SPOKEN WORD RECOGNITION USING HIDDEN MARKOV MODEL

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Technology

in

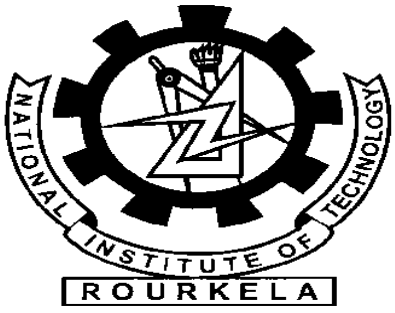
Electronics and Communication Engineering

(Communication and Signal Processing**)**

by

P Ramesh

Roll No:211EC4097



Department of Electronics and Communication Engineering

National Institute of Technology Rourkela-769008

2013

Spoken Word Recognition using Hidden Markov Model

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Technology

in

Electronics and Communication Engineering

(Communcation and Signal Processing)

by

P Ramesh

Roll No:211EC4097

under the guidance of

Dr. Samit Ari

Department of Electronics and Communication Engineering

National Institute of Technology, Rourkela-769008

2013

Declaration

I hereby declare that the work presented in the thesis entitled "*Spoken Word Recognition using Hidden Markov Model"* is a bonafide record of the research work done by me under the supervision of Prof. Samit Ari, Department of Electronics & Communication Engineering, National Institute of Technology, Rourkela, India and that no part thereof has been presented for the award of any other degree.

P Ramesh
Roll No: 211EC4097
Dept. of Electronics & Comm. Engg.
National Institute of Technology
Rourkela, India-769 008

Department of Electronics & Communication Engineering

National Institute of Technology Rourkela

Certificate

This is to certify that the thesis entitled, "**Spoken Word Recognition using Hidden Markov Model**" submitted by Mr. **P RAMESH** in partial fulfillment of the requirements for the award of Master of Technology Degree in Electronics and Communication Engineering with specialization in "**Communication and Signal Processing**" during session 2012-2013 at the National Institute of Technology, Rourkela (Deemed University) is an authentic work carried out by him under my supervision and guidance. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/ Institute for the award of any degree or diploma**.**

Dr. Samit Ari

Dept. of Electronics & Comm. Engg.

National Institute of Technology

Rourkela, India-769 008

ACKNOWLEDGEMENTS

This project is by far the most significant accomplishment in my life and it would be impossible without people who supported me and believed in me.

I would like to extend my gratitude and my sincere thanks to my honorable, esteemed supervisor Dr. **SAMIT ARI**. He is not only a great professor with deep vision but also and most importantly a kind person. I sincerely thank for his exemplary guidance and encouragement. His trust and support inspired me in the most important moments of making right decisions and I am glad to work with him. His moral support when I faced hurdles is un forgettable. My special thank goes to Prof**. S K Meher** Head of the Department of Electronics and Communication Engineering, NIT, Rourkela, for providing us with best facilities in the Department and his timely suggestions.

I want to thank all my teachers Prof**. S.K Patra, K.K Mahapatra, A.K Sahoo, Punam Singh and S.K Behera** for providing a solid background for my studies and research thereafter. I would fail in my duty if I don't think all the lab mates without whom the work would not have progressed. They have been great sources of inspiration to me and I thank them from the bottom of my heart.

I would like to thank all my friends and especially my classmates for all the thoughtful and mind stimulating discussions we had, which prompted us to think beyond the obvious. I have enjoyed their companionship so much during my stay at NIT, Rourkela. I would like to thank all those who made my stay in Rourkela an unforgettable and rewarding experience.

Last but not least I would like to thank my parents, who taught me the value of hard work by their own example. They rendered me enormous support during the whole tenure of my stay in NIT Rourkela.

P RAMESH
(211EC4097)

ABSTRACT

The main aim of this project is to develop isolated spoken word recognition system using Hidden Markov Model (HMM) with a good accuracy at all the possible frequency range of human voice. Here ten different words are recorded by different speakers including male and female and results are compared with different feature extraction methods. Earlier work includes recognition of seven small utterances using HMM with the use only one feature extraction method.

This spoken word recognition system mainly divided into two major blocks. First includes recording data base and feature extraction of recorded signals. Here we use Mel frequency cepstral coefficients, linear cepstral coefficients and fundamental frequency as feature extraction methods. To obtain Mel frequency cepstral coefficients signal should go through the following: pre emphasis, framing, applying window function, Fast Fourier transform, filter bank and then discrete cosine transform, where as a linear frequency cepstral coefficients does not use Mel frequency.

Second part describes HMM used for modeling and recognizing the spoken words. All the raining samples are clustered using K-means algorithm. Gaussian mixture containing mean, variance and weight are modeling parameters. Here Baum Welch algorithm is used for training the samples and re-estimate the parameters. Finally Viterbi algorithm recognizes best sequence that exactly matches for given sequence there is given spoken utterance to be recognized. Here all the simulations are done by the MATLAB tool and Microsoft window 7 operating system.

Contents

LIST OF FIGURE

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Speech recognition can be divided into isolated spoken word recognition, continuous speech recognition, text dependent spoken word recognition,and speaker independent and speaker dependent recognition[1]. In the isolated soken word recognition each word is spoken by different speaker during training, and anytesting signal (same word spoken by any other or same person) should be recognised. The continuous speech recognition can be further divided into connected speech word recognition and conversational speech recognition. The isolated word recognises each word but has a limited number of words (vocabulary) but in the case of continuos speech recognition focuses on understanding the sentences and has a large vocabulary of words. Spoken word recognition can also be speaker-dependant (in which case the acoustic features have to be varied for every time speaker changes) or speaker-independent, in this type recognises spoken word is irrelevant of the speaker. There are huge application for speaker independent recognition and commercial is better than others but has complexity to imolement. The above mentioned complexcity in implementation occurs due unique type of acoustic features for each persons.Therfore in speaker independent systems only isolated spoken word recognition is generally implemented[2].

Here different words are recorded separately with different speakers like each person records utterances like digits “One, Two…….Ten” are trained and during testing any one among above words is spoken by any one of the above speakers and recognition is obtained. Along with the above mentioned data the second data base used were utterances like “apple, pen, hi……move”.

1.2 LITERATURE REVIEW

Speech word recognition area has a wide range of research. Many of the researchers had proposed many different methodologies to speech recognition problems. Most of the method

used for speech recognition is Hidden Markov Models (HMM), Dynamic Time warping (DTW) and many others. Literature review concludes that among all the methodologies used for speech recognition problem, Hidden Markov models are best statistical models that are most accurate in modelling the speech parametersa. Many researchers proposed different methodologies for Hidden Markov Model based speech recognition system. Most of them had used different feature extraction methods to get the speech features like fundamental frequency, energy coefficients and others. Many of the studies shows that use of Mel frequency cepstrum coefficients as feature to speech recognition yield good results at human auditory range of frequencies but not throughout the range of human auditory frequency range. Review on other feature extraction methods like linear prediction coding, say that the extracted features using that method are good for speech processing like speech coding but not speech recognition.

Complete study of a tutorial on Hidden Markov Model and it application in Speech recognition (especially on isolated spoken word recognition) by Lawrence Rabinier where a major contribution to hidden markov model and its application are discussed clearly. Here types of HMM's like discrete and continuous HMM were studied. I have also referred a tool kit of HMM by Microsoft Corporation and Cambridge University Engineering Department where I got many helps while writing a code for the proposed model in MATLAB programming language.

1.3 SCOPE OF THE THESIS

The main aim of this project is identification or recognition of spoken word utterances among the many spoken words trained using Hidden Markov Models and obtaining a good accuracy of recognition with use different methods used for getting speech features like Mel frequency cepstral coefficients (MFCC) and Linear frequency cepstral coefficients (LFCC). To get the

good recognition rate at high end of the human voice auditory frequency we used linear frequency cepstral coefficients rather than Mel frequency cepstral coefficients which yield good results in between frequency range of human voice frequency range. A part from above achievement this proposed model also aims to increase the number of speakers as well as number of spoken utterances from seven to ten. The overall scope of the proposed model is to increase the rate of recognition in any circumstances like increased number of speakers or increased number of utterances and especially to recognise sounds consisting nasal consonants,reveberatoy sounds[3].

1.4 MOTIVATION OF THESIS

Spoken word recognition (isolated spoken word) is a very good application of Hidden Markov Model (HMM) that has many number of real world applications like security systems, telephone networks and automation. Spoken word recognition can be used to automate different works that previously required human physical work involvement, such as recognizing simple spoken commands to perform something like turning on fans or closind and opening of a gates. It can be used for security systems like opening of personal computers using voice and word recognition. To increase recognition rate, techniques such as neural networks, dynamic time warping and hidden Markov models have been used. Recent technological advances have made recognition of more complex speech patterns possible. Despite these breakthroughs, however, current efforts are still far away from 100% recognition of natural human speech. Much more research and development in this area are needed to achieve the speech recognition ability of a human being with the above mentioned recognition rate. Therefore, we proposed this a challenging and worthwhile project that can be rewarding and benefits in many ways, where there is much improvement in the

recognition of isolated words or spoken utterances especially at the high end of human voice frequency range.

1.5 THESIS OUTLINE

The outline of this is as follows

Chapter 2: It describes the basic fundamentals of speech like definition, production mechanism of speech, nature and type of speech signal. It also describes data base used in this project. Representation of speech waveform is mentioned.

Chapter 3: In this chapter feature extraction method for speech signal is described. We mostly concentrate on fundamental frequency also pitch, Mel frequency cepstral coefficients (MFCC) and linear frequency cepstral coefficients. Different types of windowing techniques that can be applied during feature extraction are discussed. Results of different feature extraction methods are compared at different levels of frequencies.

Chapter 4: Here we discuss Hidden Markov Model and basics like Bayes theorem, chain rule and markov process. A clear idea of Hidden Markov Model is discussed with a motivating example. We also concentrate on Baum Welch algorithm, and Viterbi algorithm for recognition. Parameter initialisation and their re-estimation are discussed. Finally best estimated sequence is found using Viterbi algorithm.

Chapter 5: This chapter details the results obtained. Here we compare the different differences in feature extraction methods used. Recognition rate of spoken words are compared for both male female.

Chapter 6: Here we conclude the work in this project and results obtained. Scope for future work is described.

CHAPTER 2

HUMAN VOICE

FUNDAMENTALS

2.1 DEFINING HUMAN VOICE

Human voice is a natural form of communication for human beings. The basic meaningful unit in a spoken signal is the sound. That is, a speech signal contains a sequence of sounds. The expression of or the ability to express thoughts and feelings by articulate sounds is speech. Voice signals are generated by nature. They are naturally occurring, hence they are random signals. There are several models put forth by researchers based on their perception of voice signal. The range frequencies for human voice are discussed. The mechanics behind human voice production are unique and quantifiable.

2.1.1 FREQUENCY RANGE

Generally frequencies in the range of 50Hz and above are generated in human natural voice. The majority of the energy lies in between 350Hz and 3400Hz. The human ear on the other hand, can responds to sounds over the range of frequencies from 200Hz to 2000Hz with most sensitivity in the region between 300Hz to 9KHz. Considering these factors along with functional testing the frequency range of 350 Hz to 3400Hz considered to be the most important for speech intelligibility and speech recognition.

Reducing this (350 Hz to 3400 KHz) frequency band width can predominently reduce the speech intelligibility however increasing it has been fount not significantly improve recognition or the intelligibility. Here note that increased bandwidth will increase over sound quality however incremental gains in sound quality have to be weighed against increased frequency usage.

2.2 HUMAN VOICE PRODUCTION MECHANISM

The human speech production system is shown in Figure1. The various organs responsible for producing speech are labelled. The main source of energy is the lungs with the

diaphragm. When a person speaks the air is forced through the glottis between the vocal cords and the larynx and it passes via three main cavities, namely, vocal tract, the pharynx, and the oral and nasal cavities. To produce speech air is forced by the oral and nasal cavities through the mouth.

Excitation signals are generated in the following ways – phonation, whispering[4], frication, compression, and vibration. Phonation is generated when vocal cords oscillate. These chords can close and stretch because of cartilages. The oscillations depend on the mass and tension of the cords. The opening and closing of the cords break the air stream impulses. The shape and duty cycle of these pulses will depend on loudness and pitch of the signal. The pitch is defined as the repetition rate of these pulses. At low levels of air pressure, the oscillations may become irregular and occasionally the pitch may drop. These irregularities are called vocal fry.

The V shaped opening between the vocal chords, known as glottis, is the most important source in the vocal system[5], generating a periodic excitation because of the natural frequency of vibration of vocal cords may act in several different ways during speech. Their most important frication is to modulate the air flow by rapidly opening and closing, thus causing a buzzing sound and so produce vowels and consonants. The fundamental frequency vibration of the vocal cords which function as a tuning fork depends on the mass and tension of the cords and is about 200Hz and 300 Hz with men, women, and children, respectively.

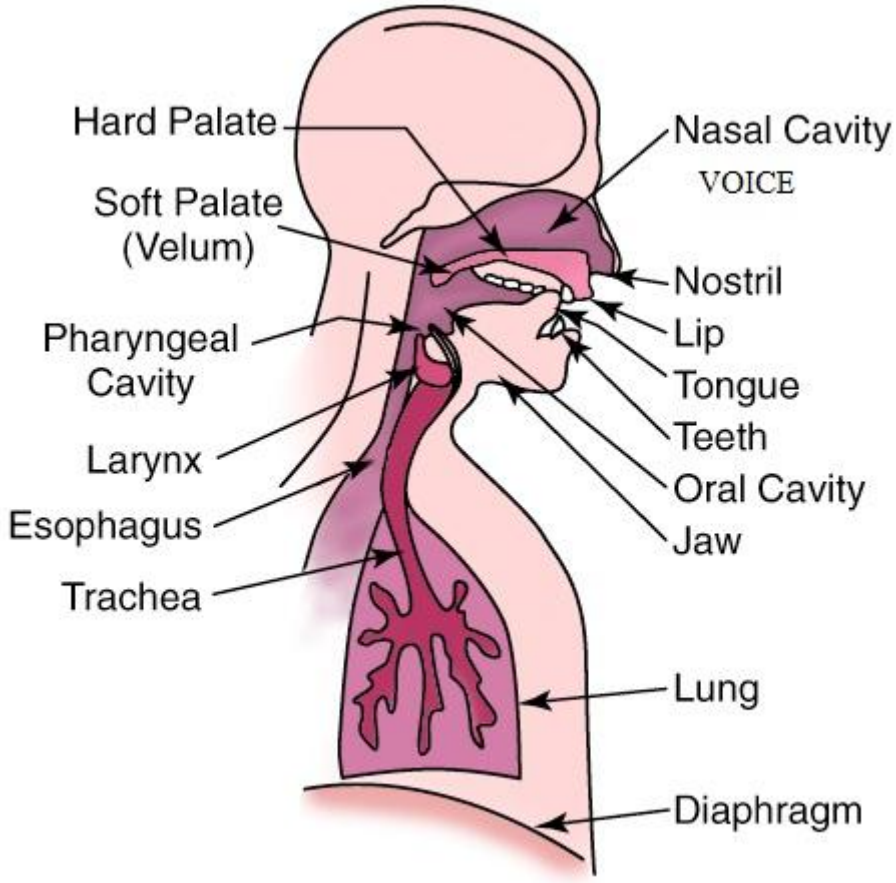


Figure 2.1: The human vocal organs[6]

In the case of whispering, the vocal cords are drawn closer with a very small triangular opening between the cartilages. The air passing through this opening will generate turbulences (wide band noise). The turbulence will work as an excitation for whispering. When the vocal tract is constricted at any other point, the air flow again becomes turbulent, generating wide band noise. It is observed that the frequency spectrum of this wide band noise effects the location of the constriction. The sounds produced with this excitation are called fricatives or sibilants. Frication can occur with or without phonation. If the speaker continues to exhale when the vocal cord responds, a small explosion will occur. A combination of short silence with a short noise burst has a characteristic sound[24]. If the release is abrupt and clean, it is called a stop or a stop or a plosive. If the release is gradual and turbulent, it is termed as an affricate.

The pharynx connects the larynx to the oral cavity. The dimension of the pharynx is fixed but its length changes slightly when larynx is raised or lowered at one end and the soft palate is raised or lowered at the other end. The soft palate also isolate the nasal cavity and the pharynx or forms of the route from the nasal cavity to the pharynx, the epiglottis and false vocal cords are at the bottom of the pharynx to prevent food from reaching the larynx and to isolate the esophagus acoustically from the vocal tract[24].

The oral cavity is one of the most important parts connected to the vocal tract. Its size, shape, and acoustics can be varied by the movements of the palate, the tongue, the lips, the cheeks, and the teeth which are called articulating organs. The tongue in particular is very flexible, allowing the tip and edges to be moved independently. The complete tongue can also be moved forward, backward, up, and down. The lips can control the size and shape of the mouth opening when speech sound is radiated. Unlike oral cavity the nasal cavity has fixed dimension and shape. Its length is about 12 cm and volume is 60 cm^3. The air stream to the nasal cavity is controlled by the soft palate.

2.2.1 LTI MODEL FOR SPEECH PRODUCTION

The basic assumption of the speech processing system source of excitation and the vocal tract system are independent. The assumption of independence allows us to discuss the transmission function of the vocal tract system separately. We can now allow the vocal tract system to get excited from any of the possible source. Based on this assumption we can put forth the digital model for speech production.

A speech is basically convolved signal. It is generated when an excitation signal gets generated at the sound box. The vocal tract is generally modelled as a filter that may be all pole filters or a pole-zero filters. The excitation signal gets convolved with the impulse response of the vocal tract and this convolved signal is the speech signal.

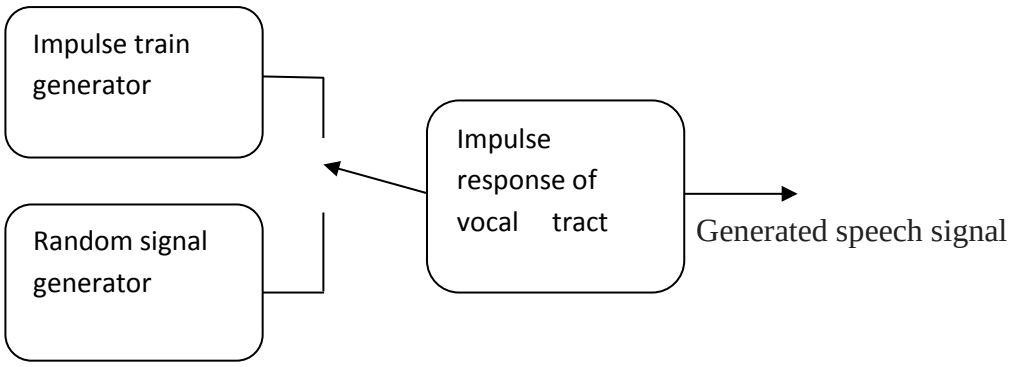


Figure 2.2 Speech production using LTI model

The impulse train generator is used as an excitation when a voiced segment is produced. The unvoiced segments are generated when the random signal generator is used as the excitation.

2.3 NATURE OF SPEECH SIGNAL

There can be as many as hundreds of speech parameters describing a speech signal. The models put forth by different researchers in the field are built on the dominant features extracted by the researcher.

2.3.1 PHONETICS

Any language can be described in terms of a set of distinctive sounds called **phonemes.** For English there are 42 phonemes which include vowels, diphthongs, semivowels and consonants. Phonetics is a branch of linguistics that involves the study of human speech. It deals with the physical properties of speech sounds[7] namely, phones, their physiological production, acoustic properties, and auditory perception. The study of phonetics is a subject of linguistics that focuses on speech. In this field of research there are three basic areas of study, namely, articulatory phonetics, acoustic phonetics, and auditory phonetics[8].

2.3.2 ARTICULATORY PHONETICS

The study of the production of speech by the articulatory and vocal tract by the speaker. The main goal of articulatory phonetics is to provide a common notation and a frame of reference for linguistics. This allows accurate reproduction of any unknown utterance which is written in the form of phonetics transcriptions.

Consonants are defined in anatomical terms using the point of articulation, manner of articulation and phonation. The point of articulation is the location of principal constructions in the vocal tract defined in the terms of the participating organs. The manner of articulation is principally a degree of constriction at the point of articulation and the manner of release. Plosives have a clean and sharp of plosives, they are also called stops. If the consonants are accompanied by voicing they are called **voice consonants.**

Vowels are well defined than consonants. This is because the tongue never touches another organ when making a vowel. So we can specify parameters like the point of articulation. In the case of vowels we specify only the distance from the mouth. Vowels are described by the following variables:

1. Tongue high or low
2. Tongue front or back
3. Lips rounded or unrounded
4. Nasalized or un –nasalized

The vowels such as a, e, i, o and u are called tense vowels. These vowels are associated with more extreme positions on the vowel diagram, as they particularly require greater muscular tension to produce.

2.3.3 ACOUSTIC PHONETICS

Acoustically, the vocal tract is a tube of non-uniform cross section approximately 16 cm long, which is the usually opened at one end close at other end. The vocal tract tube has many natural frequencies. Natural frequencies can occur at the odd multiples of 500Hz. These resonant frequencies are **formants.** These appear as dark bands in spectrograms and they are considered to be very important acoustical features.

Acoustic phonetics is a subfield of phonetics which deals with the acoustic aspects of speech sound. Acoustic phonetics investigates properties of speech sound, such as mean squared amplitude of a speech waveform, its duration and its fundamental frequency. By exposing a segment of speech to phonograph, and filtering it with a different band pass filter time, a spectrogram of the speech utterance could be generated.

2.3.4 AUDITORY PHONETICS

Auditory phonetics is related to the perception and interpretation of sounds. The transmitter and receiver are involved in the process of linguistic communication. In the case of auditory phonetics receiver is the listener, that is, human ear.

We have to discuss three components of ear, namely, the outer, the middle, and the inner ear. The outer ear consists of the auricle or the pinna, and the auditory meatus or the outer ear canal.

2.4 TYPES OF SPEECH

The speech signal is divided into two parts, a voiced segment and an unvoiced segment. A waveform is two dimensional representation of a sound. The two dimensions in waveform display are time and intensity. In case of a sound waveform, the vertical dimension is

intensity and the horizontal dimension is time. Waveforms are also called time domain representation of sound, as they are representation of changes in intensity over time.

2.4.1 VOWELS AND VOICED SEGMENTS

Vowels are components of sound, that is, /a/, /e/, /i/, /o/, /u/. The excitation is the periodic excitation generated by the fundamental frequency of the vocal cords and the sound gets modulated when it passes via the vocal tract

2.4.2 DIPHTHONG

It means two sounds or two tones. A diphthong is also known as gliding vowel, meaning that there are two adjacent vowel sounds occurring within the same syllable. A diphthong is a gliding mono syllabic speech item starting at or near the articulatory position for one vowel and moving towards the position of the other vowel. While pronouncing a diphthong, that is in the case of words like eye, hay, bay, low, and cow the tongue moves and these are said contain diphthongs. There are six diphthongs in English. Diphthongs can be characterised by time varying vocal tract area function.

2.4.3 SEMI VOWEL

A semi vowel is a sound, such as /w or /j/ in English, which is phonetically similar to a vowel sound but function often as the syllable boundary. These are called semivowels due to their vowel like structure. They occur between adjacent phonemes and are recognized by a transition in the functioning of the vocal tract area.

2.4.4 UNVOICED SOUNDS

With the unvoiced sounds, there is no fundamental frequency in the excitation signal and hence we consider the excitation as white noise. The air flow is forced through a vocal tract constriction occurring at several places between the glottis and the mouth.

2.4.5 CONSONANTS

A consonant is a speech segment that is articulated using a partial or complete closure of vocal tract. The examples of consonants are as follows /p/, /t/, /k/, /h/, /f/ etc. Consonants can be further classified as nasals, stops – either voiced or unvoiced, fricatives (again voiced or unvoiced), whisper and affricatives.

2.5 DATABASE

Data base consists of twelve different spoken words like “One, Two, Three, Four ……..Ten” or else “Apple, Ball, Cat…….Zebra” each having time in milli seconds. They are recorded using audacity software, where it is possible set sampling frequency and can be viewed spectrally. These utterances are spoken by six different speakers, among which three are male and another three are female. These utterances are represented in ‘.wav’ format. The sampling frequency of these recorded samples is taken as 16000 Hz. The representation in ‘.wav’ format makes us easy in open the samples in MATLAB programming language.

Table 2.1: Representation of data base for first set

Male	Speker1	training	11tr .wav	12tr .wav	13tr .wav	14tr .wav	15tr .wav	16tr .wav	17tr .wav	18tr .wav	19tr .wav
		testing	11te .wav	12te .wav	13te .wav	14te .wav	15te .wav	16te .wav	17te .wav	18te .wav	19te .wav
	Speker2	training	21tr .wav	22tr .wav	23tr .wav	24tr .wav	25tr .wav	26tr .wav	27tr .wav	28tr .wav	29tr .wav
		testing	21te .wav	22te .wav	23te .wav	24te .wav	25te .wav	26te .wav	27te .wav	28te .wav	29te .wav
	Speker3	training	31tr .wav	32te .wav	33tr .wav	34tr .wav	35tr .wav	36tr .wav	37tr .wav	38tr .wav	39tr .wav
		testing	31te .wav	32te .wav	33te .wav	34te .wav	35te Wav	36te .wav	37te .wav	38te .wav	39te .wav
Female	Speker4	training	41tr .wav	42tr .wav	43tr .wav	44tr .wav	45tr .wav	46tr .wav	47tr .wav	48tr .wav	49tr .wav
		testing	41te .wav	42te .wav	43te .wav	44te .wav	45te .wav	46te .wav	47te .wav	48te .wav	49te .wav
	Speker5	training	51tr .wav	52tr .wav	53tr .wav	54tr .wav	55tr .wav	56tr .wav	57tr .wav	58tr .wav	59tr .wav
		testing	51te .wav	52te .wav	53te .wav	54te .wav	55te .wav	56te .wav	57te .wav	58te .wav	59te .wav
	Speker6	training	61tr .wav	62tr .wav	63tr .wav	64tr .wav	65tr .wav	66tr .wav	67tr .wav	68tr .wav	69tr .wav
		testing	61te .wav	62te .wav	63te .wav	64te .wav	65te .wav	66te .wav	67te .wav	68te .wav	69te .wav

Each speaker records each word two times among them one is used for training and another for testing.

The sample of spoken word is shown below with a sampling rate of 16000Hz.

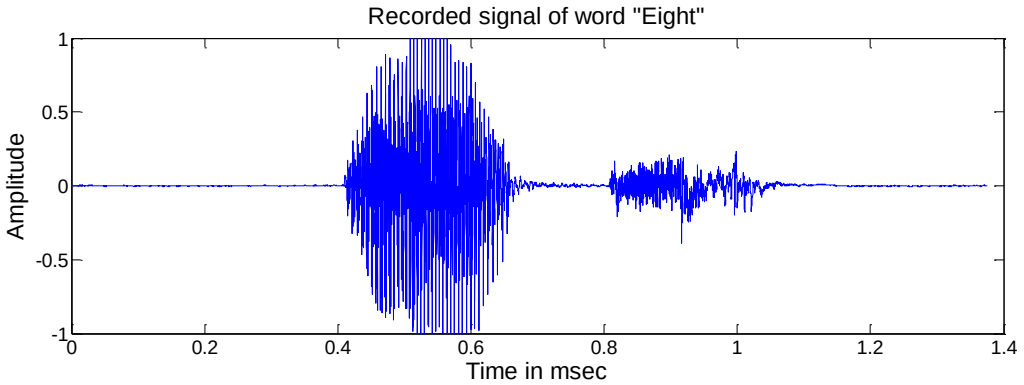


Figure 2.3: Wave form representation of Spoken signal “Eight”

The above shown figure is recorded in personal computer using micro phone.

Table 2.2: Representation of second data set

			One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Male	Speker1	training	11tr .wav	12tr .wav	13tr .wav	14tr .wav	15tr .wav	16tr .wav	17tr .wav	18tr .wav	19tr .wav
		testing	11te .wav	12te .wav	13te .wav	14te .wav	15te .wav	16te .wav	17te .wav	18te .wav	19te .wav
	Speker2	training	21tr .wav	22tr .wav	23tr .wav	24tr .wav	25tr .wav	26tr .wav	27tr .wav	28tr .wav	29tr .wav
		testing	21te .wav	22te .wav	23te .wav	24te .wav	25te .wav	26te .wav	27te .wav	28te .wav	29te .wav
	Speker3	training	31tr .wav	32te .wav	33tr .wav	34tr .wav	35tr .wav	36tr .wav	37tr .wav	38tr .wav	39tr .wav
		testing	31te .wav	32te .wav	33te .wav	34te .wav	35te Wav	36te .wav	37te .wav	38te .wav	39te .wav
Female	Speker4	training	41tr .wav	42tr .wav	43tr .wav	44tr .wav	45tr .wav	46tr .wav	47tr .wav	48tr .wav	49tr .wav
		testing	41te .wav	42te .wav	43te .wav	44te .wav	45te .wav	46te .wav	47te .wav	48te .wav	49te .wav
	Speker5	training	51tr .wav	52tr .wav	53tr .wav	54tr .wav	55tr .wav	56tr .wav	57tr .wav	58tr .wav	59tr .wav
		testing	51te .wav	52te .wav	53te .wav	54te .wav	55te .wav	56te .wav	57te .wav	58te .wav	59te .wav
	Speker6	training	61tr .wav	62tr .wav	63tr .wav	64tr .wav	65tr .wav	66tr .wav	67tr .wav	68tr .wav	69tr .wav
		testing	61te .wav	62te .wav	63te .wav	64te .wav	65te .wav	66te .wav	67te .wav	68te .wav	69te .wav

So totally there are 144 spoken utterances. Out of which halh are used for training and remaining for testing.

CHAPTER 3

FEATURE EXTRACTION

3.1 INTRODUCTION

This chapter will introduce the significance, meaning and methods of extraction of two important speech parameters, namely, pitch frequency and cepstral coefficient. For speech recognition, speaker verification, speech synthesis, etc. One must extract the features of the speech segment, such as fundamental frequency, formants, linear predictive coefficient (LPC), Mel frequency cepstral coefficient (MFCC), cepstral coefficients, line spectral pairs, 2-D and 3-D spectrogram, etc[9]. There are some time domain features and frequency domain features, such as frequency domain, cepstral domain, wavelet domain, discrete cosine transform (DCT) domain and so on. We will mainly discuss fundamental frequency and cepstral coefficients for speech recognition.

3.2 FUNDAMENTAL FREQUENCY

A speech signal consists of different frequencies which are harmonically related to each other in the form a series. The lowest frequency of this harmonic series is known as the fundamental frequency or pitch frequency. Pitch frequency is the fundamental frequency of the vocal cords. There are many different techniques available to get the fundamental frequency (f_0) like auto correlation method and FFT based extraction. The following steps are followed during pitch extraction.

1. Take the spoken word or utterance
2. Take FFT of the above signal with certain point number.
3. Track the first peak in the FFT output to find the fundamental frequency. Frequency resolution is decided by FFT point number
4. The fundamental frequency is the FFT point number (where first peak occurs) multiplied by the frequency resolution.

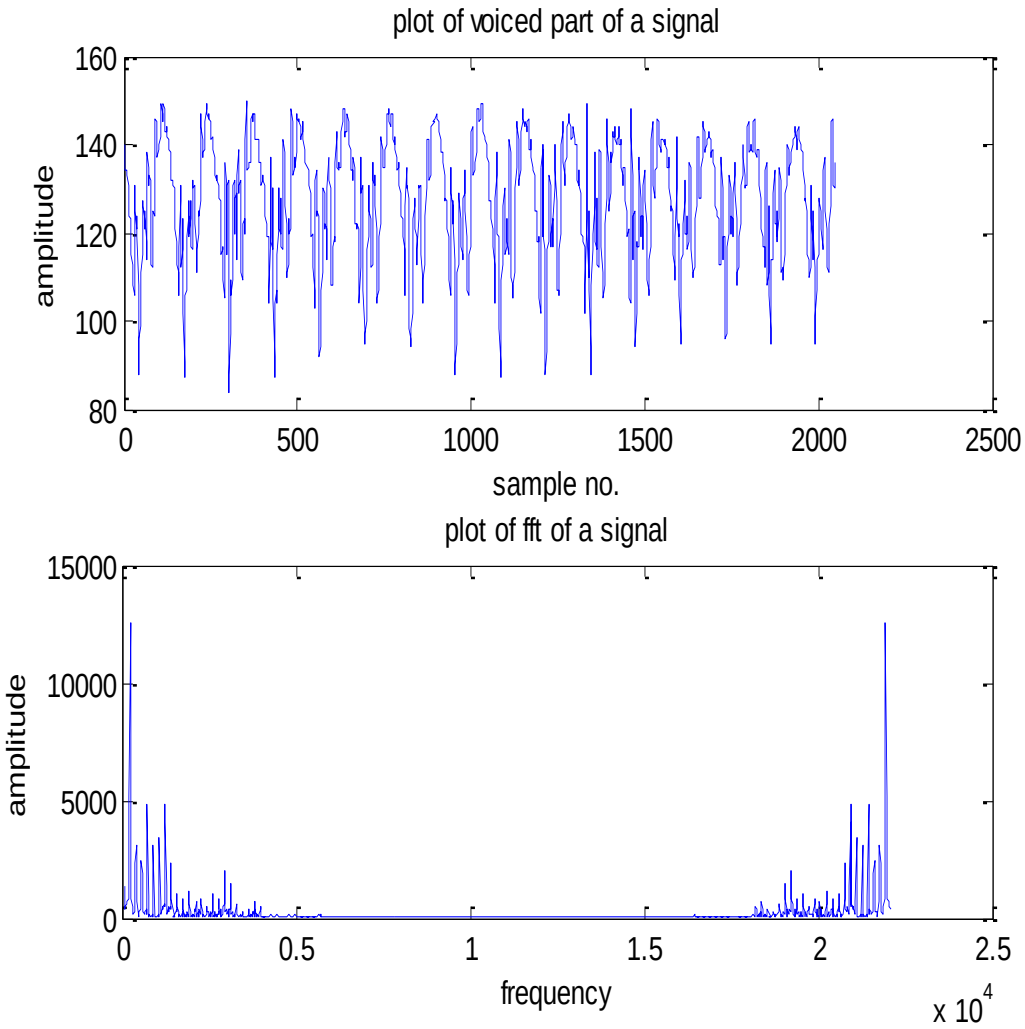


Figure 3.1: Plot of voiced signal and its fundamental frequency

The signal plot and the output plot of FFT indicates that the first peak has a value of 12560 and it occurs 16th position, and hence the fundamental frequency is 10.79*16 = 172.26 Hz where resolution is 22,100/2048 = 10.79 Hz.

3.3 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCCS)

Speech feature extraction is a fundamental requirement of any speech recognition system. It is the mathematic representation of the speech file. In a human speech recognition

system, the goal is to classify the source files using a reliable representation that reflects the difference between utterances.

CEPSTRUM

The name Cepstrum was derived from the spectrum by reversing the first four letters. We can say cepstrum is the Fourier Transformer of the log with unwrapped phase of the Fourier Transformer. Mathematically we can say Cepstrum of signal = FT (log (FT (the signal)) +j2IIm), where m is the integer required to properly unwrap the angle or imaginary part of the complex log function.

MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) :

MFCC's are used in many different areas of speech processing, speech recognition and speech synthesis. They are very similar in principle with the human ear perception, and are especially good for speech recognition and speech synthesis. The following block diagram shows the MFCC implementation[10]. It consists of six block. The first step enhances the spoken word signal at high frequencies. Then framing facilitates use of FFT.

FFT is performed to get energy distribution over frequency domain. Before applying FFT we will multiply each frame with window function, to keep the continuity between first and last point. N set of triangular bandpass filters are multiplied to calculate energy in each band pass filter.Here we use mel frequency triangular band pass filters. Then Discrete Cosinr Yransform is applied to get the MFCC. They represents the accostic features of speech.

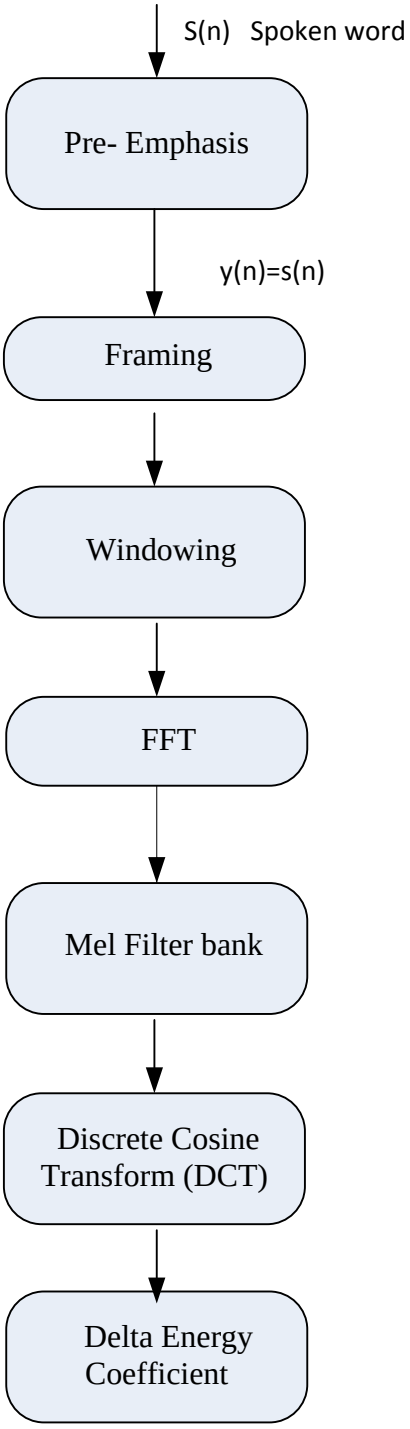


Figure 3.2 Block diagram of MFCC

The MFCCs are often used for speech recognition, and essentially mimic the functionality the human ear.

3.3.1 PRE-EMPHASIS

It is the fundamental signal processing applied before extracting features from speech signal, for the purpose of enhancing the accomplishment of feature extraction algorithms. Here we use pre-emphasis to boost the high frequencies of a speech signal which are lost during speech production[11].

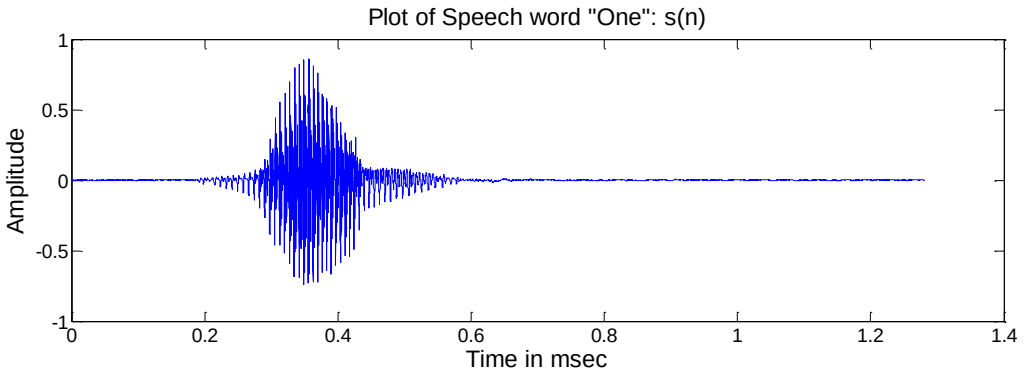


Figure 3.3: Plot of spoken utterance before pre-emphasis

$$Y(n) = X(n) - k*X(n-1) \qquad 3.1$$

Where, k is between 0.9 and 1.

X (n) is the speech signal before pre-emphasis

Y (n) is the speech signal after pre-emphasis

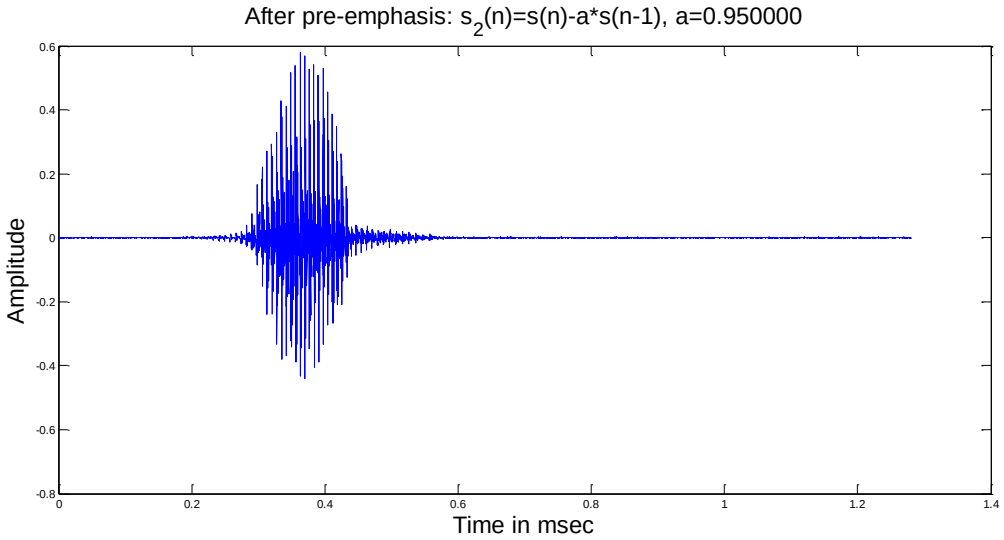


Figure 3.4: Signal after pre-emphasis

Here this step acts like high pass filter,where high frequency content of spoken signal is enhanced.

3.3.2 FRAMING

The spoken word signal is then divide into frames of N samples, with adjacent frames being separated. The starting frame consists of the first N samples of spoken word signal. The second frame begins M samples after the first frame and overlaps it bycertain choosen amount of samples. Here we take each frame consists of 256 samples of speech signal, and the subsequent frame starts from the 100th sample of the previous frame. Thus, each frame overlaps with two other sub-sequent frames. This method is called framing of signal. The spoken word sample in one frame is considered to be stationary.

Here the frame length selection is a important for spectral analysis, due to the trade-off between the time and frequency resolutions. The window should be fair enough for adequate frequency resolution, but on the other hand, it should be short enough so that it would capture the local spectral properties of spoken signals.

3.3.3 WINDOWING

In between the extremes are moderate windows, such Hamming and Hanning [12]. They are generally used in narrowband applications, such as the spectrum of a telephone channel. In summary, spectral analysis involves a trade-off between resolving comparable strength components with similar frequencies and resolving disparate strength components with dissimilar frequencies. That trade-off occurs when the window function ismaintained.

The effects of windowing on the Fourier coefficients of the filter on the result of the frequency response of the filter are as follows:

(i) A major effect is that to keep the continuity between first and end points of each frame divided during frame.

(ii) The transition band width depends on the main lobe width of the frequency response of the window function used.

(iii) As the frequency response of filter is derived through a convolution, it is clear that the resulting filters will not be optimal.

(iv) As the length of the window function rises, the main lobe width is decreased which decreses the width of the transition band, but this also generates more ripple in the frequency response of a signal.

(v) The window function eliminates the ringing effects at the band edge and does result in lower side lobes for an increase in the width of the transition band of the filter.

3.3.5 GENERALIZED HAMMING WINDOWS

Generalized Hamming windows[6] are of the form given below

$$w(n) = \alpha - \beta cos\left(\frac{2\pi n}{N-1}\right) \qquad 3.2$$

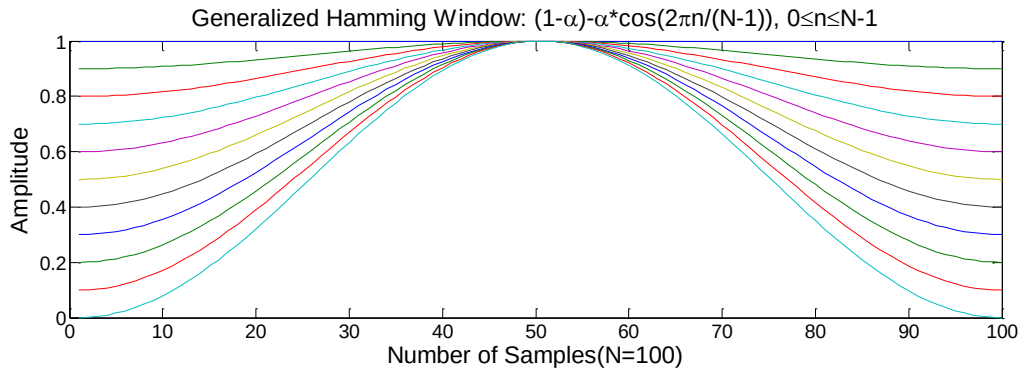


Figure 3.5: Generalised Hamming Windows funtion

3.3.6 HANN (HANNING) WINDOW

The Hann window named after Julius von Hann and also known as the Hanning (for being similar in name and form to the Hamming window), von Hann and the raised cosine window [13] is defined by

$$w(n) = 0.5\left(1 - cos\left(\frac{2\pi n}{N-1}\right)\right) \qquad 3.3$$

3.3.7 HAMMING WINDOW

The window with these particular coefficients was proposed by Richard W. Hamming. The window is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window. Hamming window is a modified version of the Hanning window. The shape of the Hamming window is similar to that of a cosine wave. The following equation defines the

$$w(n) = \alpha - \beta cos\left(\frac{2\pi n}{N-1}\right) \qquad 3.4$$

with $\alpha = 0.54$ $\beta = 1 - \alpha = 0.46$ for n = 0, 1, 2 …, N – 1

where N is the length of the window and w is the window value.

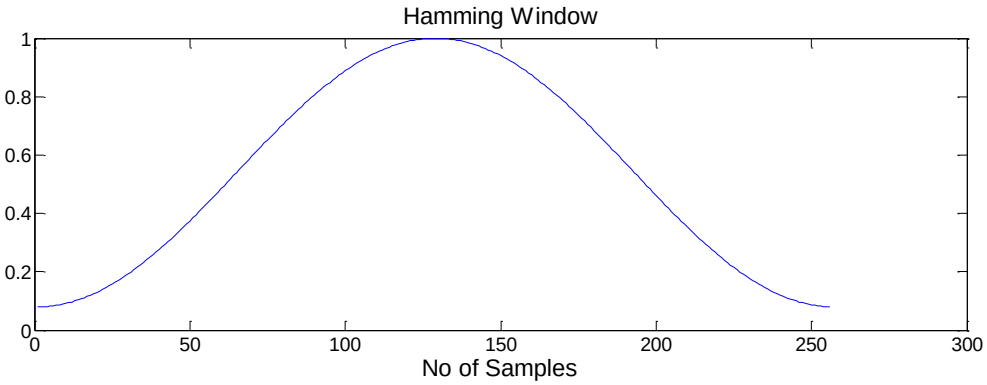


Figure 3.6: Hamming window

Approximation of the constants to two decimal places substantially lowers the level of side lobes, to a nearly equiripple condition. In the equiripple sense, the optimal values for the coefficients are α = 0.53836 and β = 0.46164.

3.4 FOURIER TRANSFORM

The Fourier transform is an operation that transforms one complex-valued function of a real variable into another domain. In such applications as signal processing and speech processing, the signal before applyinf FFT will be time domain. However FFT converts time domain signal to frequency domain.It describes which frequencies are present in the original function.

For a continuous function of one variable f(t), the Fourier Transform F(f) will be defined as:

$$F(f) = \int_{-\infty}^{\infty} f(t) e^{-j2\pi ft} dt \qquad 3.5$$

and the inverse transform as

$$f(t) = \int_{-\infty}^{\infty} F(f) e^{+j2\pi ft} df \qquad 3.6$$

where j is the square root of -1 and e denotes the natural exponent.

3.5 TRIANGULAR BANDPASS FILTERS

We multiple the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation

$$\text{mel}(f) = 1125 * \ln(1 + f/700) \qquad 3.$$

3.5.1 MEL FREQUENCY WARPING

In general human ear hears the frequencies, non linearly. Here the scaling of frequecncies linear up to 1KHz, the we usu logarithmic scale to represent. Human auditory system is scaled by frequency scale called Mel scale frequency [14]. Here Mel reprents melody. Mostly they are used as a band pass filtering for this stage of recognition. The spoken signals for each frame is allowed through Mel-Scale frequency band pass filter to mimic the human auditory perception. Thus for each tone with an actual frequency *f*, measured in Hz, a subjective pitch is measured on a scale called the mel scale. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's auditory perception level. The below example shows the relationship between the mel frequencies and the linear frequencies.

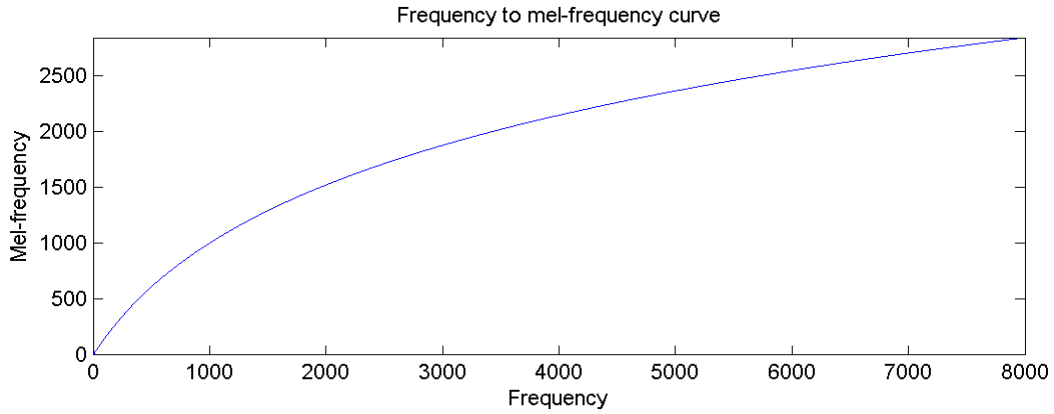


Figure 3.7: Conversion normal frequency to mel frequency

In general, we have two options for the triangular filters, as shown in the next.

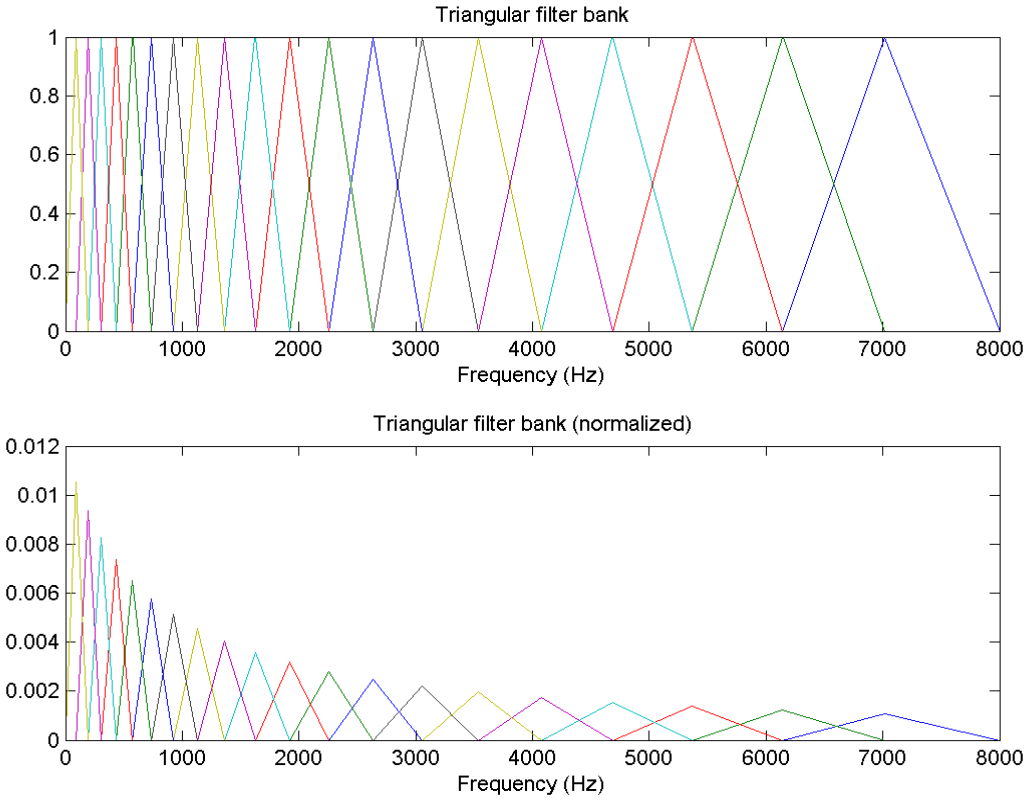


Figure 3.8: Mel frequency triangular filters

The reasons for using triangular bandpass filters

1. To smooth the amplitude spectrum such that the harmonics are flattened in order to get the envelop of the spectrum with harmonics of spoken word. It indicates that the pitch of a speech signal is generally not presented in MFCC. As a result, a speech recognition system will behave more or less the same when the input utterances are of the same timbre but with different tones/pitch.
2. Reduce the size of the features involved.

3.6 DISCRETE COSINE TRANSFORMS (DCT)

Discrete CosineTransform (DCT) to make it easier to remove noise embedded in the noisy spoken signal. This is often done as it is easier to separate the speech energy and the noise energy in the

transform domain. It will be shown in this paper, that the Discrete Cosine Transform (DCT)[25] performs the DFT in terms of speech energy compaction.

$$C_m = \sum_{K=1} N \cos[M*(K-0.5)*\pi / N]E_K \qquad 3.8$$

Where C_m represents MFCC

N represent number of triangular filters

M represents order of cepstral coefficients.

3.7 LINEAR FREQUENCY CEPSTRAL COEFFICIENTS

Some sounds like nasal consonants, reverberatory sounds, and whispered sounds [4] are not sensible to Mel frequency cepstral coefficients. As these sound frequeccies are concentrated at high end of the general auditory frequencies, their MFCC gives less values compared to genral auditory frequencies. So, instead of MFCC, we take linear frequency cepstral coefficients (LFCC). In case obtaining LFCCfor spoken word, the whole process is similar to MFCC except filter banks. Here triangular filter banks over normal frequency scale is considered[15].

At reverberatory sounds LFCC are very sensitive and gives good cepstral values at high end of the frequencies of human auditory system. But LFCC are not generally used for normal speech recognition.

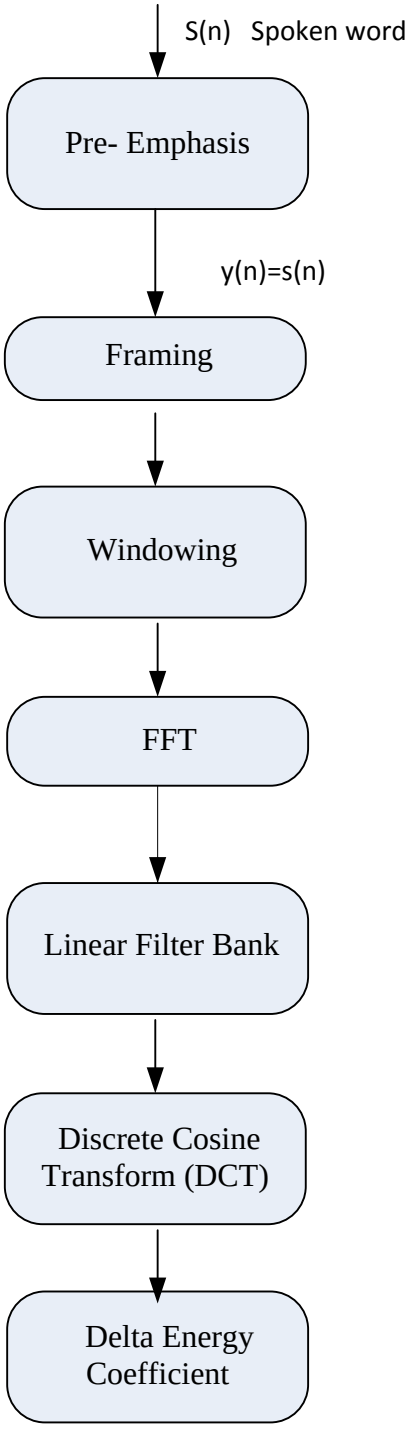


Figure 3.9 Block diagram LFCC

So from the above block we can say signal is converted into frequency domain usinf FT and again into another domain query domain by applying DCT.

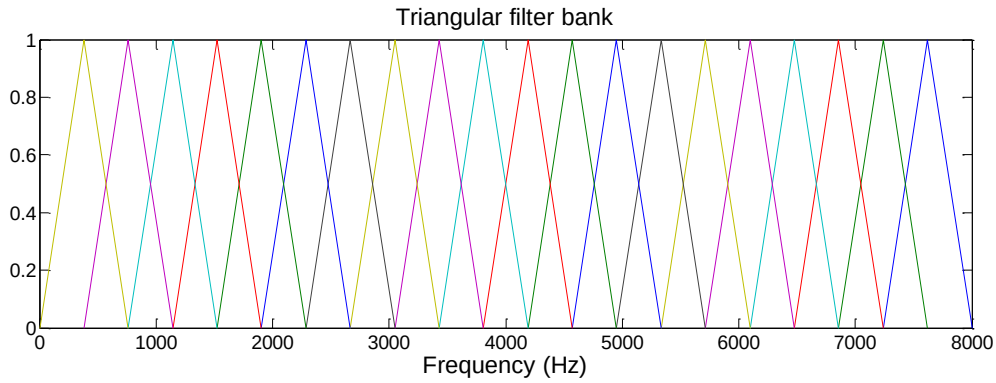


Figure 3.10: Linear triangular filters

In the above diagram triangular band filters are occupied equally over the linear frequencies,rather concentrating oly human auditory frequencies as in cases of Mel fiter bank.

3.8 COMPARISION

Mel-frequency cepstral coefficients (MFCC) have been dominantly used in speaker recognition as well as in speech recognition. However, based on theories in speech production, some speaker characteristics associated with the structure of the vocal tract, particularly the vocal tract length, are reflected more in the high frequency range of speech. This insight suggests that a linear scale in frequency may provide some advantages in speaker recognition over the mel scale. LFCC gave better performance only in nasal and non-nasal consonants [16], not in vowels. used a modified set of LFCC for speaker identification in whisper speech and found LFCC is more robust to whisper speech. Although there were efforts on comparing MFCC and LFCC, the results are inconsistent.

CHAPTER 4

HIDDEN MARKOV MODEL

4.1 DEFINING HIDDEN MARKOV MODEL (HMM)

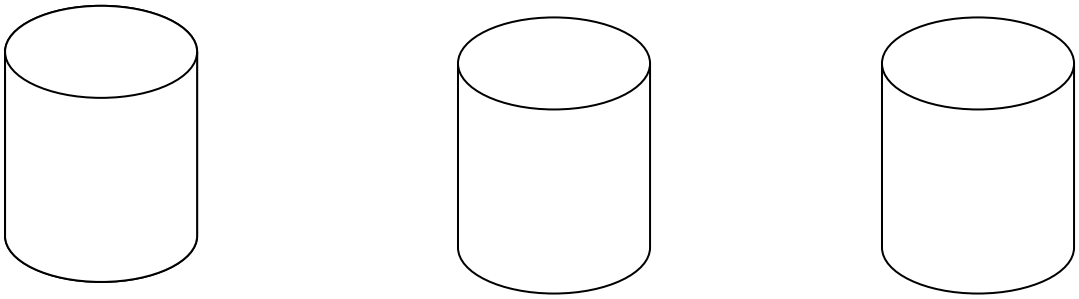
Hidden Markov Model is statistical finite state machine, in which system being modelled is assumed to be in the Markov process. It consists of states to model the sequence of observation data. Here the states are not visible but the output depending on the states are known to us. It generaly statistical Baysian dynamic network with Markov rule considering into account.

4.1.1 MARKOV PROCESS

The Markov process states that the transition to next state depends only on the current state and its output probability of current state but not on all the previous states. Generally the above process is called first order Markov process[17]. Suppose if transition to next state depends ok k previous states then the process is calle k^{th} order markov process.

4.12 MOTIVATING EXAMPLE HIDDEN MARKOV MODEL

Let us consider three containers or urns U_1, U_2 and U_3 which contains 100 balls each. And these each 100 balls are of three different colours Red, Green and Blue in some propotion as shown below.



Urn 1	Urn 2	Urn 3
Red=30	Red=10	Red=60
Green=50	Green=40	Green=10

Figure 4.1: HMM example with urns and balls

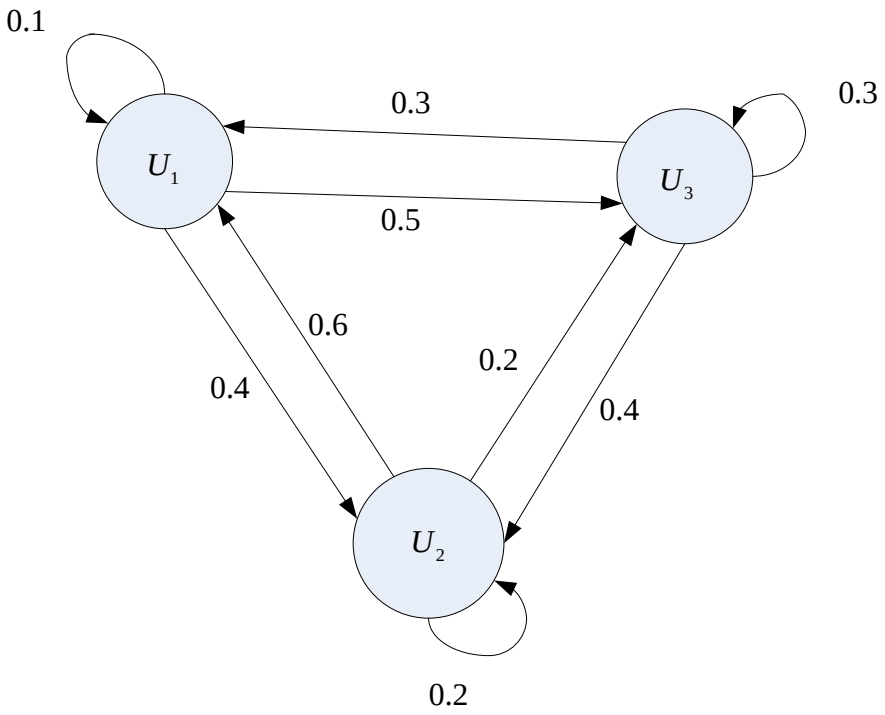


Figure 4.2: Example state transition representation of HMM

In the above example each urn contains balls with 3 different colours. Each a ball is picked up from th any urn unknown (Hidden) and is placed in the same urn from where the ball is being picked up. Then another ball is picked from any urn may be the previous or anyone from oter two depends on the output probability. So here picking up ball is known to us but from which urn it is drawn is hidden. Here transition of states are provided along with the output probabilities of each urn. Problem here is for given observation sequence suppose R G R B G B R the what will be state sequence ? ? ? ? ? ? ?. So similar problem is observed in spoken word recognition usin HMM is described in following discussion.

4.2 ISOLATED SPOKEN WORD RECOGNITION

The spoken word recogniser affectively maps between sequences of spoken vectors (observation vectors) and the wanted given symbol sequences to be recognised [26]. The system implementation is difficult with the two problems being identified. Firstly, the mapping from symbols to speech is not one-to-one

since different given symbols can give rise to similar spoken sounds. Furthermore, there are large variations in the realised spoken word waveform due to speaker speech parameter variability, speaking mood, environment in which spoken, etc. Secondly, the boundaries between symbols cannot be recognised explicitly from the speech waveform.

Consider each spoken word to be represented by a sequence of speech vectors or observations O, stated as

$$O = o_1, o_2, o_3, \ldots\ldots\ldots o_N \tag{4.2}$$

where O_t the spoke word vector observed at time t. The problem of isolated word recognition can be stated as that of evaluating

$$\arg\max_i \{P(w_i \mid O)\} \tag{4.3}$$

where w_i is the ith vocabulary word. This probability can not be calculated directly but using Bayes' Rule

$$P(w_i \mid O) = \frac{P(O \mid w_i)P(w_i)}{P(O)} \tag{4.4}$$

Thus, for a given set of prior probabilities $P(w_i)$, the most probable spoken word depends only on the happening of $P(O \mid w_i)$. For the given dimensionality[18] of the observation vector sequence O, the direct evaluation of the joint conditional probability $P(o_1, o_2, o_3, \ldots\ldots\ldots o_N \mid w_i)$ from examples of spoken words is not achievable. Anyhow, if statistical finite machine like HMM is assumed, then evaluation from data is possible since the problem of estimating the class conditional observation densities $P(O \mid w_i)$ is replaced by the much simpler problem of estimating the Markov model parameters[18].

In spoken word recognition based on HMM, it is understood that the sequence of observed spoken word vectors respective to each word that is generated by a Markov model. HMM is considered as finite state statistical machine [19] that changes state once every time, each time t that a state j is entered, a speech vector o_t is generated from the probability density $b_j(o_t)$. And furthermore, the transition from state i to state j is also probabilistic and is monitored by the discrete probability a_{ij}. The joint probability that O is generated [20] by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities of states and the output probabilities of the states. So for the state sequence

$$X \quad P(O, X \mid M) = a_{12} b_2(\mathrm{o}_1)\, a_{22} b_2(\mathrm{o}_2) a_{23} b_2(\mathrm{o}_3) \ldots\ldots \qquad 4.5$$

However, in general, only the observation sequence O is known to us (visible) and the underlying state sequence X is not visible there is hidden. Hence it is called a Hidden Markov Model. For an unknown X, the required likelihood is evaluated by combining over all the possible state sequences $X = x(1), x(2), x(3) \ldots\ldots x(T)$, that is

$$P(O \mid M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \qquad 4.6$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state of HMM. As an alternative to equation 4.6, generalised evaluation of the likelihood can be obtained by only considering the most likely state sequence.

$$\hat{P}(O \mid M) = \max_{X} \left(\sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right) \qquad 4.7$$

However the equations 4.6 and 4.6 evaluation is not practicable, simple recursive procedures exists which will allow both quantities to be evaluated very efficiently and effectively using recursion. The

problem for isolated spoken word recognition is solved if the equation 4.3 is evaluated by Markov model. For given a set of models M with respect to words w_i , equation 4.3 can be solved.

$$P(O \mid w_i) = P(O \mid M_i) \qquad 4.8$$

All of this, of course, considers that the parameters transition probabilities $\{a_{ij}\}$ and output probabilities $\{b_j(o_t)\}$ are known for each model M_i. For given a set of training sample spoken words with respect to a particular model, the parameters of that model can be computed automatically by a effective and efficient re-estimation procedure available. Thus, for a given sufficient number of representative training samples of each word can be collected then a Hidden Markov Model (HMM) can be developed which implicitly models all of the many parameters of variability inherent in real speech[21]. Fig. 4.3 briefly concludes the use of HMMs for isolated word recognition. Firstly, a HMM is trained for each vocabulary word using a number of examples of that word. In this case, the vocabulary consists of just three words: "one", "two" and "three". Furthermore, to recognise some unknown word, the likelihood of each model generating that word is calculated and the most likely model captures the word[22].

Figure 4.3: Usage of Hidden Markov Model for Spoken utterance recognition (Training)

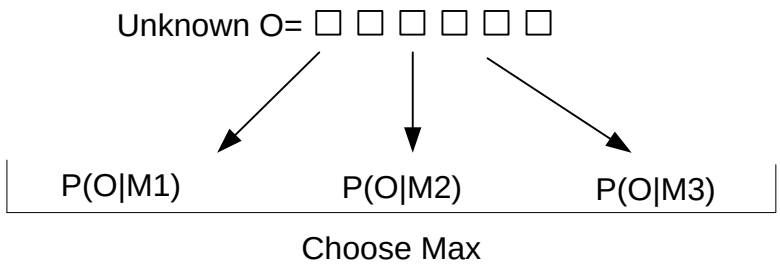


Figure 4.4: Usage of Hidden Markov Model for Spoken utterance recognition (Testing)

4.3 SPECIFICATION OF OUTPUT PROBABILITY

Before the problem of parameter estimation can be discussed in more detail, the form of the output distributions $\{b_j(o_t)\}$ needs to be made explicit. HTK is designed primarily for modelling continuous parameters using continuous density multivariate output distributions. It can also handle observation sequences consisting of discrete symbols in which case, the output distributions are discrete probabilities. For simplicity, however, the presentation in this chapter will assume that continuous density distributions are being used. In common with most other continuous density HMM systems, here the system represents output distributions by Gaussian Mixture Densities. In this system, however, a further generalisation is made. Here the HMM allows each observation vector at time t to be split into a number of S independent data streams o_{st}. The formula for computing is then $\{b_j(o_t)\}$

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} N\left(o_{st}; \mu_{jsm}, \Sigma_{jsm}\right) \right]^{\gamma_s} \qquad 4.9$$

where M_s is the number of mixture components in stream s, c_{jsm} is the weight of the m'th compo-nent and $N(\cdot; \mu, \Sigma)$ is a multivariate Gaussian with mean vector ¹ and covariance matrix Σ, that is

$$N(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1} (o-\mu)} \qquad 4.10$$

where N is the dimensionality of 0.

The exponent γ_s is a stream weight1. It can be used to give a particular stream more importance, however, it can only be set manually. Multiple data streams are used to enable separate modelling of multiple information sources. Here, the processing of streams is completely general. However, the speech input modules assume that the source data is split into at most 4 streams. It is necessary to remark that the default streams are the basic parameter vector, first (delta) and second (acceleration i.e., delta delta) difference coefficients and log energy.

4.4 RE-ESTIMATION OF PARAMETERS USING BAUM-WELCH METHOD

For computing the parameters of a HMM, the first need to make a rough guess at what they would be. After doing this, the more accurate (in the maximum likelihood sense) parameters can be found by applying the so-called Baum-Welch re-estimation method.

Here the basis of the formulae will be presented in a very informal way. At first, it is to be noted that the consideration of multiple data streams does not alter matters effectively as each stream is taken to be statistically independent. Furthermore, mixture components can be viewed to be a special form of sub-state in which the transition probabilities are the mixture weights [23]. Therefore, the most important problem is to compute or estimate the means and variances of a HMM in which each state output distribution is a single component Gaussian, that is

$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^n \left|\sum_j\right|}} e^{-\frac{1}{2}(o_t-\mu_j)^T \sum_j^{-1}(o_t-\mu_j)} \qquad 4.11$$

If we consider only one state j in the HMM, this estimation of parameter would be easy. The maximum likelihood estimates of μ_j and $\sum_j$ would be just the simple averages, that is

$$\mu_j = \frac{1}{T}\sum_{t=1}^{T} o_t \qquad 4.12$$

and

$$\Sigma_j = \frac{1}{T}\sum_{t=1}^{T} (o_t - \mu_j)(o_t - \mu_j)^T \qquad 4.13$$

In general practice, there are many states and there is no direct alignment of observation vectors to the individual states since the underlying state sequence is not known to us. Anyhow, if some approximate assignment of vectors to states can be made then equations 4.11 and 4.12 could be used to give the required initial values for the parameters. Here we first divides the training observation vectors equally amongst the model states and then uses equations 4.11 and 4.12 to give initial values for the mean and variance of each state. We then calculates the maximum likelihood state sequence using the Viterbi algorithm discussed below, reassigns the observation vectors to states and then uses equations 4.11and 4.12 again to get better initial values. This process is repeated until the estimates do not change. Since the full likelihood of each observation sequence is based on the summation of all possible state sequences, each observation vector o_t contributes to the computation of the maximum likelihood parameter values for each state j . In other words, instead of assigning each observation vector to a specific state as in the above approximation, each observation is assigned to every state in proportion to the probability of the model being in that state when the vector was observed. Thus, if $L_j(t)$ denotes the probability of being in state j at time t then the equations 4.11 and 4.12given above become the following weighted averages

$$\mu_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)} \qquad 4.14$$

$$\Sigma_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t - \mu_j)(o_t - \mu_j)^T}{\sum_{t=1}^{T} L_j(t)} \qquad 4.15$$

where, the summations in the denominators are taken to give the required normalisation. Equations 4.14 and 4.15 are the Baum-Welch re-estimation formulae for computing the means and covariance's of a HMM. A similar but slightly more complex formula can be derived for the transition probabilities (see chapter 8).Of course, to apply equations 4.13 and 4.14, the probability of state occupation $L_j(t)$ must be calculated. This is done efficiently using the so-called Forward-Backward algorithm. Let the forward probability $\alpha_j(t)$ for some model M with N states be defined as

$$\alpha_j(t) = P\left(o_1, ...o_t, x(t) = j \mid M\right) \qquad 4.16$$

That is, $\alpha_j(t)$ is the joint probability of observing the first t speech vectors and being in state j at time t. This forward probability can be efficiently calculated by the following recursion

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1)\alpha_{ij}\right] b_j(o_t) \qquad 4.17$$

The above recursion depends on the fact that the probability of being in state j at time t and seeing observation o_t can be deduced by summing the forward probabilities for all possible predecessor states i weighted by the transition probability α_{ij} . The slightly odd limits are caused by the fact that states 1 and N are non-emitting3. The initial conditions for the above given recursion are

$$\alpha_1(1) = 1$$

$$\alpha_j(1) = \alpha_{1j} b_j(o_t) \qquad 4.18$$

for $1 < j < N$ and the final condition is given by

$$\alpha_N(T) = \left[\sum_{i=2}^{N-1} \alpha_i(T)\alpha_{iN} \right] \quad 4.19$$

The calculation of the forward probability also yields the total likelihood $P(O \mid M)$.

The backward probability $\beta_j(t)$ is defined as

$$\beta_j(t) = P\left(o_{t+1}, ...o_T \mid x(t) = j, M\right) \quad 4.20$$

As in the forward case, this backward probability can be computed efficiently using the following recursion

$$\beta_j(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_j(t+1) \quad 4.21$$

with initial condition given by

$$\beta_j(T) = \alpha_{iN} \quad 4.22$$

for $1 < i < N$ and final condition given by

$$\beta_1(t) = \sum_{j=2}^{N-1} a_{1j} b_1(o_1) \beta_j(1) \quad 4.23$$

Notice that, in the definitions mentioned above, the forward probability is a joint probability whereas the backward probability is a conditional probability [28]. This somewhat asymmetric definition is deliberate since it allows the probability of state occupation to be determined by taking the product of the two probabilities. From the definitions,

$$\alpha_j(t)\beta_j(t) = P(O, x(t) = j \mid m) \quad 4.24$$

Hence,

$$L_j(t) = P(x(t) = j \mid O, m) \qquad 4.25$$

$$= \frac{P(O, x(t) = j \mid m)}{P(O \mid M)}$$

$$= \frac{1}{P} \alpha_j(t) \beta_j(t)$$

where $P = P(O \mid M)$.

All of the information needed to perform HMM parameter[24] re-estimation using the Baum-Welch algorithm is now in place. The steps in this algorithm may be summarised as follows

1. For every parameter vector/matrix requiring re-estimation, allocate storage for the numerator and denominator summations of the form illustrated by equations 4.13 and 4.14. These storage locations are referred to as accumulators
2. Calculate the forward and backward probabilities for all states j and times t [29].
3. For each state j and time t, use the probability $L_j(t)$ and the current observation vector o_t update the accumulators for that state.
4. Use the final accumulator values to calculate new parameter values.
5. If the value of $P = P(O \mid M)$ for this iteration is not higher than the value at the previous iteration then stop, otherwise repeat the above steps using the new re-estimated parameter values.

All of the above assumes that the parameters for a HMM are re-estimated from a single observation sequence, that is a single example of the spoken word. In practice, many examples are needed to get good parameter estimates. However, the use of multiple observation sequences adds no additional complexity to the algorithm. Steps 2 and 3 above are simply repeated for each distinct training sequence. One final

point that should be mentioned is that the computation of the forward and backward probabilities involves taking the product of a large number of probabilities. In practice, this means that the actual numbers involved become very small. Hence, to avoid numerical problems, the forward-backward computation is computed using normal way.

4.5 RECOGNITION AND VITERBI DECODING

The previous section has described the basic ideas underlying HMM parameter re-estimation using the Baum-Welch algorithm. In passing, it was noted that the efficient recursive algorithm [25] for computing the forward probability also yielded as a by-product the total likelihood $P(O|M)$. Thus, this algorithm could also be used to find the model which yields the maximum value of $P(O|M_i)$, and hence, it could be used for recognition.

In practice, however, it is preferable to base recognition on the maximum likelihood state sequence since this generalises easily to the continuous speech case whereas the use of the total probability does not. This likelihood is computed using essentially the same algorithm as the for-ward probability calculation except that the summation is replaced by a maximum operation. For a given model M, let $\phi_j(t)$ represent the maximum likelihood of observing speech vectors o_1 to o_t and being in state j at time t. This partial likelihood can be computed efficiently using the following recursion.

$$\phi_j(t) = \max_i \left\{ \phi_i(t-1) a_{ij} \right\} b_j(o_t)$$

4.26

$$\phi_1(t) = 1$$

$$\phi_j(1) = a_{1j} b_j(o_1)$$

where

for $1 < j < N$. The maximum likelihood $\hat{P}(O \mid M)$ is then given by

$$\phi_N(T) = \max_i \left\{ \phi_i(T) a_{iN} \right\} \qquad 4.27$$

As for the re-estimation case, the direct computation of likelihoods [27] leads to under flow, hence, log likelihoods are used instead. The recursion of equation then becomes

$$\psi_j(t) = \max_i \left\{ \psi_i(t-1) + log(a_{ij}) \right\} + log(b_j(o_t)) \qquad 4.28$$

This recursion forms the basis of the so-called Viterbi algorithm. This algorithm can be visualised as finding the best path through a matrix where the vertical dimension represents the states of the HMM and the horizontal dimension represents the frames of speech (i.e. time). Each large dot in the picture represents the log probability of observing that frame at that time and each arc between dots corresponds to a log transition probability. The log probability of any part is computed simply by summing the log transition probabilities and the log output probabilities along that path. The paths are grown from left-to-right column-by-column. At time t, each partial path $\psi_i(t-1)$ is known for all states i, hence equation 4.28 can be used to compute $\psi_i(t)$ thereby extending the partial paths by one time frame.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 INTRODUCTION TO RESULTS

Several experiments with different spoken utterances are taken into consideration while obtaining the overall recognition rate using Hidden Markov Model (HMM). Here in this chapter, results obtained for speaker independent, small vocabulary isolated spoken word recognition are discussed. The main objective of the thesis is to recognise or detect set of ten spoken utterances or words. We considered two set of words consists of :{ one, two, three, four, five, six, seven, eight, nine and ten}, {apple, pen, hi , calm, eight, vain, hello, poem, key, move}. These words recorded with micro phone using audacity software through which recorded signals are directly represented in '**.wav'** format. Many different models were taken into consideration like frame length variation, order of Mel frequency cepstral coefficients and linear frequency cepstral coefficients and their delta coefficients.

Feature extraction	:	MFCC and LFCC and their delta coefficients
Order features	:	13 (12 MFCC/LFCC + Log energy)
Filter banks	:	Triangular (both Mel and linear)
Frame size	:	256

Confusion matrix is used for comparining the results of finite machine classification methods like Hidden Markov Model used for word classification.

Table 5.1: Confusion matrix for spoken word recognition using HMM with MFCC method for first data set.

Word	**Rate of Classification of spoken words in %**									
	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten
One	**62.5**	2.0911	1.2533	2.9832	2.0221	1.9892	2.0101	1.8989	1.9789	2.0231
Two	0.5423	**90.7**	0.8798	0.8976	0.5986	0.9837	0.3273	0.9823	1.0021	0.9919
Three	0.5687	0.9889	**87.4**	0.1290	0.9213	1.1021	0.5476	0.8916	0.9812	1.9801
Four	2.0912	1.9801	1.5198	**72.9**	0.9867	1.1432	1.5643	0.9989	1.4321	1.0131
Five	0.5658	0.4356	0.2189	0.6721	**90.3**	0.6721	0.4521	0.5121	0.5426	0.1296
Six	0.1287	0.9978	0.6412	0.6712	0.7121	**88.6**	1.2132	0.5412	0.3217	0.9874
Seven	2.5412	3.1654	2.7896	2.8970	2.5698	2.7645	**70.1**	3.1643	3.1219	2.0012
Eight	2.0152	2.0891	1.9867	1.8934	2.0141	1.9847	1.7832	**75.5**	2.0213	1.8791
Nine	2.9892	2.6753	2.0541	1.9989	2.0972	1.9834	1.8985	2.0012	**59.4**	2.0234
Ten	1.1534	1.5354	1.9837	2.7474	2.0284	1.9462	1.3735	2.0125	1.2251	**60.9**
Average classification rate is 80.21										

The above confusion matrix tells the classification of words using Hidden Markov model with MFCC as feature extraction method for spoken word recognition. A good recognition rate is achieved for general spoken words. Here the recognition rate achieved is 80.21and is achieved using MFCC but use of other methods does not gives this rate of recognition.

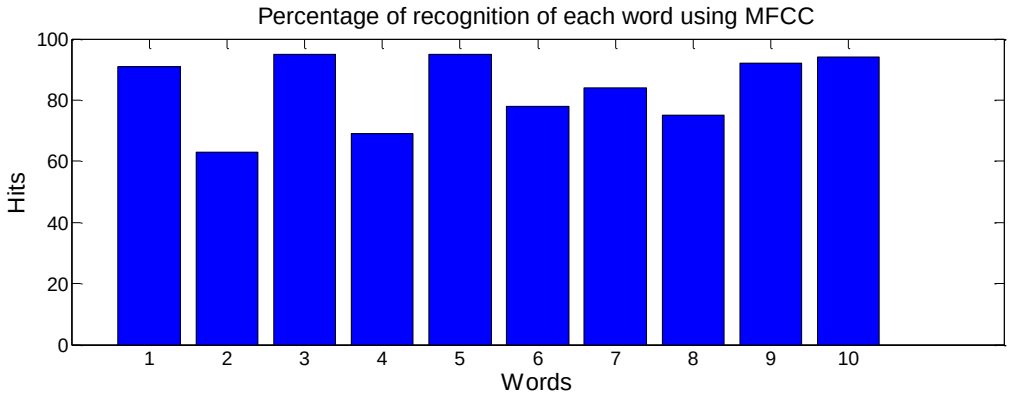


Figure 5.1: Percentage of recognition of each word using MFCC for first data set

In above figure percentage recognition is more at general spoken words like three,five but rate of recognition less for nasal consonants like nine and ten.

Table 5.2: Confusion matrix for spoken word recognition using HMM with LFCC method for first data set

Word	**Rate of Classification of spoken words in %**									
	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten
One	**75.9**	0.9282	0.8932	0.3661	0.9842	1.8462	0.8946	0.9742	0.7345	0.2572
Two	1.9383	**65.3**	1.9836	1.0253	1.0993	1.9832	1.7326	1.3781	1.7361	0.9984
Three	1.4740	1.4702	**55.4**	1.8492	2.7409	1.8372	1.8421	0.9942	1.0937	1.7402
Four	2.8923	2.8934	2.9032	**65.8**	1.9373	1.8932	1.9735	2.3728	1.3810	1.3809
Five	3.7459	2.9845	2.4632	1.9989	**54.3**	3.4874	2.3579	2.5403	2.1293	2.1281
Six	2.7420	2.9484	2.8392	2.0028	1.9467	**60.2**	1.2937	2.7503	2.8311	2.6472
Seven	0.2837	0.5412	0.6728	0.3782	0.3782	0.6378	**92.8**	0.7892	0.6478	0.7432
Eight	0.2783	1.3809	1.3833	1.0228	1.0321	0.9845	0.9283	**70.3**	1.3732	0.9221
Nine	0.2121	0.1282	0.8391	0.8922	0.9372	0.1243	0.1276	0.4231	**80.9**	0.5312
Ten	0.9937	1.2873	1.2618	1.6381	1.3789	1.3791	1.1791	1.8974	1.4397	**70.7**
Average classification rate is 68.19										

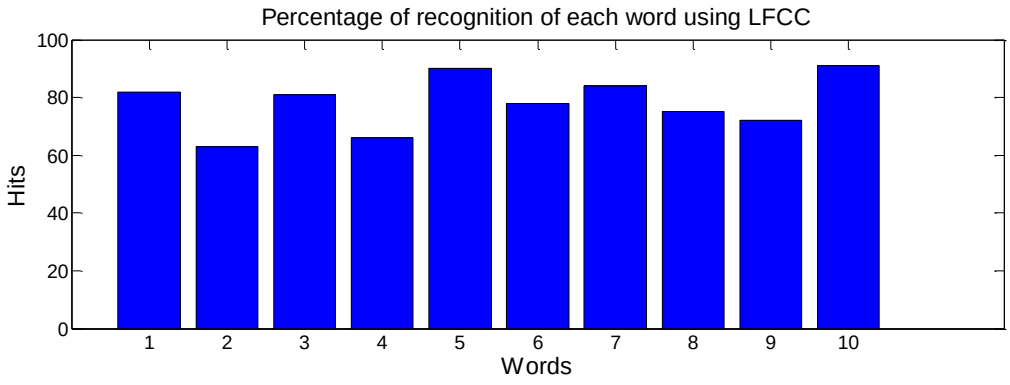


Figure 5.2: Percentage of recognition of each word using MFCC for first data set

SECOND DATA SET MATRIX

Table 5.3: Confusion matrix for spoken word recognition using HMM with MFCC method for second data set.

Word	**Rate of Classification of spoken words in %**									
	Apple	Pen	Hi	Calm	Eight	Vain	Hello	Poem	Key	Move
Apple	**90.6**	3.2135	2.3675	0.4175	0.3658	0.8288	0.2658	0.2418	0.2758	0.2058
Pen	0.3342	**77.3**	1.7958	0.3250	2.3292	2.3500	0.5625	1.9090	2.8787	0.8945
Hi	0.2345	0.4657	**92.5**	0.3419	1.3647	1.4235	0.8568	0.7783	0.3912	1.9987
Calm	0.3467	0.7779	0.4389	**63.4**	1.7856	1.0989	0.8964	0.7841	0.9999	3.7857
Eight	2.8998	2.6576	3.8758	1.8475	**65.5**	0.5788	1.3849	1.9059	2.8789	1.8898
Vain	0.8998	0.6677	0.8756	0.7889	0.7847	**50.1**	1.8589	0.89859	0.8985	1.7588
Hello	3.8598	2.8587	1.5878	2.4989	1.4988	3.9038	**70.5**	1.3899	2.4989	2.9488
Poem	0.4948	0.3984	1.9394	0.4898	2.1090	0.9093	1.9499	**52.4**	0.9849	0.3984
Key	2.2930	3.5940	2.0903	1.4788	2.9399	3.2889	2.9939	0.3989	**77.8**	2.4939
Move	3.2903	3.9029	2.9189	3.3930	2.9030	3.9039	1.3092	0.3929	3.1909	**59.4**
Average classification is 65.95										

Table 5.4: Confusion matrix for spoken word recognition using HMM with LFCC method for second data set.

Word	**Rate of Classification of spoken words in %**									
	Apple	Pen	Hi	Calm	Eight	Vain	Hello	Poem	Key	Move
Apple	**69.2**	3.2135	2.3675	0.4175	0.3658	0.8288	0.2658	0.2418	0.2758	0.2058
Pen	0.3342	**86.5**	1.7958	0.3250	2.3292	2.3500	0.5625	1.9090	2.8787	0.8945
Hi	0.2345	0.4657	**60.5**	0.3419	1.3647	1.4235	0.8568	0.7783	0.3912	1.9987
Calm	0.3467	0.7779	0.4389	**92.3**	1.7856	1.0989	0.8964	0.7841	0.9999	3.7857
Eight	2.8998	2.6576	3.8758	1.8475	**65.5**	0.5788	1.3849	1.9059	2.8789	1.8898
Vain	0.8998	0.6677	0.8756	0.7889	0.7847	**85.7**	1.8589	0.89859	0.8985	1.7588
Hello	3.8598	2.8587	1.5878	2.4989	1.4988	3.9038	**59.8**	1.3899	2.4989	2.9488
Poem	0.4948	0.3984	1.9394	0.4898	2.1090	0.9093	1.9499	**85.3**	0.9849	0.3984
Key	2.2930	3.5940	2.0903	1.4788	2.9399	3.2889	2.9939	0.3989	**60.2**	2.4939
Move	3.2903	3.9029	2.9189	3.3930	2.9030	3.9039	1.3092	0.3929	3.1909	**50.6**
Average classification is 79.13										

5.2 COMPARISION OF RECOGNITION RATE FOR MFCC AND LFCC

Here we used two feature extraction methods to obtain better recognition rate at different ranges of human voice frequency.

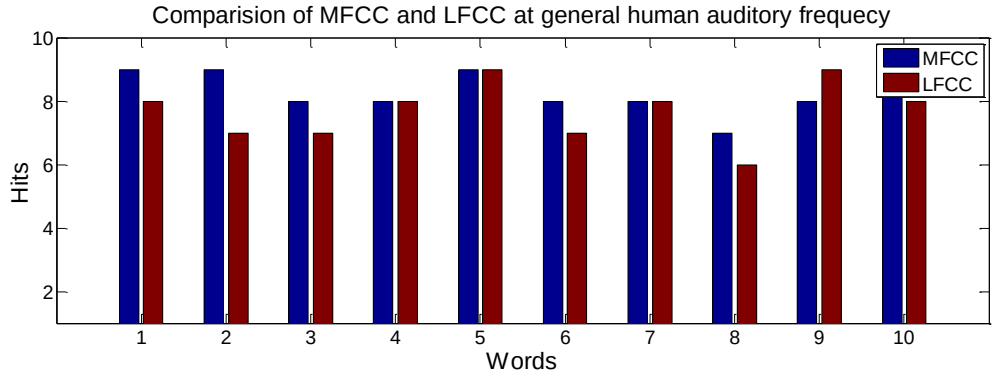


Figure 5.3: Comparision of MFCC and LFCC at general human auditory frequency.

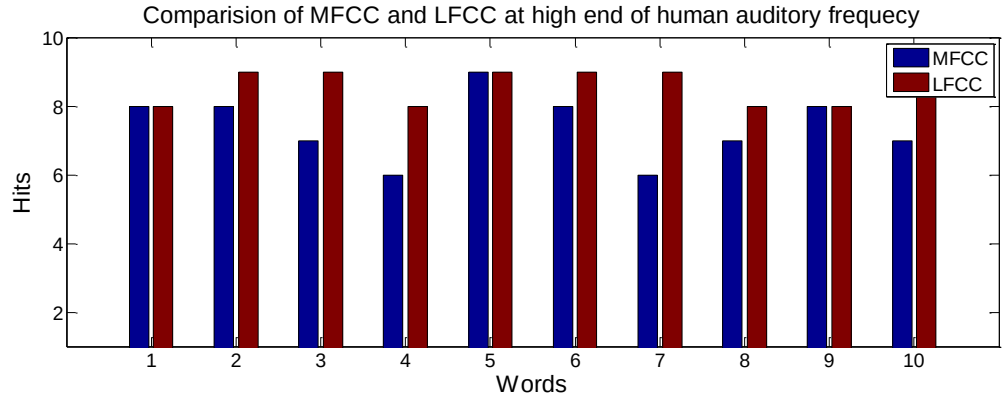


Figure 5.4: Comparision of MFCC and LFCC at high end of human auditory frequency.

In the above figure recognition of words spoken at high end of the human auditory frequencies like vain, poem are detected with good accuracy rate with LFCC method of feature extraction. But for the above words MFCC performance is very less compared to LFCC.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

This thesis work is an extension of the work done by LAWRENCE RABINIER, developer of HMM tool kit, Naval post graduate school and others. The main objective of this project is to get good recognition rate at high end of the human voice generally female voice, to increase the number of spoken words, and finally to improve the overall recognition rate of isolated spoken word utterances. The above showed results conclude that the work in this project improves the performance of recognition especially at the high end of the human voice. All the objectives proposed in this thesis were successfully achieved by implementing MFCC, LFCC and discrete hidden markov models in MATLAB programming language.

6.2 FUTURE WORK

Even though recognition rate is improved compared to previous models still there is room for improvement in recognition especially when number of speakers increased and no of spoken words to be recognised increases. Here in this proposed model recognition spoken utterance ‘two’ made difficult with poor recognition rate since it contains air flow from the mouth of the speaker while recording. A drawback with the use of HMM is time consuming. The proposed model takes much time during the training of the signals, especially during speech parameter estimation and re-estimation. So, there is possibility of reducing the time for while simulating the results. One can develop a model for recognition of speech such that automatic dictation should intern produce a text file, which is not yet done with good accuracy for sure because if it had been developed there is no need of typing a text.

REFERENCE

[1] Xing Fan and John H.L. Hansen, "*Speaker Identification with Whispered Speech based on modified LFCC Parameters and Feature Mapping*", in ICASSP 2009, Taipei, Taiwan.

[2] C. Miyajima, H. Watanabe, K. Tokuda, T. Kitamura, S. Katagiri, "*A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction,*" Speech Communication, Volume 35, Issues 3-4, October 2001, Pages 203-218.

[3] S. Thomas, S. Ganapathy, and H. Hermansky, "*Recognition of Reverberant Speech Using Frequency Domain Linear Prediction*," IEEE Signal Processing Letters, vol. 15, pp. 681–684, 2008.

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

[5] H. Lei & E. Lopez "*Mel, Linear, and Anti-Mel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition*", Inter speech 2009, Brighton, UK, 2009.

[6] F. Harris, "*On the use of windows for harmonic analysis with the discrete Fourier transform,* "Proc. IEEE, vol. 66, no. 1, pp. 51–83, Jan.1978.

[7] *Distributed real time speech recognition* http://www.google.com/patents/US6633846

[8] A.Oppenheim, S.Willsky, and S.Nawab, Signals and Systems, 2nd ed. New Delhi, India: PHI Learning, 2009.

[9] L.Rabiner, B.H. Juang: "*Fundamentals of Speech Recognition*," PTR Prentic-Hall, 1993

[10] J. Sandberg, M. Hansson Sandsten, T. Kinnunen, R. Saeidi, P. Flan Drin, and P. Borgnat, "*Multitaper estimation of frequency-warped cepstra with application to speaker verification*,"IEEE Signal Process.Lett. vol. 17, no. 4, pp. 343–346, Apr. 2010.

[11] F.Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska Delacreetaz, and D.A. Reynolds, "*A tutorial on text-independent speaker verification*," EURASIP Journal on Applied Signal Processing, Hindawi Publishing Cor-poration, vol. 4, pp. 430–451, 2004.

[12] D.K. Kim and N.S. Kim, "*Maximum a posteriori adaptation of HMM parameters based on speaker space projection*," Speech Communication, vol. 42, no. 1, pp. 59–73, Jan. 2004.

[13] WU CHOU, "*Discriminant-Function-Based Minimum Recognition Error Rate Pattern Recognition Approach to Speech Recognition*," PROCEEDINGS OF THE IEEE, VOL. 88, NO. 8, AUGUST 2000.

[14] J. R. Deller, Jr., J. G. Proakis, J. H. Hansen. (1993). Discrete-Time Processing of Speech Signals. New York: Macmillan.

[15] "Phonetics and Theory of Speech Production" http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap3.html

[16] F. Rosdi and R. N. Ainon, "*Isolated malay speech recognition using Hidden Markov Models*," in Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on, 2008, pp. 721-725.

[17] Huang, X., Alex, A., and Hon, H. W. (2001). *Spoken Language Processing*; A Guide to Theory, Algorithm and System Development. Prentice Hall, Upper Saddle River, New Jersey.

[18] Jelinek F. "*Statistical Methods for Speech Recognition*". The MIT Press, Cambridge, Massachusetts.

[19] Rabiner, L. R. and Juang, B. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey.

[20] Ting, H. N., Jasmy, Y. and Sheikh, Hussain, S. S. "Speaker-Independent Malay Syllable Recognition Using Singular Multi-layer Perceptron Network." Proceedings of the 2002 International Conference on Artificial Intelligence In Engineering and Technology (ICAIET 2002), Kota Kinabalu (Malaysia), 17-18 June 2002.

[21] Dimov, D., and Azmanov "Experimental specifics of using HMM in isolated word speech recognition." International Conference on Computer Systems and Technologies Comp Sys Tech'2005.

[22] N. Najkar, F. Razzazi, and H. Sameti, "*A novel approach to HMM-based speech recognition system using particle swarm optimization*," in BIC-TA 2009 - Proceedings, 2009 4th International Conference on Bio-Inspired Computing: Theories and Applications, 2009, pp. 296-301.

[23] Antanas Lipeika, Joana Lipeikiene, Laimutis Telksnys, "*Development of Isolated Word Speech Recognition System,*" Informatica 2002, Vol 13, 37-46, 2002.

[24] I. Bazzi and J. R. Glass, "*Modelling out-of-vocabulary words for robust speech recognition,*" in Proceedings of Int. Conf. Spoken Language Processing, Beijing, China, October 2000, pp. 401–404.

[25] D., Jurafsky and J.H., Martin, "*Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*". Prentice Hall, Upper Saddle River, NJ, USA, 2000.

[26] M. A. M. Abu Shariah, R. N Ainon, R. Zainuddin, and O. O. Khalifa, "*Human Computer Interaction Using Isolated-Words Speech Recognition Technology*," IEEE Proceedings of The International Conference on Intelligent and Advanced Systems (ICIAS'07),Kuala Lumpur, Malaysia, pp. 1173 – 1178, 2007.

[27] Speech recognition system interactive enthttp://www.google.com/patents/US7277854

[28] Steve Young, Gunnar Evermann, Mark Gales "*The HTK Book*" Microsoft Corporation
http://www.ee.columbia.edu/~dpwe/LabROSA/doc/HTKBook21/node5.html

[29] Armin Sehr, Roland Maas, and Walter Kellermann, "*Reverberation Model-Based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition.*" IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, No. 7, September 2010.