set, $c(m)$ is the LPC Cepstral coefficients, the recursion method is used. The Cepstral coefficients, which are the coefficients of the Fourier transform representation of the log magnitude spectrum [6]

2.10 Classification Methods

There are two major types of models for classification: stochastic models (parametric) and template models (non-parametric) .

In stochastic models, the pattern matching is probabilistic (evaluating probabilities) and results in a measure of the likelihood, or conditional probability, of the observation given the model. Here, a certain type of distribution is fitted to the training data by searching the parameters of the distribution that maximize some criterion. Stochastic models provide more flexibility and better results. [8].

2.10.1 Hidden Markov Model (HMM)

We designed and implemented a hidden Markov model (HMM) that optimally matches the behavior of a set of training sequences that will be provided as part of this project. The goal will be to use the standard set of forward estimation algorithms to optimally determine the best (maximum likelihood) HMM that matches a given set of training data. We also used the Viterbi algorithm for estimating the model parameters and comparing and contrasting the results for the two methods.

As mentioned in the introduction part the technique used to implement the speech recognition system was the Hidden Markov Model, HMM. The technique is used to train a model which in our case should represent a utterance of a word (Allah mofakhem, Allah mouraqeq and sunny, moony لام). This model is used later on in the testing of an utterance (Allah mofakhem, Allah mouraqeq and sunny, moony لام)

In speech proceeding, both deterministic and stochastic models have had good success, one type of stochastic model, namely hidden Markov model (HMM) .(these models are referred to as Markov source s or probabilistic functions of markov chains in the communications literature .[10]

Until now, this is the most successful and most used pattern recognition method for speech recognition. [13]

2.10.1.1 Basic Concepts

A Hidden Markov Model is a collection of states connected by transitions, as illustrated in Figure 2.32 It begins in a designated initial state. In each discrete time step, a transition is taken into a new state, and then one output symbol is generated in that state. The choice of transition and output symbol are both random, governed by probability Distributions.

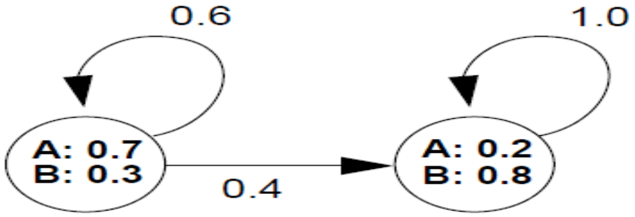


Fig2.28: A simple Hidden Markov Model, with two states and two output symbols, A and B.

Speech always goes forward in time; transitions in a speech application always go forward (or make a self-loop, allowing a state to have arbitrary duration). Figure 2.30 illustrates how states and transitions in an HMM can be structured hierarchically,. .[9]

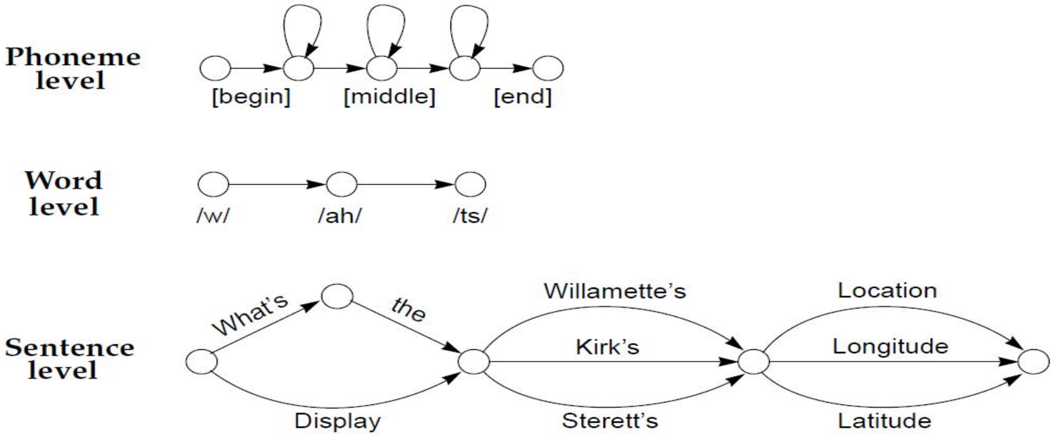


Fig2.*29*: illustrates how states and transitions in an HMM can be structured hierarchically

The hidden Markov Model is represented by λ = (π, A, B).

π = initial state distribution vector.

A = State transition probability matrix.

B = continuous observation probability density function matrix. .[9]

2.10.1.2 The three fundamental problems in the Hidden Markov Model design are the following

- **Problem one - Recognition**

Given the observation sequence O = (o_1, o_2,...,o_T) and the model λ = (π, A, B), how is the probability of the observation sequence given the model, computed? That is, how is P(O|λ) computed efficiently?

Problem two - Optimal state sequence

Given the observation sequence O = (o_1, o_2,...,o_T) and the model λ = (π, A, B), how is a corresponding state sequence, q = (q_1, q_2,...,q_T), chosen to be optimal in some sense (i.e. best “explains” the observations)?

Problem three – Adjustment

How are the probability measures, λ = (π, A, B), adjusted to maximize P(O|λ)?

2.10.1.3 The forward algorithm

Formulae:

Step 1: Initialization

$$\alpha_1(i) = \pi_i b_i(\mathrm{X}_1) \qquad 1 \le i \le N$$

………… …… (2.16)

Step 2: Induction

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(X_t) \qquad 2 \le t \le T;\ 1 \le j \le N$$

……………… … (2.17)

Step 3: Termination

$$P(\mathbf{X}|\Phi) = \sum_{i=1}^{N} \alpha_T(i) \quad \text{If it is required to end in the final state, } P(\mathbf{X}\,|\,\Phi) = \alpha_T(s_F)$$

……… (2.18)

2.10.1.4 The Viterbi Algorithm

An alternative approach to the forward method, for re-estimation of HMM model parameters, is the Viterbi algorithm which inherently finds the best matching path that best aligns each training sequence with the current model. By keeping track only of the best sequence, the computation can be reduced considerably. However, in cases where the best path is not much better than alternative paths, the Viterbi algorithm can lead to highly sub-optimal solutions. This occurs more often with randomly created HMM model parameters than with skewed model parameters, or with left-right models which have strong constraints on the state prior density and the state transition density. [29]

2.11 Neural Networks

Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. Roughly speaking, a neuron in an artificial neural network is

1. A set of input values (xi) and associated weights (wi).
2. A function (g) that sums the weights and maps the results to an output (y).[11]

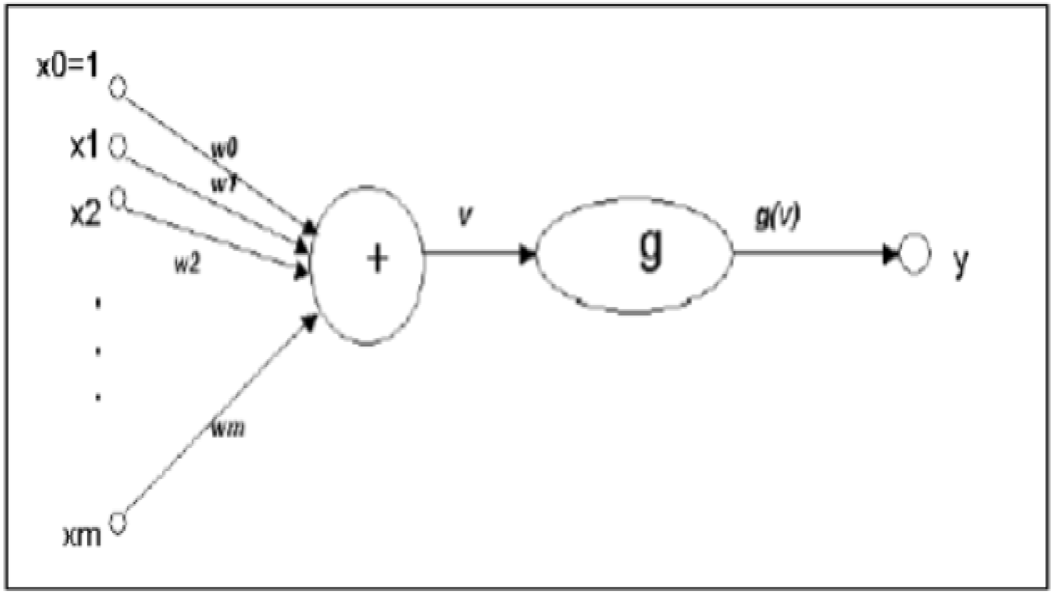


Fig 2.30. Neurons in Neural network

Pattern matching problem that is amenable to neural networks; therefore we use neural networks for acoustic modeling, while we rely on conventional Hidden Markov Models for temporal modeling, two different ways to use neural networks for acoustic modeling, namely prediction and classification of the speech patterns. Prediction is shown to be a weak approach because it lacks discrimination, while classification is shown to be a much stronger approach..[9]

Neural networks have many similarities with Markov Models. Both are statistical models which are represented as graphs. Where Markov models use probabilities for state transitions, neural networks use connection strengths and functions. A key difference is that neural networks are fundamentally parallel while Markov chains are serial. Frequencies in speech occur in parallel, while syllable series and words are essentially serial. .[12]

The neurons in the feed forward networks can always be split into layers which are ordered (e.g. arranged one over another) so that the connections among neurons lead only from lower layers to upper ones and generally, they may skip one or more layers. Especially, in a multilayered neural network, the zero (lower), input layer consists of input neurons while the last (upper), output layer is composed of output neurons. The remaining, hidden (intermediate) layers contain hidden neurons. In the topology of a multilayered network, each neuron in one layer is connected to all neurons in the next layer. Therefore, the multilayered architecture can be specified only by the numbers of neurons in particular layers, An example of a three-layered neural network 3-4-3-2 with an indicated path is in Figure 2.35 which, besides the input and output layers, is composed of two hidden layers.[12]

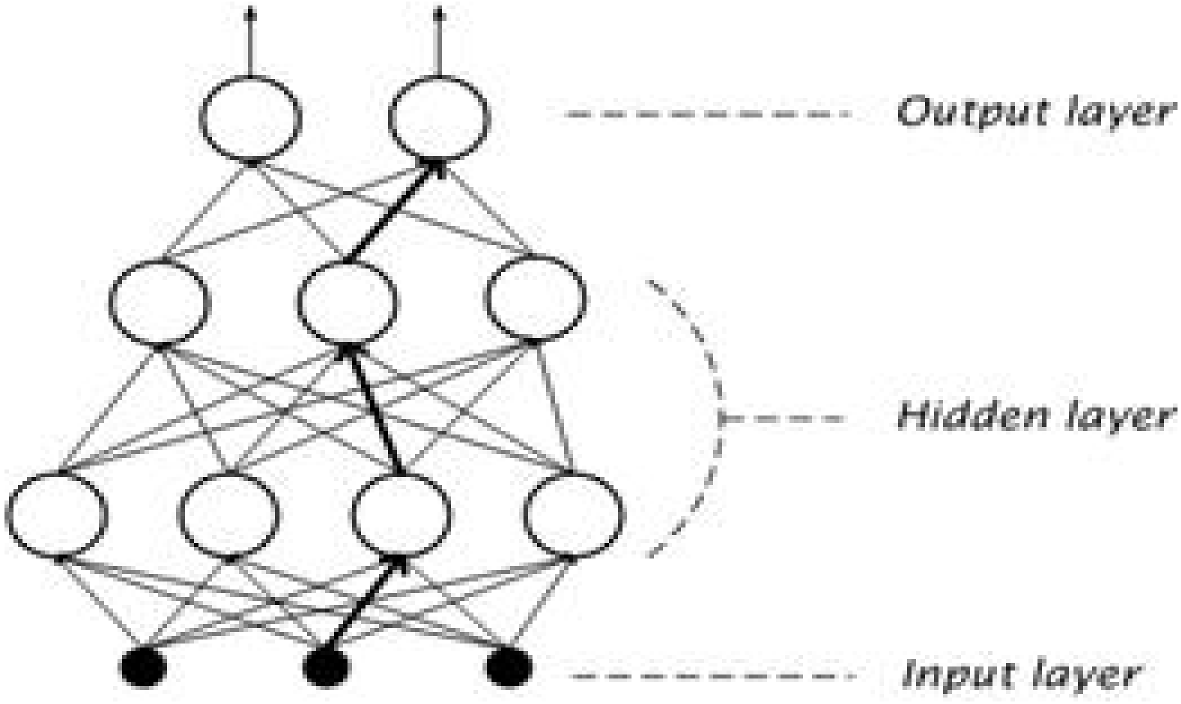


Fig 2.31. Example of architecture of multilayered (three-layered) neural network

2.11.1 Fundamentals of Neural Networks

In this section we will briefly review the fundamentals of neural networks. There are Many different types of neural networks, but they all have four basic attributes:

- A set of processing units;
- A set of connections;
- A computing procedure;

- A training procedure.

Let us now discuss each of these attributes.

2.11.1.1. Processing Units

A neural network contains a potentially huge number of very simple processing units, roughly analogous to neurons in the brain. The units in a network are typically divided into input units, which receive data from the environment (such as raw sensory information); hidden units, which may internally transform the data representation; and/or output units, which represent decisions or control signals (which may control motor responses, for example). In drawings of neural networks, units are usually represented by circles. [12]

2.11.1.2 Connections

The units in a network are organized into a given topology by a set of connections, or Weights, shown as lines in a diagram . Each weight has a real value.

A network can be connected with any kind of topology. Common topologies include Unstructured, layered, recurrent, and modular networks, as shown in Figure 2.35 Each kind of topology is best suited to a particular type of application. For example:

- Unstructured networks are most useful for pattern completion (i.e., retrieving stored patterns by supplying any part of the pattern);
- Layered networks are useful for pattern association (i.e., mapping input vectors to Output vectors);
- Recurrent networks are useful for pattern sequencing (i.e., following sequences of Network activation over time).
- Modular networks are useful for building complex systems from simpler components.[12]

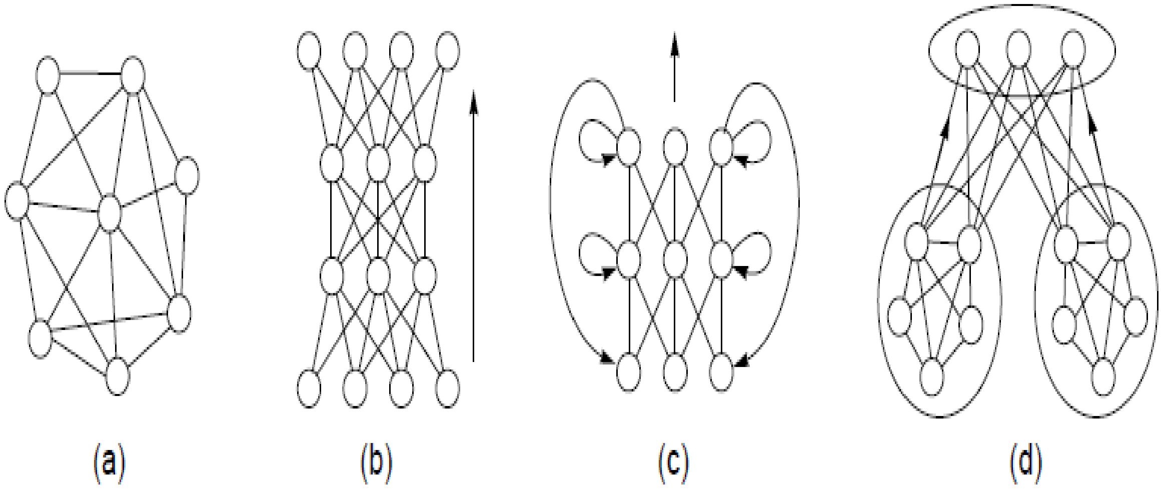


Fig 2.32. Neural network topologies: (a) unstructured, (b) layered, (c) recurrent, (d) modular.

2.11.1.3 Computation

Computation always begins by presenting an input pattern to the network, or clamping a pattern of activation on the input units. Then the activations of all of the remaining units are computed, either synchronously (all at once in a parallel system) or asynchronously (one at a time, in either randomized or natural order), as the case may be. In unstructured networks, this process is called spreading activation; in layered networks, it is called forward propagation, as it progresses from the input layer to the output layer.[9]

2.12 Altajweed Definition

Altajweed in language the perfection and improvement and idiotically: it is the science that shows rules and grammar that must be applied when reading the Holy Quraan based on what muslims received from god's messenger (peace be upon him) by giving each letter it's right in exit and vowel and adjective without any addition.

The main advantage of learning Altajweed is in maintaining tongue from errors and music when reading god's words, Altajweed is being learnt by listening to a sheikh, or to read in front of the sheikh and the sheikh will correct it for the reader, the best is to combine between those methods considering that this science is not available within books but you have to go back to Altajweed scientists to learn from them, where there are many rules that cannot be learned unless through communicating with them.

Scientists have disagreed in learning and mastering Altajweed in two ways: first: Altajweed the Holy Quran and considering its rules and grammar is a Sunna of Recitation and we must stick to it when reading the Quran and it is not a must, second: learning Altajweed is a sufficient farad, while reading it is a must on ever Muslim and this is said by most Altajweed scientists: [17]

2.12.1. اللامrules (the tense, none, proposition, command and holiness)

Have many rules based on its position and we will speech about four different cases of لاَم

- The لام
- The tense لام
- noun, propositions and command لام
- Holiness nameلام

1. The لام:

it is an excessive one from the word structure and is specialized in alnakra names only to define them, whether it was right to strip it from this ل like (المؤمنون) or not like (الذي والتي) and when this ل which we entered alhamza to it to ease its utterance in case it started the sentence and the rule here is specialized with this L which is proper to strip it from the word and here we have 2 cases - [19]

If it was followed by one of alhijaa letters-they are

A- Moon diphthong

It is the process of revealing: the: when it comes after it one of the following letters (ابغ حجك وخف عقيمةthey are 14 letters as the لام is shown clearly and the showing is pronounced by uttering the first letter which is constant لام and the letter which follow it is softened like:(الأو ّل ، الباسط ، الغفور ، الحكيم ، الجليل ، الكريم ، الودود ، الخبير ، الفصل ، العليم ، القاهر ، اليقين ، الملك ،الهادي) [16]

- **Note:**

  Most people fall for the same error which is the لام that preceed ج where most pronounce it as if it is a soften لامthus we must pay more attention to it.

2- noun (التي والذي) are not labeled as sunny nor moon because they are of the word origin but if we considered the لام that entered on اسم الموصول such as الذي والتي واللذان... it is considered as an additional and thus it has no role as it this followed by the لامis being identified then this لام is sunny [16]

B– Sunny diphthong

: It is the process of deleting the in utterance if it came before one of those 14 letters which are in the following:

طب ثم صل رحِـماً تفز ضف ذا نـعم   دع سوء ظنٍّ زر شريفاً للكرم

it is named like this because we must diphthong اللام as it appears correctly writhing الشمس where there is no sign of the اللام into this word and thus we only read one stressed letter in order to lessen the utterance due to the difficulty of returning the tongue to the first exit, scientists chose diphthong because it is easier to be pronounced plus this type of diphthong is not used unless are the النون because it will be stressed like الناس,النارand to give a closer look we show the اللام as stars and moon letters as the moon and the sunny letters as the sun, we notice that those stars are revealed clearly in front of the moon while they disappear in front of the sun. [16]

Altajweed scientists have found a mark which is alshada after اللام and which is placed after sunny letters to recognize weather this diphthong is sunny like الطّارق ، الصّابرين ، الرّحمة ، الظّالمين ، السّماء ، الزّكاة ، الشّيطان

This mark does not exist on moon letters where it is replaced by sukoun like

الأرض ، الحليم ، الكافرين ، المؤمنين ، العدل ، الحرث ، الخيل ، البنين ، القناطير ، الفضة ، الهاوية

لام tense:

It must be shown within the tense, as it is within the origin and the structure of this tense whether it was a past tense like: **ألْفيا ، ألْهاكم ، الْتقى ، جعلْنا..**

Or present like:﴿ ﴿يَلْتَقِيَانِ يلْتَقِطْهُ﴾﴾.

Or future tense likes:

.. ﴿ قُلْ نَعَمْ ﴿ أَلْقِهَا يَا مُوسَى ﴾ ..

we must mention that the future tense لام came at the end of the word like (قل and if it was followed by)ل,{ then it is not shown and if there is a constantلام and then followed by a vowel لام then we must place a diphthong because of similarity and they will be combined as one stressly pronounced letter like

قل لّا أملك لنفسي ضرّاً ولانفعاً ،هلْ لّكم,

As if it came between a constant لام and followed by a vowel راء like

وقلْ رّبِّي زدْني علماً، بلْ رّفعه [16]

اللام has been shown within the tense and it was not diphthong because النون do not diphthong any letter and it is found in this word(يرملون) and if it was diphthong then it will remove the harmony between it and other letters, as for اللام and النون**النّاس** ، **النّار**..

Plus this **لام** is an additional one and it is not independent like the tense like أ [16]

3-noun and letter and command لام:

A- noun: it is and original one and not additional one and it must be shown all the Time like لام

﴿أَلْفَافًا﴾ ﴿سُلْطَانٍ﴾ ﴿أَلْسِنَتُكُمْ﴾ ، ﴿أَلْوانكم﴾ ...

B-letter: it must be shown all the time like لام:

[16] ...﴿هَلْ عَسَيْتُمْ﴾ ، ﴿بَلْ أَنتُمْ﴾ ، ﴿هَلْ أَدُلُّكُمْ﴾ ، ﴿بَلْ طَبَعَ﴾ ، ﴿هَلْ يَسْتَطِيعُ﴾

C- : command it is the one that is shown into present tense and it must be shown all the time likeلام ، ﴿وَلْيَطَّوَّفُوا﴾ ﴿وَلْيَكْتُبْ﴾[16]

4- Holiness nameلام:

It has two cases maximaizing and softenin.

A – Maximizing:

It is used with the لام for mazimizing and with الراء for softness it is some kind of voiced soundness that enters on the letter to give and echo to this name.

It is maximized within the following three cases: [16]

- If it was followed by damma or fatha like:

:﴿قَالَ اللّهُ﴾ ﴿مَا﴾ ﴿قَامَ عَبْدُ اللَّهِ﴾ ، ﴿هُوَ اللَّهُ أَحَدٌ﴾ ﴿نَصْرُ اللَّهِ﴾-

- If it was followed by silence after fatha or sielence after damma

﴿إِلَى اللّهِ﴾ ﴿قَالُوا اللَّهُمَّ﴾...

- If it was the beginning of the sentence like:

[16]﴿اللّهُ لاَ إِلَهَ إِلاَّ هُوَ الْحَيُّ الْقَيُّومُ﴾ ..

B- Softening

It is the thinning language and idiomatically it is a tone that enters on the name so the mouth is not filled with echo".

اللام is being soften within the following three case:

1- if it is followed by kassra like:

﴿بِاللّهِ﴾﴿بِسْمِ اللّهِ﴾﴿يَفْتَحِ اللَّهُ﴾﴿قُلِ اللَّهُمَّ ﴾، ﴿فِي دِينِ اللَّهِ﴾

2-if it is followed by sukoun after kasra like

﴿ينجِي الله ﴾

3-if it is followed by tanoien like

إ﴿قَوْمًا اللّهُ﴾[16].

CHAPTER 3

CHAPTER THREE

IMPLEMENTATION AND RESULTS

3.1 Introduction

This chapter shows the implementation details of practical steps of research. It also shows the steps required to achieve the complete (Speech recognition : case study learning the rules of Tajweed) process. Also, it introduces the project testing using different testing environments and the results after testing. at the beginning described and explained the Matlab Package Which have been selected to be a programming tool in this research due to its recognized in the field of signal processing , neural network and hidden markov models to facilitate interaction with the system and make it very comfortable, and easy to use.

3.2 Matlab Package

MATLAB [1] (Matrix Laboratory) is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. Furthermore, MATLAB is a modern programming language environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research.

MATLAB has many advantages compared to conventional computer languages for solving technical problems. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. The software package is now considered as a standard tool at most universities and industries worldwide.

the language dealing with data in the form of matrices this language contains a large number of mathematical functions in the areas of programming can also write programs or private functions to solve a particular problem There are Specific applications are collected in packages referred to as toolbox. Such as :

- .Signal processing,
- Control systems
- Neural Networks
- Hidden models
- Simulation
- And other fields

In this research we used Matlab Package in signal processing and in building neural networks and hidden markov models. Due to which we mentioned before about characterized of the language has powerful built-in routines that enable a very speed of programming processing. It also has easy to use graphics commands that make the visualization of results immediately available.

3.2.1 Digital Signal Processing in Matlab

This System includes groups of tools depended on the calculation system. This application is characterized by existence of a wide range of special functions to signal processing. Like

- Create signal
- Design filter
- Spectral analysis
- And another.

There is much Function for signal processing in Matlab program these tools divided into signal processing Function and graphical interactive tools in Matlab program file called (M.file). In this research there was May Function they were used.

3.2.2 Neural Network in Matlab

Neural Network Toolbox provides algorithms, functions, and apps to create, train, visualize, and simulate neural networks. You can perform classification, regression, clustering, dimensionality reduction, time-series forecasting, and dynamic system modeling and control.

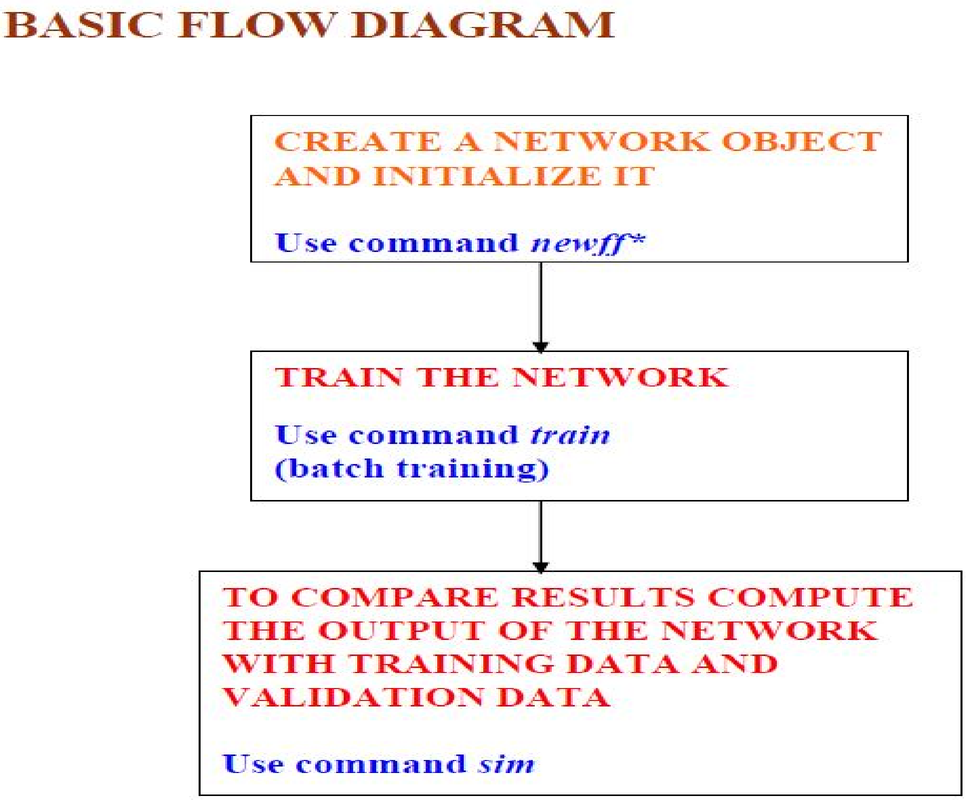

BASIC FLOW DIAGRAM
CREATE A NETWORK OBJECT AND INITIALIZE IT
Use command *newff**
TRAIN THE NETWORK
Use command *train*
(batch training)
TO COMPARE RESULTS COMPUTE THE OUTPUT OF THE NETWORK WITH TRAINING DATA AND VALIDATION DATA
Use command *sim*

Fig 3.33. Basic Flow Diagram for Neural network

The command newff both defines the network (type of architecture, size and type of training algorithm to be used). It also automatically initializes the network. The last two letters in the command newff indicate the type of neural network in question: feed forward network. We used some function concerning with feed forward network: [28] .

- **Create a Network**
- **hiddenLayerSize = 40;**
- **net = newff(inputs,targets,hiddenLayerSize);**
- **Training Function train(net ,p ,t)**

net: network created before

p : input matrix

t : output matrix (target)

- **Simulated Function sim(net ,p)**
  Resulting from this function is the output after training
- **net.trainFcn = 'trainscg';**
- **net.performFcn = 'mse';**
- **plotFcns = {'plotperform','plottrainstate','ploterrhist', ...**
- **'plotregression', 'plotfit'};**
- **Special transaction for toolbox for create NN in matlab :**

```
net.trainParam.max_fail = 6;
net.trainParam.min_grad=1e-5;
net.trainParam.show=10;
net.trainParam.lr=0.9;
net.trainParam.epochs=1000;
net.trainParam.goal=0.00;
```

- **Patternnet**

Syntax : patternnet (hidden sizes, trainfcn)

Pattern recognition networks are feed forward networks that can be trained to classify inputs according to target classes .the target data for pattern recognition networks should consist of vectors of all zero values except for a1 in element I, where is the class they are to be represent , and takes this arguments, hidden sizes :

Row vector of one or more hidden layer sizes(default =10), trainfcn: training function(default =train scg) , and return pattern recognition neural network .

3.2.3 Hidden Markov in Matlab

This package contains functions that model time series data with HMM. It Includes Viterbi, HMM filter, HMM smoother, EM algorithm for learning the parameters of HMM, etc.

- **function VITERBI(observations of len T,state-graph)** returns best-path

3.3 Sound Feature Extraction

The purpose is to convert the speech waveform into a set of features or rather feature vectors for further analysis.

The speech signal is a called a quasi-stationary signal i.e. a slowly timed varying signals. An example of a speech signal is shown in Figure 3.38.

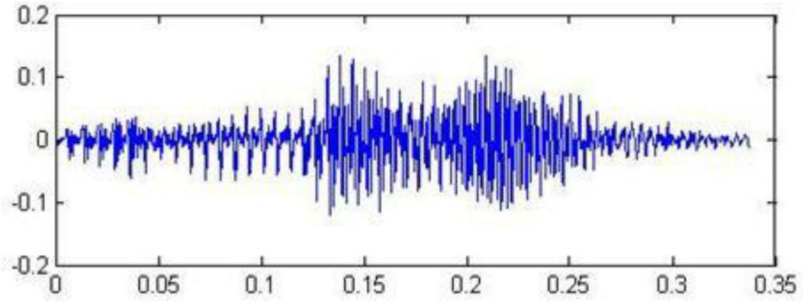


Fig 3.34: *Example of a Speech Signal*

In these parts described the practical parts for sounds feature extraction, in many levels, the first level is spectral shaping for recorded recitation for man’s readers and two women’s readers this excitation included the four rules discussed in this project. The second level is spectral analysis with a verity of methods used in speech recognition such as (MFCC&LPCC). The last level is removing the noise from the signal by used MATLAB package its rich with the functions for signal processing and another mathematical functions.

feature extraction for Allah (moufakham , mourqeq) for three readers' (sherif , umahmed ,umhajer) with 60 samples ., and two famous readers Alsudies and Alhuzafin with 50 samples , used (MFCC ,LPC) to extract the features .for Allah (moufakham,mouraqeq) extracted just the name of Allah by using trimming tool and the Audacity Program to do that . for moony and sunny لام we cutting the words which collected just extracted the لام for each clip alone (one by one) ,the result is different lengths between (8000 to 12000) , so we can't save it in one matrix , we saved it in the file ,then used the (MFCC&LPC) to extract the features , we save those extracted features in one matrix , to used it after with NN & HMM in identification level .

3.3.1 Mel Frequency Cepstrum Coefficients (MFCC)

By applying the procedure mel-frequency cepstrum coefficients (MFCC) The set of coefficients is called an acoustic vector. Thus each input speech utterance is transformed into a sequence of acoustic vectors. The tables for features extractions explained the differences in the value of feature vectors between Allah"(moufakham,mouraqeq) .

3.3.2 Linear Predictive Coding (LPC)

There are 12 parameters and 60 samples for three readers and 2cases , there is matrix with 3 dimension , the third dimension consist of two layers they are concerning with allah (moufakham & mouraqeq) the table for feature extraction to all samples for 3 readers 20 samples for each readers explained the differences in the value of feature vectors.

3.4 Signal Modeling

3.4.1Spectral Shaping

I recorded recitation for three readers (one man & two women) for Allah (moufakhum , moureqeq) and (sunny , moony) لام hokums in Altajweed in special room with no noise at the background , with the Golden wave V5.70 program these samples recorded with computer include sound card , the samples converted from analog-to-digital (A/D) form . All samples recorded with

- Sample frequency 44100 HZ fixated in all audio clip recorded
- Resolution 15 bit
- Single mono channel
- Head phone set microphone for record

These samples saved in the hard disk in audio form (wav). Like tables below:

Words for Allah Moufakhum with three readers (Sherief/ umahmed/umhajer):

1. الله
2. والله
3. أعبدوا الله

Words for Allah Moureqeq with three readers (Sherief/ umahmed/umhajer):

1. لله
2. قوماً الله
3. بسم الله
4. قل الله

Words in Quran for Sunny لام with three readers (Sherief/ umahmed/umhajer):

1. **الطامة**
2. **الثواب**
3. **الصابرين**

4. **الرحمن**
5. **الضالين**
6. **النور**
7. **الدين**
8. **السماء**
9. **الظالمين**
10. **الزكاة**
11. **الشمس**
12. **الليل**

Words in Quran for moony لام with three readers (Sherief/ umahmed/umhajer):

1. **الأول**
2. **البر**
3. **الغنى**
4. **الحكيم**
5. **الكبير**
6. **الودود**
7. **الخبير**
8. **الفتاح**
9. **ذو الجلال**

3.4.2 Spectral Analysis

There are two methods for express the signal, the primary way description and representation method in time domain, the second way description and representation method in frequency domain .was signal representation Algebraic sum of sinusoidal signals of the number of different frequencies, different lengths, these representation called spectral representation signal consist of mixture Of frequency signal, any frequency any signal has own spectral to distinguish it from another signal.

There is many, there are a set of methods of spectral analysis regard to the speech recognition such as

- Mel Frequency Cepstral Coefficients Processor(MFCC):
- Linear Predective coding function (LPC)
- Fast Fourier Transform Function (FFT)
- Reflection Coefficient
- LP-derived cepstral coefficients
- And anothers

For spectral analysis two methods chosen to the system :

- Mel Frequency Cepstral Coefficients Processor(MFCC):
- Linear Predective coding function (LPC)

• Mel Frequency Cepstral Coefficients Processor(MFCC)

MFCC's is a type of algorithm i.e. basically used to define relationship between human ear's critical bandwidths with frequency. This method is basically used for analyzing and extraction of pitch vectors Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is shifted by 10 ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. [7]

• Linear Predective coding function (LPC)

Calculate predictive parameter for signal in the form of a matrix

$$a=LPC(X)$$

a: value of transaction matrix
X: signal

3.5 System Implementation

In this section all the implementation details are presented including the software used associated with some facilitation figures were used with the system.

In the collect data base the first step through the golden program wave is the readers speak a word into a microphone. The electrical signal from the microphone is digitized by an "analog-to-digital (A/D) converter", so the program must first be "trained" with a word from a reader voice input after that the word which determined by the system can be recognized During a testing session, the program displays a printed word or phrase of the recognized word .

3.5.1 Software Components

We used gold wave program to record the samples of data base and Audacity program to cut the word included any four hakem from the long passage of Quran. Also we used special script like (file_seg) for cutting each moony and sunny (لام) from the word. Because it’s easy to recognition and classify it, we invented special trimming tools to cut the words included each of the hakem,s from the long passage of Quran .

3.5.2 Hardware Components

Headset and Microphone

It is used as voice input unit to computer and also as speaker for output voice.

The headset is configured as wanted using special computer software (Gold wave) to record the samples words from the readers .

- **Helping Tools (File_ seg)**

This is special script to deal with moony and sunny لام .for cutting sound clip , just extract the moony and sunny لام and forming matrices for each readers , for that we can’t collect all readers samples in one matrix , because the samples is different from on reader to another reader , for cutting 30.000 samples from the beginning for all readers , when we run the script , we must determine the file sound for any reader , the script will read all the file samples automatically ,and save it in special folder with the name determined in script .

- **Trimming Tool**

We used this tools to deal with Allah name to extract exact the name of Allah ., and لام moony sunny to cut the لام , .determine the file of the clip and put start point and end point , ply the new clip ,if its accurate , save the file. This tool gives best more accurate result in cutting process. To get accurate features .this is the most important process in the accuracy of the system work.

3.6 Data Base

I recorded recitation for three readers (one man & two women) with Golden wave V5.70 and Audacity 2.1.2 program , Allah (moufakhum , moureqeq) and (sunny , moony) لام to four hokums in Altajweed , for Allah mufakhum there is 3 state in Quran each state recorded 10 times, for Allah moureqeq there is 4 states in Quran each state recorded 10 times, for sunny لام there is 12 states in Quran each state recorded 10 times and for moony لام there is 9 states in the Quran each state recorded 10 times .the total data base for each readers is 280 samples , the amount for three readers collected is 840 samples .

There was additional data base collected to increase the accuracy of the system for famous two readers (Alhuzafi&Alsudes) , The data for Alsedes reader Allah moufakhum (85 samples) , Allah muraqeq (50 samples) and for sunny لام (170 samples) , moony لام (98 samples).,the all data collected for alsudes = 403 .

And for Alhuzafi reader Allah moufakhum (60 samples), Allah muraqeq (30 samples) and for sunny لام (50 samples), moony لام (60 samples), the all data collected for alsudes = 200 samples from words samples just لامextracted, Used Audacity program to cut the samples (sunny, moony لام).

The amount data for two readers collected is 603 samples. The all data base for all readers used in the system = 1443 samples, used 80% of data for training, 20 % of data for testing.

Table 3.1: Words for (Allah Moufakham)

Word	No of audio record	Names of sheikh
الله	10	Sherief/ umahmed/umhajer
والله	10	Sherief/ umahmed/umhajer
أعبدو الله	10	Sherief/ umahmed/umhajer

Table 3.2: Words for (Allah Moureqeq)

Word	No of audio record	Names of sheikh
لله	10	Sherief/ umahmed/umhajer
قوماً الله	10	Sherief/ umahmed/umhajer
بسم الله	10	Sherief/ umahmed/umhajer
قل الله	10	Sherief/ umahmed/umhajer

Table 3.3: Words for (sunny) لام

Word	No of audio record	Names of sheikh
الطامة	10	Sherief/ umahmed/umhajer
الثواب	10	Sherief/ umahmed/umhajer
الصابرين	10	Sherief/ umahmed/umhajer
الرحمن	10	Sherief/ umahmed/umhajer
الضالين	10	Sherief/ umahmed/umhajer

النور	10	Sherief/ umahmed/umhajer
الدين	10	Sherief/ umahmed/umhajer
السماء	10	Sherief/ umahmed/umhajer
الظالمين	10	Sherief/ umahmed/umhajer
الزكاة	10	Sherief/ umahmed/umhajer
الشمس	10	Sherief/ umahmed/umhajer
الليل	10	Sherief/ umahmed/umhajer

Table 3.4: Words for (moony) لام

Word	No of audio record	Names of sheikh
الاول	10	Sherief/ umahmed/umhajer
البر	10	Sherief/ umahmed/umhajer
الغنى	10	Sherief/ umahmed/umhajer
الحكيم	10	Sherief/ umahmed/umhajer
الكبير	10	Sherief/ umahmed/umhajer
الودود	10	Sherief/ umahmed/umhajer
الخبير	10	Sherief/ umahmed/umhajer
الفتاح	10	Sherief/ umahmed/umhajer
ذو الجلال	10	Sherief/ umahmed/umhajer

3.7 Neural Network (NN)

After feature extraction, we want to train the network to identify the 4 hokhams , And how they are processed and how they are trained and what input and target matrix .Input matrix Are matrix that we want to training , each column represents coefficients for one-input situations (MFCC&LPC) ,target matrix It is a matrix linking transaction input for train network and right output. We must learn the network with 4 hokham by label them with no 1. There is 4 networks 2 for man’s reader and 2 for the women’s reader, and another 4 networks for alsudies reader and 4 networks for Alhuzafi reader.

I restarted trained the network what called pattern matching neural network its suitable for Matching Module, and gave good results .and better than common Neural network .create the network with 15 entry nodes and these entry nodes changed with each sheikh, also the performance I changed depend on the last result in performance it was not change and no of iteration = 12 changed depend on the samples.

net1 = patternnet(15);

And performance more than 0.57

```
while perf>0.25
```

net1 for man

```
net_allah_man_LPC
```

and net2 for women

```
net_allah_man_LPC
```

and the same things for MFCC , two network one for man , another for women .

```
net_allah_man_MFCC  , net_allah_woman_MFCC
```

Perf2 lpc

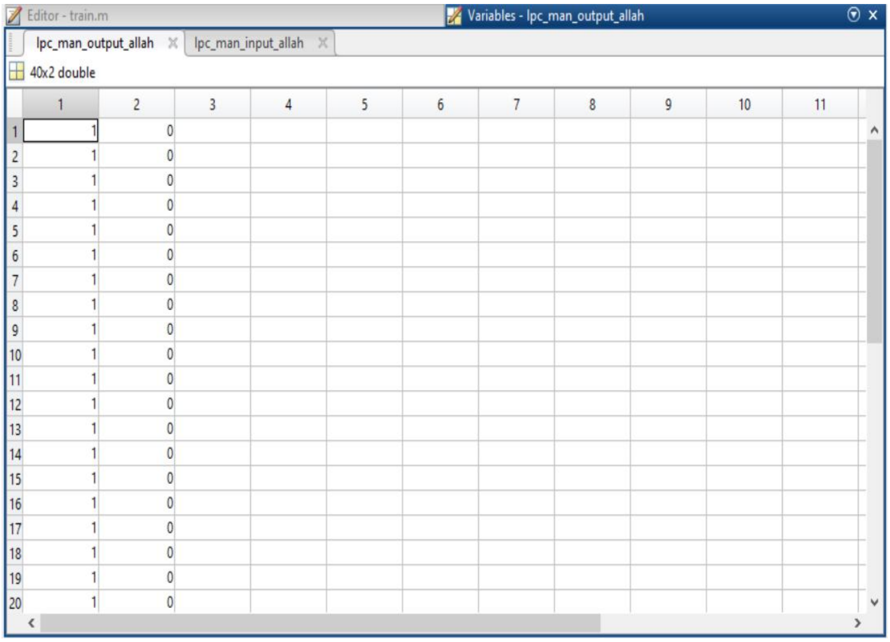
- **LPC Optimization**

To Optimized the audio clip (moony & sunny لام) the Values are very small and this affects the training network process so we need to re-representation is into another shape , so that we can train them Improving data to give the best results , Taking the logarithm of values becomes larger values Coefficients (moony & sunny لام) expressed in another way .

3.7.1 Input /Output MFCC _Allah

Before training the network we need to prepare input/output for each network , must be there input/output matrix (MFCC,LPC) for man reader and input/output matrix (MFCC,LPC) for women’s readers .there was 60 samples the first 20 samples for man reader , the 40 samples for women’s readers ,we rearrange those coefficients the matrix size for man reader MFCC (10,40) ,LPC(12,40) and for women’s reader same things except the samples were 80 samples ,and for Asudies reader were 50 samples and 50 samples for Alhuzafi reader the first 30 samples for Allah moufakham and the 20 samples for Allah morqeq . Thus we have formed input matrix for all readers .we defined two output matrices in the first readers one for man reader and one for women’s readers. The same things for Asudies reader and Alhuzafi reader there was two matrices .in the first column from (1-20) we put no 1 to indicate that this for Allah moufakham and in the second column from (20-40) we put no 1 to indicate that this columns for Allah mouraqeq .the same things for women’s readers

Table 3.5 LPC man_output_Allah



	1	2	3	4	5	6	7	8	9	10	11
1	1	0									
2	1	0									
3	1	0									
4	1	0									
5	1	0									
6	1	0									
7	1	0									
8	1	0									
9	1	0									
10	1	0									
11	1	0									
12	1	0									
13	1	0									
14	1	0									
15	1	0									
16	1	0									
17	1	0									
18	1	0									
19	1	0									
20	1	0									

3.7.2 Training & Testing:

We have trained network in particular type called pattern matching neural network is more convenient with a rating models which most appropriate to our work. It gives best results than general networks.

There is 2 networks one for man’s and one for women’s and 6 networks for (Alstudies,Alhuzafi) we want to trained it , we load matrices , and determine 15 hidden layers after saved we obtain one matrix for man’s and one matrix for women’s ,and 4 matrices for (Alstudies,Alhuzafi) we used it in Identification level .

- **pattern_net_allah_MFCC**

MFCC_man_input_allah.mat
MFCC_man_output_allah.mat
net2 = patternnet(22)
perf2>0.10
No of Iteration = 15

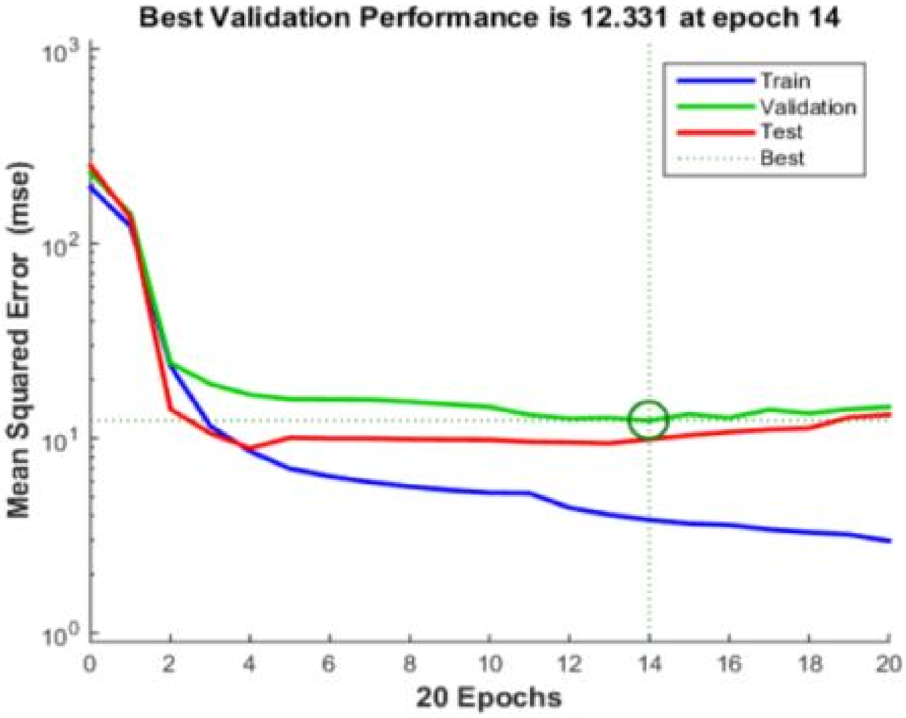


Fig 3.35: MFCC Neural Network training Allah (moufakhum,mouraqeq)

- **pattern_net_allah_LPC**

lpc_man_input_allah
lpc_man_output_allah
net1 = patternnet(15)
perf>0.25
No of iteration = 43

- **input_output_MFCC_A(moony sunny لام)**
  MFCC_man_input_A =12*180
  MFCC_woman_input_A = 12* 48
  MFCC_woman_input_A
  No of Iteration = 8

- **pattern_net_A_LPC**

lpc_man_input_A
lpc_man_output_A

patternnet(23)
perf>0.60

lpc_woman_input_A
lpc_woman_output_A
patternnet(32)
perf2>0.40

- **input_output_MFCCallah(alsudes)**

mfcc=12*40
$mfcc_2$=12*40

- **pattern_net_allah_MFCC(alsudes)**

net2 = patternnet(15);

perf2>0.05

- **pattern_net_allah_LPC**

lpc_man_input_A
lpc_man_output_A
net1 = patternnet(10);
perf>0.4
No of iteration =7

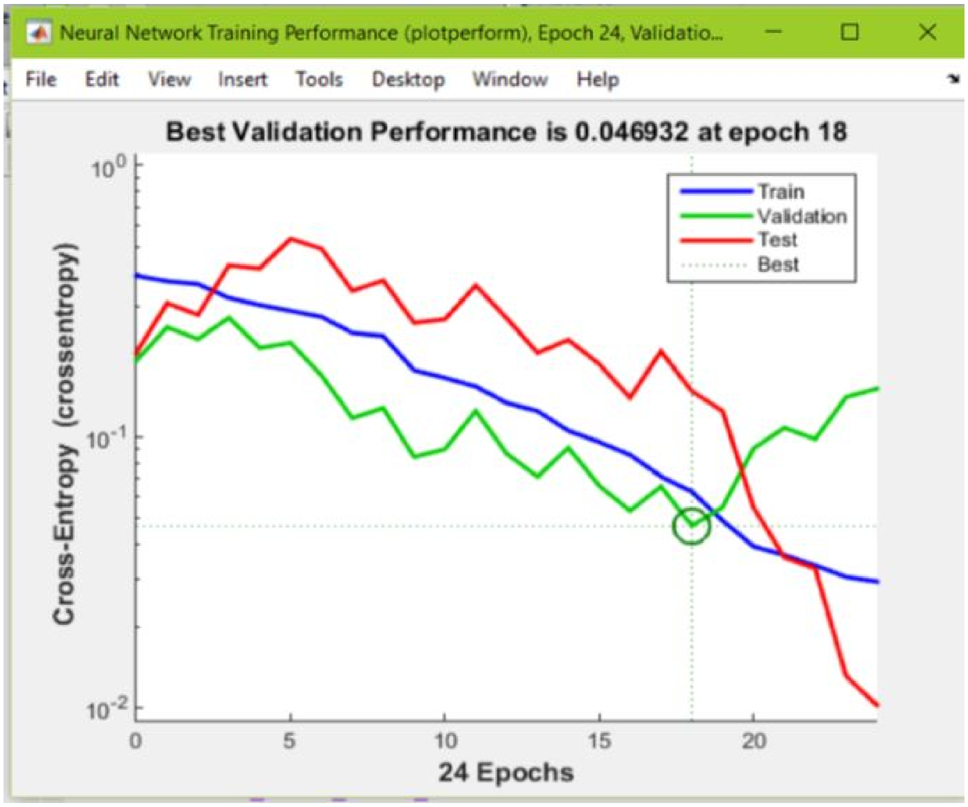

Neural Network Training Performance (plotperform), Epoch 24, Validatio...
File Edit View Insert Tools Desktop Window Help
Best Validation Performance is 0.046932 at epoch 18
Cross-Entropy (crossentropy)
10^0
10^{-1}
10^{-2}
0
5
10
15
20
24 Epochs
Train
Validation
Test
Best

Fig 3.36: MFCC Neural Network Histogram training moony &sunny لام

Table 3.6 Results for Allah _man :(sherief)

Features types	Moufakhum/Accuracy	mouraqeq/Acurecy
MFCC	50%	48%
LPC	40%	45%

Table 3.7 Results for Allah _woman:

Features types	moufakhum/Accuracy	mouraqeq/Acurecy
MFCC	80%	78%
LPC	43%	40%

Table 3.8 Results for Allah _man: Alsedes reader

Features types	moufakhum /Accuracy	mouraqeq/Accuracy
MFCC	95%	94 %
LPC	68%	67%

Table 3.9: Results for A moon _sun Alsedes reader:

Features types	moon/Accuracy	sun/Accuracy
MFCC	99%	98%
LPC	98%	97%

Table 3.10: Results for A moon _sun man:

Features types	moon/Accuracy	sun/Accuracy
MFCC	45.71%	63.33%
LPC	74.28	53.33

Table 3.11 Results for Allah _man: for Ahuzafi reader

Features types	moufakhum/Accuracy	mouraqeq/Accuracy
MFCC	95%	94 %
LPC	94%	67%

Table 3.12 Results for (sunny&moony)لام: for Ahuzafi reader

Features types	moon/Accuracy	sun/Accuracy
MFCC	94%	93%
LPC	93%	92.3%

Note from the above that (MFCC) transactions give the best results than (LPC) transaction with all kinds of readers.in training and testing.

3.8 Hidden Markov Model

3.8.1 Introduction

As mentioned in the introduction part the technique used to implement the speech recognition system was the Hidden Markov Model, HMM. The technique is used to train a model which in our case should represent a utterance of Allah (moufakhum,mouraqeq) and (moony ,sunny لام). This model is used later on in the testing of a utterance (moufakhum,mouraqeq) and (moony ,sunny). لام and calculating the probability of that the model has created the sequence of vectors ,

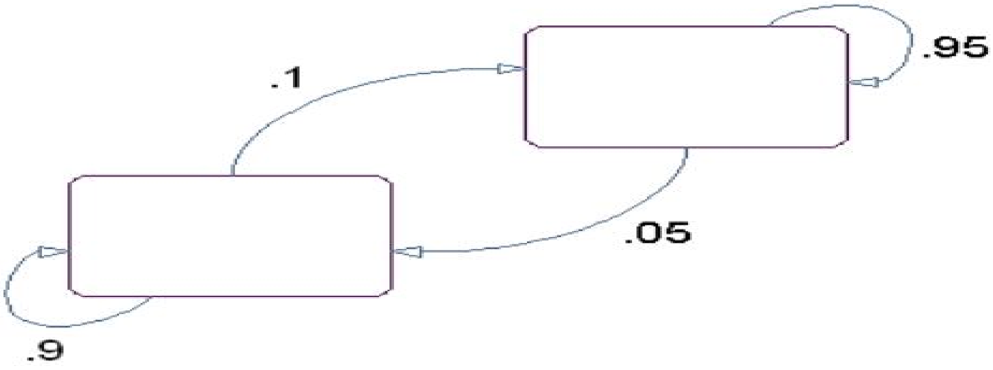
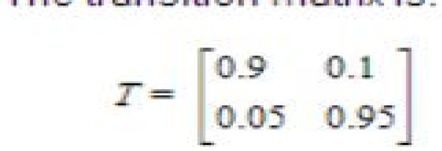


Fig 3.37:Markov Model with two state



The transition matrix is:

$$T = \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix}$$

3.8.2 Training & Testing:

Description of any Model depending on Numbers of state, there was 5 states in this Model.

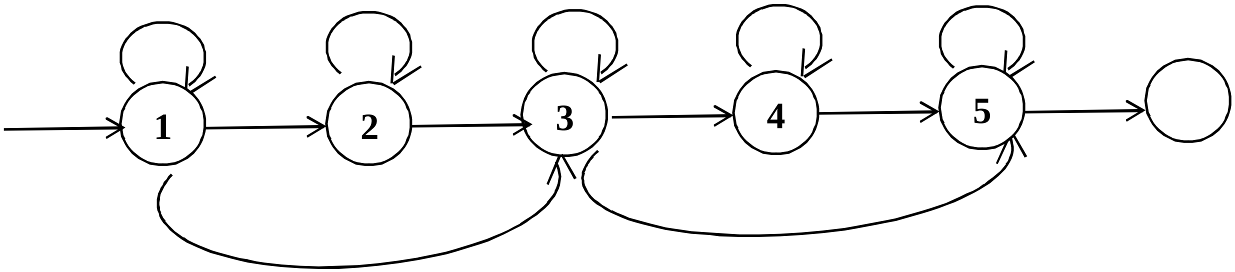


Fig 3.38:Markov Model with Five state

- **Steps of working**
    1. Create primary model
    2. Trained the model
- **This model has two matrices**
    1. Transition matrix (consist of the probability value of moving from state to another state).
    2. Observation matrix (consist of values we want to train the model by).

We entered beams Features (observation matrix) we want Markov model train them, it passes from state to state, and get in output trainer observation matrix resulting from training process which express the probability of the state that we train them .and observation matrix resulting from training process.

Specific value at the resulting of output beams. , the model going from beginning to end and there is no possibility of going back.

- **Like hood**

Within the test we calculate the like hood, we enter beam features, during the test there is an algorithm that calculates the matching of beam with model. We rely on this ratio in the classification input beams.

- **Quantization**

The observation matrix which we train system by must contain integer numbers, and our existing of samples is not limited samples, we must do Quantization process to get a limited number of value, each value has a specific number, We approximate the value of the ray features to the nearest value of corresponding value in the Quantization then we entered it in markov model , we enter the index for each beam .to the model ., new value saved in the codebook which it changes from training data to another .we found that code book consist of 956 samples which was depended on training data , these samples distributed by division between max value and min value each parts was 0.002 .like below

```
num_of_levels=fix((maxx-minn)/0.002);
```

3. Markov models for each of the Allah (moufakhum,mouraqeq) and (moony , sunny) لام , two models for each type of features. For (sheriff,umahmed,umhajer) ., each state must have special markov model ,And the extra data base code process training 4 Markov models for each of the Allah (moufakhum,mouraqeq) and (moony , sunny) لام , two models for each type of features, for (Alsudes ,Alhuzafi) , in every time extraction results depending on a specific number of cases within the Markov model and depending on the specific number of quantization levels., In our model does training code conducting training operations and successive test each time it is training and testing in accordance with a specified number of quantization levels so as to 8 different numbers of cases of model Markov (3, 4, 5, 6, 7, 8, 9 and 10).

Every time we were given the value of the number of quantization levels dependent with the codebook_gen:

3.8.3 Summary

We created Markov model with five states, it was going from beginning to end and there is no possibility of going back .I generate a random sample of the first five cases with the transaction matrix ,then entered observation matrix which special for samples then entered the observation matrix for the samples that we want to trained them with Markov model, Get transaction matrix resulting by the training process , and That reflect the transition probabilities and stayed with concerning to the cases that we train them , and observation matrix resulting by the training process , Within the testing I calculated what is called Like hood , we enterd beam features to it , There was an algorithm to calculated the extent of matching beam with our model Features founded , I rely on this ratio in the classification of rays income .Through the testing process .

- **Viterbi decoding**

Expressed track for each sample, If we go back to the trained results, To describe how stability tracks,. Samples of track eg

123344445555

- **hmm_quantize_trainingdata_MFCC**

Consist of 4 markov model :

HMM FOR allah man larg
quantized_data =12*20
HMM FOR allah woman larg
quantized_data2=12*40
HMM FOR allah man Smal
quantized_data3=12*20
HMM FOR allah woman Smal
quantized_data4=12*40

- **No of iteration was 10**

```
iteration 1, loglik = -3402.907057
iteration 2, loglik = -868.713074
iteration 3, loglik = -785.948500
iteration 4, loglik = -727.229173
iteration 5, loglik = -679.781066
iteration 6, loglik = -648.676549
iteration 7, loglik = -642.455890
iteration 8, loglik = -641.127931
iteration 9, loglik = -640.240457
iteration 10, loglik = -639.596872
```

No of state equal to no of samples in the ray feature , in each track there was 12 state .

- **hmm_quantize_testingdata_MFCC**

all_data=12*120

hmm_quantize_trainingdata_LPC

Consist of 4 markov model :

HMM FOR allah man larg
quantized_data =12*20
HMM FOR allah woman larg
quantized_data2=12*40
HMM FOR allah man Smal
quantized_data3=12*20
HMM FOR allah woman Smal
quantized_data4=12*40

- **No of iteration was 10**

```
iteration 1, loglik = -3402.907057
iteration 2, loglik = -868.713074
iteration 3, loglik = -785.948500
iteration 4, loglik = -727.229173
iteration 5, loglik = -679.781066
iteration 6, loglik = -648.676549
iteration 7, loglik = -642.455890
iteration 8, loglik = -641.127931
iteration 9, loglik = -640.240457
iteration 10, loglik = -639.596872
```

- **hmm_quantize_testingdata_LPC**

all_data=12*120

- **Results tables with HMM :**

Table 3.13Results for Allah _man(sherief):

Features types	moufakhum/Accuracy	mourqeq/Accuracy
MFCC	18.84%	81.16%
LPC	33.73%	66.26%

Table 3.14 Results for Allah_woman:

Features types	moufakhum/Accuracy	mourqeq
MFCC	28.20%	71.6%
LPC	42.42%	57.58%

Table 3.15: Results for A moon _sun man (moony&sunny)لأم:

Features types	moon/Accuracy	sun/Accuracy
MFCC	60 %	83.33%
LPC	68.5%	80%

Table 3.16 Results for Allah _man: for Alsudes reader:

Features types	moufakhum/Accuracy	mourqeq/Accuracy
MFCC	95%	94%
LPC	90%	92%

Results for Table 3.17Allah _man: for Alhuzafi reader

Features types	moufakhum/Accuracy	mourqeq/Accuracy
MFCC	98%	97%
LPC	95%	90%

Results for Table 3.18: A moon _sun man (moony&sunny)لام man:

Features types	moon/Accuracy	sun/Accuracy
MFCC	90%	60%
LPC	85.3%	50%S

- **Sudis_allah_Large_training**

98 samples Allah _Large and Allah_Small 80% for training and 20% for testing

170 samples A_moon and A_ sun 80% for training and 20% for testing

- **No of iteration was 20**

```
>> hmm_quantize_testingdata_MFCC
%%%%% accuracy for MFCC %%%%%
Accuracy for large = 95
Accuracy for small = 94
```

- **Then run code hmm_quantize_trainingdata**:

This is where the code process training and testing, and the results show for eight different values for the number of cases in the Markov model depending on the number of quantization levels that have been identified in the codebook_gen:

- **Test Code Search Allah(Moufakhum,mouraqeq)**

This code we used it in classification level, is used in an audio clip to search for Allah (moufakhum,mouraqeq) , after running it we must determine the file place , it will determines the no of samples contains and numbers of transaction for MFCC .

CHAPTER 4

4. CONCLUSION

4.1 Conclusion

The conclusion of this thesis degree project is that the theory for classification for 4 hukoms of tajweed (Allah ‘moufakhum , mouraqeq’,’moony , sunny’ لام) by Mel Frequency Cepstrum Coefficient(MFFC) and Linear Predictive Coding (LPC) from a waveform signal of specialists readers of Quran, training Hidden Markov Models , Neural Network and testing utterances against the trained models has been successfully implemented in the MatLab environment Read more on further work. The recognition rate is 80% in MatLab for 4 hukoms by using Mel Frequency Cepstrum Coefficient (MFFC) in training and testing Markov Models , Neural Network and 85% for Allah(moufakhum) , (moony) لام and less aquracey which it was mentioned above for the other techniques Linear Predictive Coding (LPC) in training and testing Markov Models , Neural Network, The environments It affected the accuracy of samples So Committees Sampling in an environment more sophisticated where noise free.

4.2 Future Work

There is several future suggestions recommend apply:

Use other methods of analyzing the audio signal such as (FFT), (LFCC) for the purpose of extracting the qualities rather than (MFCC) and (LPC)

Use the another classifier. Neural Network (NN) and hidden Markov Model (HMM)

We recommend following up work in this system to includes all the terms of rules of tajweed in a Holy Quran.

ARABIC REFERENCES

[1] دكتور كمال بشر ، 2000 ،“ علم الأصوات " ، دار غريب للطباعة والنشر والتوزيع ، القاهرة.

[2] يحي محمد الحاج ، منصور الغامدي ، محمد الكنهل ، عبد الله الأنصاري ، 2010 ،“ التعرف الآلى على الأصوات القرآنية.

[3]الدكتور أحمد مختار عمر ، 1997 م 1418 – هـ ،" دراسة الصوت اللغوى " عالم الكتب ، القاهرة ، ص51-111.

[4]الدكتور سليمان حسين جوير ، 1444 هـ ،" الإنسجام الصوتى " مجلة كلية المعارف الجامعية ،العدد 17 ، ص337-340.

ENGLISH REFERENCES

[5] Thomas F. Quatieri,NChandrasekaran,Ting-Peng Liang, “ Discrete Time Speech Signal Processing Principles and Practice”, Pearson Education, Academic Texts, 2003.

[6] Lawrence Rabiner, “Fudementals of Speech Recognition”, Inc Asimon & Schuster Company, 1993, PP 96–102.

[7] Bidoor Noori Ishaq, Bharti W. Gawali, “Comparative Analysis of MFCC, DTW&ANN for Arabic Speech Recognition”, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, VOL 1, Issue 11, NOVEMBER 2014, PP 57.

[8] Anu L B , Dr Suresh D , Sanjeev kubakaddi, “ Person Identification using MFCC and Vector Quantization”, IPASJ International Journal of Electronics & Communication (IIJEC) Volume 3, Issue 6, June 2015, PP 20.

[9] Joe Tebelskis, “Speech Recognition using Neural Networks” , Carnegie Mellon University , Pittsburgh, Pennsylvania 15213-3890, MAY 1995, PP 16–30

[10] Dr.Raj Reddy, “Spoken Language Processing”, PTR Prentice = Hall, 211, PP 383–385.

[11] Dr.S.P.Victor, C.RajKumar, “Modular Implementation of Neural network Structures in Marketing Domain Using Data Mining Technique”, International Journal of Engineering and Computer Science, ISSN: 2319-7242 VOL 5 ISSUE 1, JANUARY 2016, PP. 15624–15630.

[12] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, “Neural Networks used for Speech Recognition”, Journal of Automatic Control, University of Belgrade, VOL. 20:1-7, 2010.

[13] Matthew N. O. Sadiku,Warsame H. Ali, “Signals and Systems” ,CRC Press, 2016 , PP 222.

[14] K.R.Rao,D.N.Kim,J.J.H, “Fast Fourier transforms”, CRC Press, PP 1.

[15] David Houcque, “Introduction to Matlab for Engineering”, Northwestern University,version 1.2, August 2005, PP 1.

[16] Jamaliah Ibrahim Noor , Yamani Idna Idris Mohd , Razak Zaidi , Naemah Abdul Rahman Noor , “Automated Tajweed Checking Rules Engine for Quranic Learning”, Multicultural Education & Technology, Vol. 7 Iss: 4, 2013, pp.275 – 287.

[17] http://www.almaaref.org/books/contentsimages/books/almaaref_alislameya/tajweed_alquran/page/lesson8.htm

[18] Rahmi Yuwan,Dessi Puji Lestari, “Automatic Extraction Phonetically Rich and Balanced Verses for Speaker – Dependent Quranic Speech Recognition System”, School of Electrical Engineering and Information’s Institute Technology ,2014.

[19] Ahsiah Ismail , Mohd Yamani Idna Idris , Noorzaily Mohamed Noor , Zaidi Razak, ZulkifliMohd Yusoff , “ MFCC-VQ Approach for Qalqllah Tajweed Rule Checking”, Malaysian Journal of Computer Science. Vol. 27(4), 2014, pp 275-293.