HMM-BASED ISOLATED AND CONNECTED WORD SPEAKER INDEPENDENT SPEECH RECOGNITION USING DIFFERENT ACOUSTIC MODELS

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Shagun
(Roll No. 801132026)

Under the supervision of:
Mr. Ravinder Kumar
Asst. Professor, CSED
&
Mr. Karun Verma
Asst. Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

July 2013

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*HMM-Based Isolated and Connected Word Speaker Independent Speech Recognition Using Different Acoustic Models*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Mr. Ravinder Kumar and Mr. Karun Verma*, and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Shagun Bansal
Signature: (15-07-2013)
(Shagun)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. Ravinder Kumar)
Asst. Professor, CSED,
Thapar University,
Patiala

15/7/13
(Mr. Karun Verma)
Asst. Professor, CSED,
Thapar University,
Patiala

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude towards THAPAR UNIVERSITY, for providing me a platform to do my thesis work at such an esteemed institute.

I wish to express my respect, deep sense of gratitude and indebtedness to my guides **Mr. Ravinder Kumar,** Assistant Professor, Department of Computer Science and Engineering, Thapar University, Patiala, and **Mr. Karun Verma,** Assistant Professor, Department of Computer Science and Engineering, Thapar University, Patiala for their invaluable and enthusiastic guidance, useful suggestions, unfailing patience and sustained encouragement throughout this work.

I would like to thank **Dr. Maninder Singh,** Head, Department of Computer Science and Engineering, Thapar University, Patiala for kind help, guidance, encouragement and providing the necessary facilities to carry out my research. I am indebted to the faculty members of the department for valuable suggestions, friendly support and full cooperation rendered by all of them.

I would like to acknowledge the help and support of Abhishek Bansal, Shelly, Shivi, Hitesh, Virender, Puneet, Priya, Prince, Shini, Kanika, Tejinder and Rinky for collection of speech data.

I am very grateful for the support I got from my family and friends. I would also like to express my gratitude to my parents for everything they have done for me. I have no words to mention the support and patience of my parents. Mom, thank you for your prayers, encouragement and your sincere belief in me.

Last, but not the least, I am thankful the Supreme Power "The God", one who has always guided me to work on the right path of the life. Without his grace, this would never come to be today's realty. With special thanks, I dedicate this thesis to God.

Shagun

ABSTRACT

Speech recognition is the independent, computer-driven transcription of spoken language into readable text in real time. It is the technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although automatic speech recognition (ASR) technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services.

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech. Firstly, speech recognition engine converts the speech signal into a sequence of vectors which are measured throughout the duration of the speech signal. Then, using a syntactic decoder it generates a valid sequence of representations.

In the work presented a HMM-based isolated and connected word speaker independent speech recognition system has been developed. Two different approaches are used for modeling the system. One is word-based approach as acoustic model and another one is triphone-based approach as acoustic model.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Speech Recognition is the process of converting spoken speech into text. Speech recognition system is thus sometimes referred to as speech-to-text system. Speech recognition, also referred to as voice recognition, allows you to provide input to an application with your voice. Just like clicking with your mouse, typing on your keyboard, or pressing a key on the phone keypad provides input to an application, speech recognition allows you to provide input by talking. In the desktop world, you need a microphone to be able to do this. So, Speech recognition provides an alternative to traditional methods of interacting with a computer, such as textual input through a keyboard. An effective system can replace or reduce the reliability on standard keyboard and mouse input. This can especially assist the following:

- People who have little keyboard skills or experience, who are slow typists, or do not have the time or resources to develop keyboard skills.
- Dyslexic people or others who have problems with character or word use and manipulation in a textual form.
- People with physical disabilities that affect either their data entry, or ability to read (and therefore check) what they have entered.

The research work on automatic speech recognition (ASR) has been done for almost four decades. Based on major advances in statistical modeling of speech in the 1980s, automatic speech recognition systems today find widespread application in tasks that require a human-machine interface. While we are still far from having a machine that converses with humans on any topic like another human, many important scientific and technological advances have taken place, bringing us closer to the machines that recognize and understand fluently spoken speech.

There are many issues behind the speech recognition like in which environment recognition to be done, recognizing the speech is speaker independent or speaker dependent, size of vocabulary to be used by the speech recognition system, ability to recognize isolated words and/or continuous words and many more.

1.1 Basic Terminology

Following are a few of the basic terms and concepts that are fundamental to speech recognition.

- **Utterance**

When the user says something, this is known as an utterance. An utterance is any stream of speech between two periods of silence. Silence, in speech recognition, is almost as important as what is spoken, because the start and end of an utterance is determined from the silence timing in the speech.

Utterances are sent to the speech engine to be processed. The speech recognition engine is "listening" for speech input. When the engine detects audio input, the beginning of an utterance is signaled. Similarly, when the engine detects a certain amount of silence following the audio, the end of the utterance occurs. If the user doesn't say anything, the engine returns what is known as a silence timeout. Utterances can be a single word, a few words, a sentence, or even multiple sentences. For example, "start", "click", "click Recycle Bin" or "refresh speech commands" are all examples of possible utterances.

- **Dictionary**

A text file used by a speech engine that defines the phonemes corresponding to the words present in the grammar. The dictionary or vocabulary is made up of all the words in all active grammars. Generally, smaller vocabularies are easier to recognize, while larger vocabularies are more difficult. Unlike normal dictionaries, each entry doesn't have to be a single word. There can be multiple words in a single entry. For example Stand Up, it is a multi word entry in vocabulary file.

- **Phoneme**

The basic unit of sound is phoneme. In the same way that written words are composed of letters, a spoken word is composed of various phonemes. For example, the English word "six" has four phonemes (the "s" sound, the "ih" sound, the "k" sound and the "s" sound) but three letters. A speech engine uses its dictionary to break up words and utterances into phonemes, and compares them to one another to perform ASR.

- **Language Model or Grammar**

A grammar is defined for a Speech-to-Text system to govern, which combination of words the system can recognize. Grammar defines the domain, or context, within which the recognition engine works. The engine compares the current utterance against the words and phrases in the active grammar. If the user says something that is not in the grammar, the speech engine will not be able to recognize it correctly.

A grammar uses a particular syntax, or set of rules, to define the words and phrases that can be recognized by the engine. For example, for the voice dialing application, a suitable grammar can be

$digit = ONE|TWO|THREE|FOUR|FIVE|SIX|SEVEN|EIGHT|NINE|OH|ZERO;

$name = SHAGUN [BANSAL] | KARUN [VERMA] | RAVINDER [KUMAR];

(SENT-START (DIAL <$digit> | (PHONE|CALL) $name) SENT-END)

Where the vertical bars denote alternatives, the square brackets denote optional items and the angle braces denote one or more repetitions.

- **Acoustic Model**

Acoustic model contains a statistical representation of the distinct sounds that make up each word in the language model or grammar. Each distinct sound corresponds to a phoneme.

- **Decoder**

Decoder is a software program that takes the sounds spoken by a user and searches the acoustic model for the equivalent sounds. When a match is made, the decoder determines the phoneme corresponding to the sound; it keeps track of the matching phonemes until it reaches a pause in the user's speech. It then searches the language model or grammar file for the equivalent series of phonemes. If a match is made it returns the text of the corresponding word or phrase to the calling program.

- **Accuracy**

The performance of a speech recognition system is measurable. The most widely used measurement is accuracy. It is typically a quantitative measurement and can be

calculated in several ways. The most important measurement of accuracy is whether the desired end result occurred. This measurement is useful in validating application design. For example, if the user says "call", the engine returns "call" and the "CALL" action executes, it is clear that the desired end result is achieved.

Recognition accuracy is an important measure for all speech recognition applications. It is tied to grammar design and to the acoustic environment of the user. To enhance the system accuracy, the simple way is, adjust the application and its grammars based on the results obtained when testing the application with typical users.

1.2 Reducing Extraneous Factors

There are a number of methods for increasing the accuracy and ease of use of speech recognition systems.

- Using a high-performance computer gives better results. If making the real time application, there is need in the computer to contain a fast processor and a large amount of RAM in order to work efficiently.
- It is better to use a good quality microphone. Microphones with “Active Noise Reduction” or “Active Noise Cancellation” can reduce the amount of background noise that can “confuse” the recognition system.
- Installing a good sound card also improves the recognition accuracy. Soundcards that come installed in computers are often of variable quality. For serious use of speech recognition systems, high quality duplex (input and output) sound cards are to be use.
- Working in a sound-free environment is best policy. Setting up the computer in a place where there is a minimum of background noise, such as in a separate room can reduce instances where the microphone picks up and translates other speech.
- Using the same environment for testing as used for training gives highest recognition performance.
- Recognition accuracy is operating system dependent. Even the different versions of an operating system affect the system performance. Some speech recognition systems are designed to run on only one, or a narrow set, of operating systems. This can cause problems in educational establishments with a mixture of operating systems.

1.3 Classification of Speech Recognition Systems

Speech recognition systems can be classified into various categories based upon the various parameters such as the type of speech utterance, type of speaker model and the type of vocabulary that they have the ability to recognize. The speech recognition system classifications are shown in figure 1.1. Speech recognition is becoming more complex and a challenging task because of the variability in the signal. These challenges are briefly explained below.

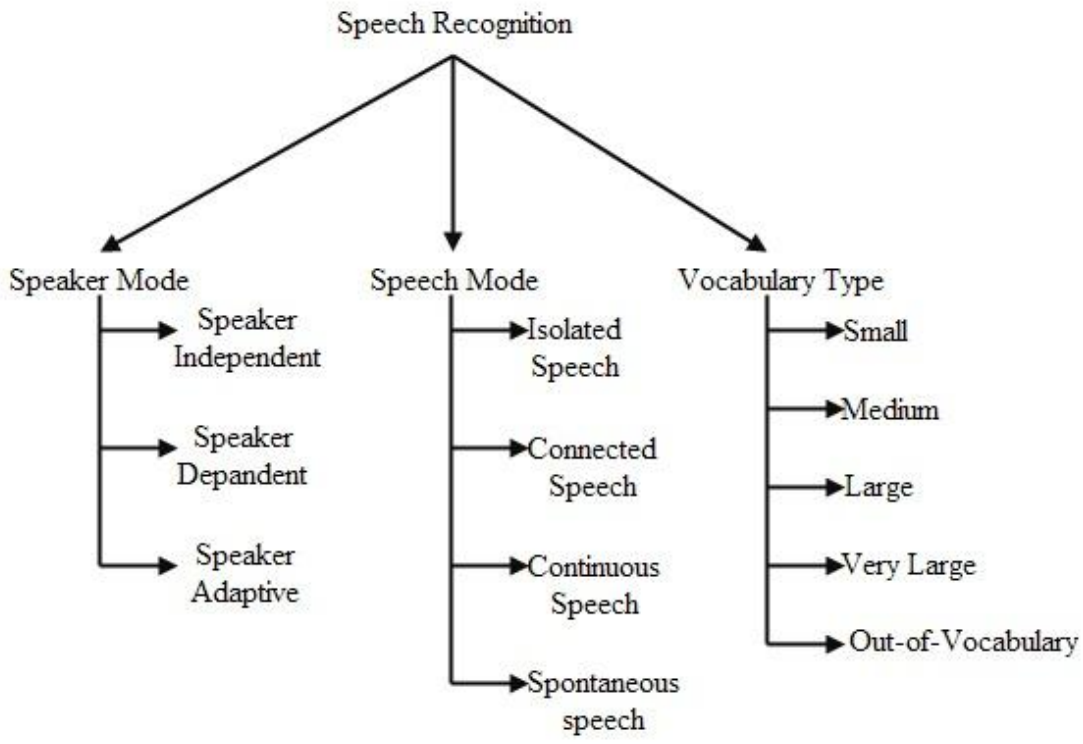


Figure 1.1: Speech Recognition System Classifications

1.3.1 Types of Speech Utterance

Based on the types of utterances they have the ability to recognize, speech recognition system can be classified into various categories such as isolated word, connected word, continuous word and spontaneous word speech recognition system. A brief of each is given below [36].

- **Isolated Word Speech Recognition System**

Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but does require a

single word at a time. As, on both sides of word there is silence (lack of an audio signal) these systems have “listen/non-listen states”. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. These types of systems are easier to construct as end points are easier to find and the pronunciation of a word does not affect others. The disadvantage of this type is choosing different boundaries affects the results.

- **Connected Word Speech Recognition System**

Connected word recognition is similar to isolated word recognition, but allows more than one word combined with the short pause between them.

- **Continuous Word Speech Recognition System**

Continuous speech recognition handles the speech in which words are connected without the pause. This speech recognition is the natural form of recognition in which users speak almost in most natural form. Recognizers which take continuous speech are somewhat difficult to create as they have to use special methods to find the start and end points of words.

- **Spontaneous Speech Recognition System**

This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and non-words. This speech recognition takes the speech in which not necessarily the words are there but also the natural speech words like “ums”, “ahs” and many more.

1.3.2 Types of Speaker Model

Speech recognition system is broadly classified into three main categories based on speaker models namely speaker dependent, speaker independent and speaker adaptive.

- **Speaker Dependent System**

Speaker dependence describes the degree to which a speech recognition system requires knowledge of a speaker’s individual voice characteristics to successfully

process speech. Speaker dependent systems are designed for a specific speaker. They are generally more accurate for the particular speaker, but much less accurate for other speakers. They assume the speaker will speak in a consistent voice and tempo. These systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems. Speaker dependent system can be used for very large vocabularies, but is limited to understanding only selected speakers.

- **Speaker Independent System**

Speaker independent systems are designed for variety of speakers. It recognizes the speech patterns of a large group of people. This system is most difficult to develop, most expensive and offers less accuracy than speaker dependent systems. However, they are more flexible. Speaker independent system generally limits the number of words in the vocabulary used for recognition system.

- **Speaker Adaptive System**

A speaker adaptive system is developed to adapt its operations to the characteristics of new speakers. It lies somewhere between speaker dependent and speaker independent systems.

1.3.3 Types of Vocabulary

The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications only require a few words, others require very large dictionaries. In speech recognition systems the types of vocabularies can be classified as follows.

- Small vocabulary - tens of words
- Medium vocabulary - hundreds of words
- Large vocabulary - thousands of words
- Very-large vocabulary - tens of thousands of words
- Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word

Apart from the above characteristics, the environment variability, channel variability, speaking style, sex, age, speed of speech also makes the speech recognition systems more complex. But the efficient speech recognition system must cope with the variability in the signal.

1.4 Basic Model of Speech Recognition

The basic model of speech recognition is shown in figure 1.2. It includes the preprocessing phase, feature extraction phase, acoustic and language model and the recognition phase.

- **Preprocessing**

Signal recorded from microphone goes to the preprocessing phase. Preprocessing phase includes transformation of input speech into a form that can be understandable by the machine. For this, input speech signal in the form of analog signal is converted into the digital speech signal. The digitized (sampled) speech signal is then processed through the first-order filters to spectrally flatten the signal. This process, known as preemphasis, increases the magnitude of higher frequencies with respect to the magnitude of lower frequencies. This process is necessary because during the sound capturing and analog to the digital conversion process certain side effects are immersed in the speech in the form of noise. The preemphasized speech is converted into the frames of subsequent samples. For this frame size ranges from 10 to 25 milliseconds and an overlap of 50% to 70% between neighboring frames. Each individual frame is windowed to minimize the signal discontinuities at the beginning and at the end of each frame.

- **Feature Extraction**

Feature extraction is the process of parameterization of the speech i.e. representation of speech utterances in terms of feature vectors, which can be used for making the acoustic models. Feature extraction is expected to discard the irrelevant information while keeping the useful one. Basically, the purpose of a feature extractor is to identify, within the data what information is needed to perform accurate classification. The speech signal contains the characteristics information of the speaker and environment in addition to signal message. A feature extractor for speech recognition

needs to maximally discard the speaker and environment information and only allow the signal message information to pass; on the other hand a feature extractor for speaker recognition needs to filter the speaker related information from the speech signal. The capability of a feature extractor for speech recognition is measured in terms of how well feature extractor can find out the actual speech signal message information.

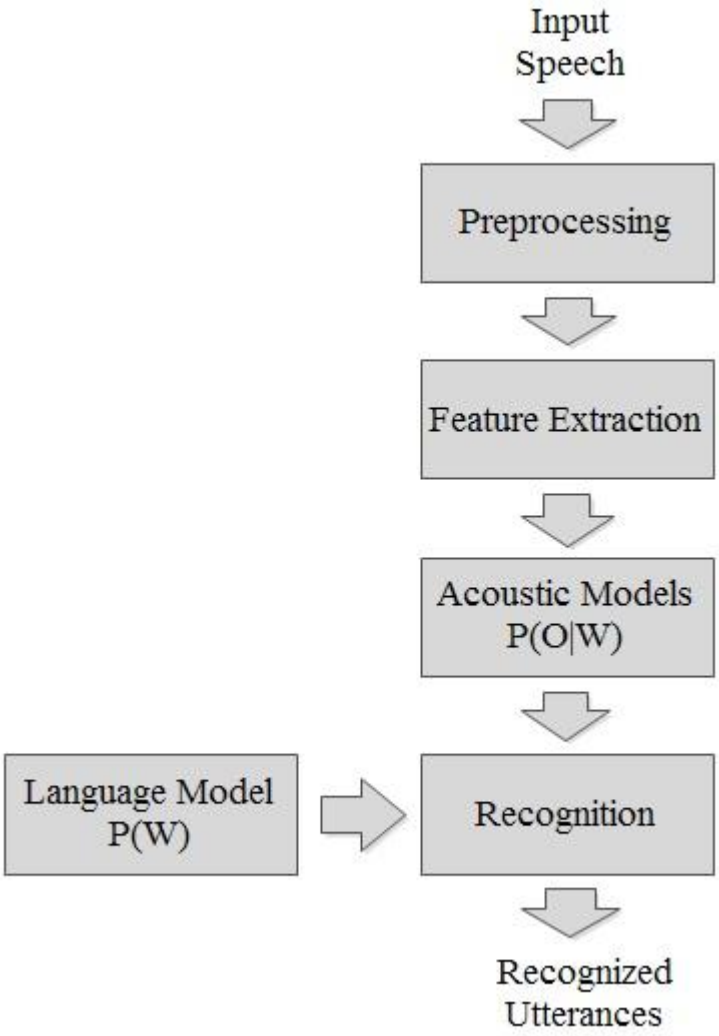


Figure 1.2: Basic Model of Speech Recognition

- **Acoustic Models**

To recognize an unknown utterance, it has to be compared with some reference models called acoustic models. Using these models, most probable sound is identified. There are two kinds of acoustic models-Word model and phoneme model. Word

model and phoneme model are used for small vocabulary system and large vocabulary system respectively. Acoustic models can be generated using various approaches such as Hidden Markov Model (HMM) [56], Artificial Neural Network (ANN) [3], Dynamic Bayesian Network, Support Vector Machine and hybrid approaches [12].

- **Language Model**

Language model incorporates both syntactic and semantic constraints of the language and recognition task. It is a file containing the probabilities of sequence of words. Humans can easily recognize the words having same phonetic structure as they know the context and also have a good idea about what words or phrases can occur in the context, but words having same phonetic structure are main hurdle to the speech recognition system. To handle this hurdle in speech recognition system, language model is used. Mainly, context is provided with the help of language model. So, language model specifies what are the valid words and in what sequence they can occur.

- **Recognition**

This is the phase where unknown utterances are recognized with the help of acoustic model generated from the trained databases and the language model.

1.5 Fundamental Problems in Speech Recognition System

The problem of automatic speech recognition (ASR) is to program a computer to take digitized speech samples and print the words that a human would recognize when listening to the same sound. There are several fundamental problems that must be overcome in any speech recognition system. These are [40]:

- Acoustic representation: How will the information in the acoustic signal be represented?
- Word representation: How will words be represented? Words are linguistic units and as such can be represented as an atomic unit in a speech recognition system. But words are also composed of syllables and phonemes and these can also be used as the atomic units.
- Training: In order to recognize unknown utterances, there is need of linkage between acoustic and word representation. But due to the great variability in

speech signal this linkage must be made by considering many examples for which the association is known. How exactly should this be done?

- Recognition: Once the training is done, how exactly can the system are used for recognition?

Below, each of the stated problems is discussed in little detail.

Acoustic Representation:

As a first step in ASR, the acoustic signal is processed to extract features that are higher level than the raw sound wave itself. A feature extractor for speech recognition discards the speaker and environment information and only allows the signal message information to pass. There are various methods for extracting the features which includes MFCC (Mel Frequency Cepstral Coefficient) [53], PLP (Perceptual Linear Prediction) [14], LPCC (Linear Predictive Cepstral Coefficient) [19], temporal patterns and many more. MFCC has generally obtained a better accuracy and a minor computational complexity with respect to alternative processing as compared to other feature extraction techniques [53].

Also, to recognize an unknown utterance, it has to be compared with some reference models called acoustic models. Acoustic models can be generated using various approaches such as Hidden Markov Model, Artificial Neural Network, Dynamic Bayesian Network, Support Vector Machine and hybrid approaches. HMM based statistical speech recognition systems have been popular in the past decade or so and have shown better results as compared to the other recognition techniques.

Word Representation:

The simplest way to represent words is atomically. In this representation, words are modeled as a whole. The advantage of representing speech at the word level is that such representations are direct and can be easily made. However, when large vocabulary recognition systems are used, using word models gives rise to problems. This is because a separate model is required for each word and often of a set of variants of each word. Consequently such systems require a large amount of training data, need a lot of training time and take up a lot of memory. Also, adding a new word to the vocabulary requires many examples of that word in order to satisfactorily train its new word model.

Another way to represent the words is sub-word units. The simplest sub-word units that can be used are syllables. Syllables are relatively intuitive and are often defined as consisting of a vowel and optional surrounding consonants. Syllables also suffer from some of the same drawbacks as whole word units. Other sub-word units that can be used are phonemes. Every word is made up of a number of phonemes and the system using phoneme representation for the representation of the words defines one acoustic model for each phoneme of the word. The acoustic model of a certain word is made up of the concatenation of the acoustic models of the phonemes that constitute the word. The advantage of representing speech at the phoneme level is that such representations allow the representation of speech signal to be less redundant than the original acoustic representation, and consequently storage requirements would be reduced. Also, adding a new word in the vocabulary is easy when using phonemes as word representation.

Training and Recognition:

The training task consists of taking a collection of utterances with associated word labels and learning an association between the specified word model and observed acoustics. When using phonemes as word representation, figuring out the phoneme boundaries is specific to the modeling approach used. A larger part of the training is to figure out, implicitly or explicitly, which word representation is to be applied to each frame of the utterance.

Recognizing phase recognizes the test samples based on the acoustic properties which are calculated in the training phase. This phase also has the same problem as in training phase, which word representation is to be associated with each frame of the utterance.

1.6 Relevant Issues of ASR Design

Developing a high quality automatic speech recognition system is really a difficult problem. The difficulty of speech processing technology can be broadly characterized along a number of dimensions [34]. The relevant issues on which recognition accuracy of ASR depend are shown in table 1.1.

Table 1.1: Relevant Issues of ASR Design

No.	Issue	Characteristics
1	Environment	Type of Noise, Signal/Noise Ratio, Working Conditions
2	Transducer	Microphone, Telephone
3	Channel	Band Amplitude, Distortion, Echo
4	Speech Styles	Voice Tone (quiet, normal, shouted), Speed (slow, normal, fast)
5	Speech Mode	Isolated, Connected, Continuous, Spontaneous
6	Speaker Mode	Speaker Dependence/Independence, Sex, Age, Physical and Psychical state
7	Vocabulary Size	Small, Medium, Large, Very Large

Environment:

The recognition task is made harder by the presence of background noise and signal distortions. The type of background noise for example, stationary, nonhuman noise versus background speech and crosstalk by other speakers also contribute to the difficulty of speech recognition system.

Microphone Characteristics:

The microphone comes in various qualities based upon performance such as good, average, below average etc. The distance between the mouth of the speaker and microphone is also a main factor for deciding the accuracy of the system.

Channel Characteristics:

Channel quality is also an important dimension. High bandwidth with full frequency range in case of human speech versus low bandwidth with limited frequency range in case of telephone speech. The latter is harder to recognize.

Speech Style:

Read speech with formal style versus spontaneous and conversational speech with casual style. The latter is harder to recognize.

Speech Mode:

Isolated words, connected words, continuous speech and spontaneous speech. Continuous speech recognition is much harder than isolated word recognition. This is because with the former, word boundaries are difficult to locate, and there are co-articulation effects between words. In addition content words are often emphasized whereas function words are often poorly articulated.

Speaker Dependency:

Speaker independent operation is considered to be one of the hardest problems in ASR, and most commercial recognizers are speaker dependent.

Size of Vocabulary:

Speech recognition becomes more difficult as the vocabulary of the speech recognizer increases for several reasons.

- Firstly, as the number of words increases, the opportunity to confuse words with similar sounding words increases. Consequently, the probability of correctly identifying a given word will often decrease with increasing vocabulary size. The nature of the fall in recognition accuracy for a given word will depend on the confusability of the vocabulary.
- Secondly, as the vocabulary becomes larger, the computational load required to make recognition will increase.
- Another important issue with large vocabularies is the difficulty posed by the labeling of the suitable databases required to train the speech recognizers.

The present state-of-the-art speech recognition systems make the problem more manageable by constraining it in a variety of ways. They demand that the recordings be carried out in good low-noise conditions and that the utterances are produced by co-operative speakers. The vocabulary is often no more than 1000 words, and in the case of continuous speech recognition, a grammar is almost always used to constrain the choice of words. Generally, ASR systems are made in noise free environment to make the system more accurate. If ASR is implemented in natural environment, it is must to identify and filter out noises from the speech signal.

1.7 Speech Recognition Approaches

Research on automatic speech recognition by machine has fascinated much attention over the past four decades. It is due to the technological curiosity about understanding the mechanisms for mechanical realization of human speech capabilities. Desire to automate simple tasks requiring human machine interactions also motivated the researchers to the work on this appealing field. In general, there are three classical approaches for recognizing the speech [29]. They are:

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Artificial Intelligence Approach

Among the three approaches, the acoustic-phonetic approach has been studied and researched more in the past 40 years. This approach is the oldest speech recognition approach originating from the 1950s. The AI approach is the youngest approach and least known. The pattern recognition approach is the most common approach and is applied in most current ASR systems.

1.7.1 Acoustic Phonetic Approach

Acoustic phonetic approach is also known as rule-based approach [55]. This approach uses knowledge of phonetics & linguistics to guide search process. In this approach, some rules are defined which express everything or anything that might help to decode based in "blackboard" architecture i.e. at each decision point it lays out the possibilities and apply rules to determine which sequences are permitted. This approach identifies individual phonemes, words, sentence structure and/or meaning.

Basically, the acoustic phonetic approach considers the existence of finite and distinctive phonetic units in spoken language and considers that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. The acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (co-articulation effect) but acoustic phonetic approach assume that the rules governing the variability are straightforward and can be readily learned by a machine.

Figure 1.3 shows the block diagram of acoustic phonetic approach to speech recognition. The first step in the acoustic phonetic approach is a spectral analysis of

the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is segmentation and labeling. In this step, system finds the feature stable regions and then labels those regions accordingly in order to match each individual phonetic unit [36]. This results in a phoneme lattice characterization of the speech. The last step is validation, which determines a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling. In the validation process, linguistic constraints on the task (vocabulary, syntax and other semantic rules) are invoked in order to access the lexicon for word decoding based on the phoneme lattice.

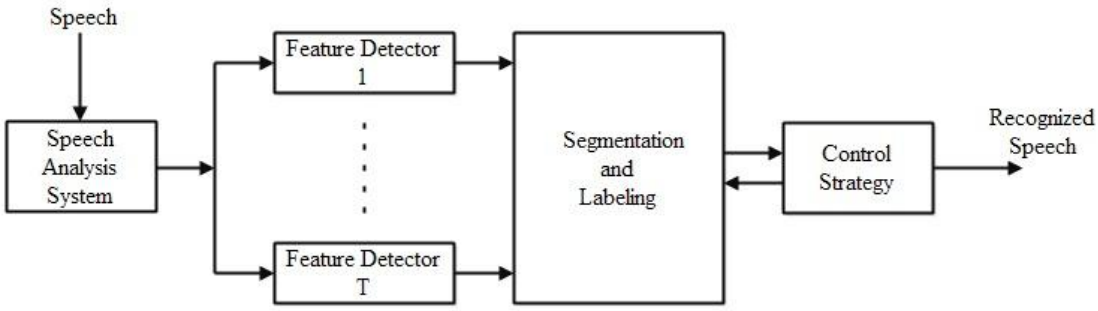


Figure 1.3: Block Diagram of Acoustic Phonetic Approach to Speech Recognition

The acoustic phonetic approach also faces problems. Some of the problems are:

- Among the difficulties of this method is need for extensive knowledge of acoustic properties of phonetic units.
- Features are often based on non-optimal ad hoc considerations rather than based on intuition.
- The choice of features is likely based on suboptimal and so optimal implementation of classification and regression tree (CART) methods is rarely achieved.
- Furthermore, there is no well-defined, automatic procedure for tuning the labeled speech.
- Moreover, no standard way in labeling the training speech.

These problems need to be solved for acoustic phonetic approach to be utilized practically. Due to these problems, acoustic phonetic approach still needs much more

research and understanding before it can successfully implemented in actual speech recognition systems.

1.7.2 Pattern Recognition Approach

In pattern-recognition approach, the speech patterns are used directly without explicit feature determination and segmentation. Pattern recognition approach mainly has two steps-training of patterns and recognition of pattern via pattern comparison [4, 29]. Pattern can be speech samples, image files etc. This approach uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm.

Figure 1.4 shows the block diagram of the pattern recognition approach to speech recognition. In the parameter measurement phase, a sequence of measurements is made on the input signal to define the test pattern. After that, a direct comparison is made between the unknown test pattern and each reference pattern using the decision rule which determines the identity of the unknown according to the goodness of match of the patterns.

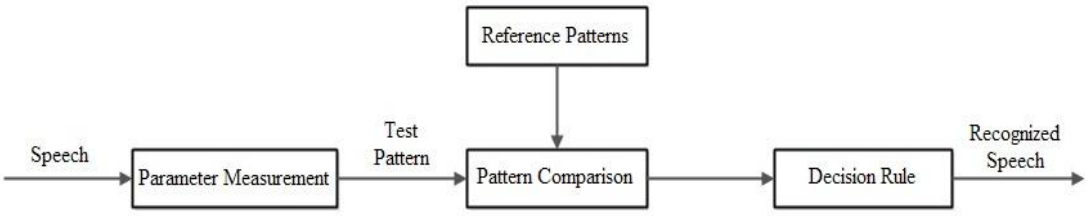


Figure 1.4: Block Diagram of Pattern Recognition Approach to Speech Recognition

Strengths of pattern recognition approach are [5]:

- This approach is insensitive to sound class so can be used with wide range of speech sound, including phrases, whole words, and sub-word units. A basic set of techniques developed for one sound class can generally be directly applied to different sound classes with little or no modifications to the algorithms.
- It is relatively straightforward to incorporate syntactic (and even semantic) constraints directly into the pattern recognition structure, thereby improving recognition accuracy and reducing computation.

The pattern recognition approach also has some problems associated with it. Few of them are:

- Performance of the system is directly dependent over the training data provided for creating sound class reference patterns. The more training, the higher the performance of the system for virtually any task.
- The reference patterns are sensitive to the speaking environment and transmission characteristics of the medium used to create the speech; this is because the speech spectral characteristics are affected by transmission and background noise.
- Computational load for pattern trained and classification is proportional to number of patterns being trained.

In spite of above stated problems, this approach has become the predominant method for speech recognition [7]. Based upon the representation of speech patterns, this approach can be divided further into two approaches.

1.7.2.1 Template Based Approach

Template based approach stores a collection of prototypical speech patterns as reference patterns which represent the dictionary of candidate words. Recognition is carried out by matching an unknown speech utterance with each of these reference templates and selecting the category of the best matching pattern [36].

In template based approach, templates for entire words are usually constructed. By doing this, errors due to segmentation or classification of phonemes can be avoided. But then, each word must have its own full reference template. Also, when vocabulary size increases beyond a few hundred words, template preparation and matching become prohibitively expensive or impractical. One solution is to derive typical sequences of speech frames for a pattern via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another solution is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker. When using entire words as templates there is a problem that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical.

1.7.2.2 Stochastic Approach

Stochastic approach can be seen as extension of template based approach, using some powerful and statistical tools and sometimes seen as anti-linguistic approach [37]. In stochastic approach, variations in speech are modeled statistically, using automatic learning procedures. Modern general-purpose speech recognition systems are based on statistical acoustic and language models. Effective acoustic and language models for ASR in unrestricted domain require large amount of acoustic and linguistic data for parameter estimation. Processing of large amounts of training data is a key element in the development of an effective ASR technology now days. Also, speech recognition usually has uncertainty and incompleteness due to the confusable sounds, speaker variability, contextual effects, and homophones words. Stochastic approach uses the probabilistic models to deal with uncertain or incomplete information [47]. Thus, stochastic approach is suitable for speech recognition.

The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model (HMM) is characterized by a finite state markov model and a set of output distributions. HMM is a popular tool for modeling a wide range of time series data. In speech recognition, HMMs have been applied with great success to problem such as part of speech classification.

Compared to template based approach, hidden Markov modeling is more general and has a firmer mathematical foundation. A template based model is simply a continuous density HMM, with identity covariance matrices and a slope constrained topology. Although templates can be trained on fewer instances, they lack the probabilistic formulation of full HMMs and typically underperform HMMs. However, the problem with using HMM is that HMMs do not provide much insight on the recognition process. As a result, it is often difficult to analyze the errors of an HMM system in an attempt to improve its performance. Also, hidden Markov modeling must make a priori modeling assumptions, which are liable to be inaccurate.

1.7.3 Artificial Intelligence Approach (Knowledge Based Approach)

The artificial intelligence (AI) approach recognizes the speech according to the way a person recognizes the speech by applying his intelligence in visualizing, analyzing and characterizing speech and making decision on the acoustic knowledge. The basic

idea of artificial intelligence approach to speech recognition is to compile and incorporate knowledge from a variety of knowledge sources and to apply it on the problem at hand. The different types of knowledge sources are [29]:

- **Acoustic Knowledge:** Evidence of which sounds are spoken on the basis of spectral measurements and presence or absence of features.
- **Lexical Knowledge:** The combination of acoustic evidence so as to postulate words as specified by a lexicon that maps sounds into words (or equivalently decomposes words into sounds).
- **Syntactic Knowledge:** The combination of words to form grammatically correct strings (according to a language model) such as sentences or phrases.
- **Semantic Knowledge:** Understanding of the task domain so as to be able to validate sentences that are consistent with the task being performed or which are consistent with previously decoded sentences.
- **Pragmatic Knowledge:** Inference ability necessary in resolving ambiguity of meaning based on ways in which words are generally used.

The Artificial Intelligence approach is combination of the acoustic phonetic approach and pattern recognition approach. This approach uses the information regarding linguistic, phonetic and spectrogram. Normally, there are three alternative ways often used in AI speech recognition system [29].

- Bottom Up Approach
- Top Down Approach
- Blackboard Approach

Bottom Up Approach: In the standard bottom up processor as shown in figure 1.5, the lower-level processes are applied before the higher-level processes. Lower level processes include feature extraction and phonetic decoding and higher level processes include lexical decoding and language model.

Top Down Approach: Another alternative way is top down process. The processor integrates the word hypothesis matching, lexical decoding and syntactic analyses blocks into a consistent framework as shown in figure 1.6.

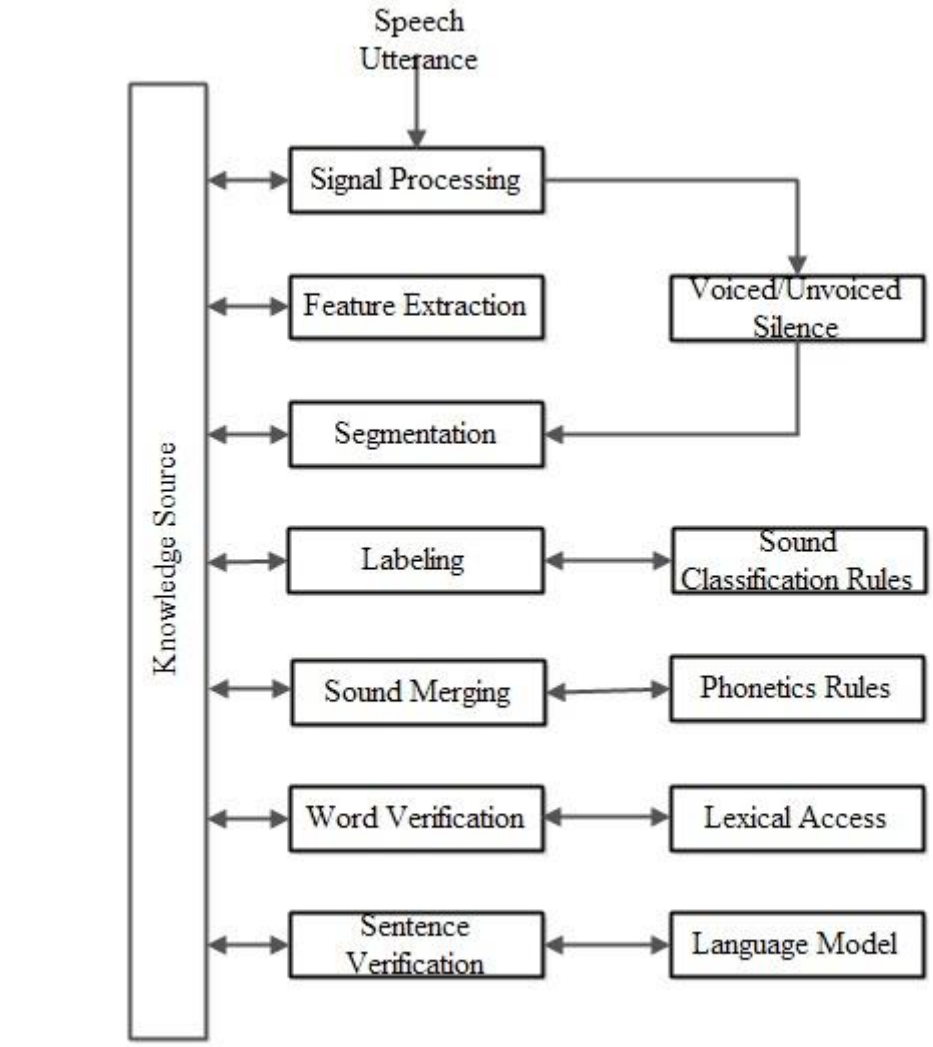


Figure 1.5: Bottom Up Approach to Knowledge Integration

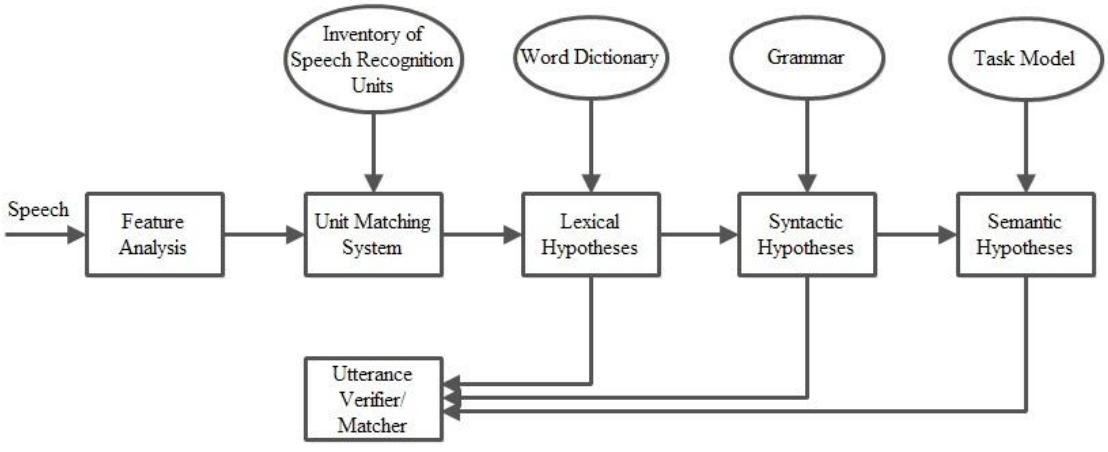


Figure 1.6: Top Down Approach to Knowledge Integration

Blackboard Approach: Figure 1.7 shows the third alternative way, blackboard approach. In this approach all knowledge sources (KS) are considered independent. A hypothesis-and test paradigm is applied as the main communication medium among KSs that are data driven and based on the patterns on the blackboard. The system operates asynchronously and assigned cost and utility considerations are distributed across all levels. The approach was extensively studied at Carnegie Mellon University (CMU) in the 1970s and it has been further researched for dialogue-based expert systems especially at Massachusetts Institute of Technology (MIT).

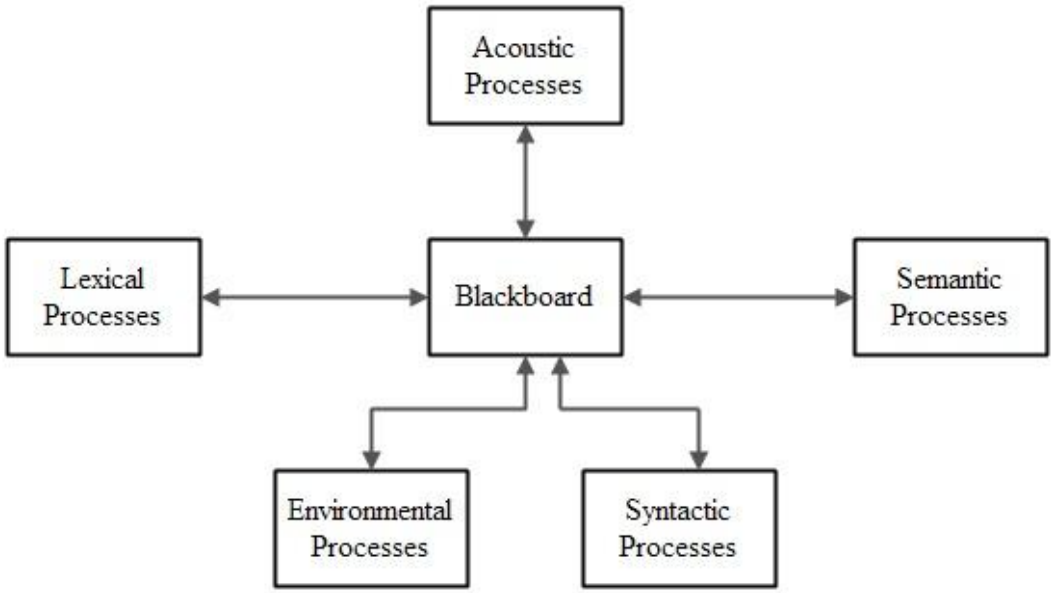


Figure 1.7: Blackboard Approach to Knowledge Integration

While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. Basically, AI approach involves the direct and explicit incorporation of expert speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. However, this approach had only limited success, due to the many reasons.

- The main reason is difficulty in quantifying expert knowledge.
- Another reason is there is difficulty in the integration of many levels of human knowledge phonetics, lexical access, syntax, semantics and pragmatics.

1.8 Speech Recognition Methods

There are many types of techniques that can be applied to recognize the speech. Among them Hidden Markov Model (HMM), Neural Network (NN), Vector Quantization (VQ), Dynamic Time Warping (DTW) and Support Vector Machine (SVM) are commonly the classical method for speech recognition.

1.8.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is one of the common methods in speech recognition systems. It is one of the oldest, methods since 30 years ago and most important algorithms by matching the unknown speech input template to a pre define reference template in speech recognition [15]. Dynamic time warping (DTW) approach is used for measuring similarity between two patterns which may vary in time or speed [36]. For example, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW can be applied to video, audio and graphics. To cope with different speaking speeds during the recognition of speech, DTW is a best approach. In general, DTW is a method that allows a computer to find an optimal match between two given sequences with certain restrictions. One example of the restrictions imposed on the matching of the sequences is on the monotonicity of the mapping in the time dimension.

DTW stretches and compresses various sections of utterance so as to find alignment that result in best possible match between template and utterance on frame by frame basis. Frame is the short segment of speech signal which is basis of parameter vector computation and match defined as sum of frame-by frame distances between template and input utterance [54]. Template with closest match is chosen as recognized word.

Traditionally, DTW algorithm needs long processing time and large pattern storage, which become a major problem for real time application as the number of speech patterns increases. As a result, DTW is widely used in the small scale speech recognition systems. For example, it provides a good recognition performance in small vocabulary, isolated word and speaker dependent. Also, DTW approach is limited to word template.

Compared to HMM, DTW required high computational requirement. HMM can capture the statistical characteristics of word and sub-word units among different speakers even in large vocabulary and thus it is better than DTW in speaker independent large vocabulary speech recognition. The DTW techniques have been generally superseded by the more powerful and flexible HMM models.

Continuity is less important in DTW than in other pattern matching approaches. Also, DTW is particularly suited to matching sequences with missing information, provided there are long enough segments for matching to occur. The optimization process is performed using dynamic programming, hence the name.

1.8.2 Vector Quantization

Quantization is an important aspect of data compression. The purpose of data compression is to reduce the bit rate to minimize communication channel capacity or digital storage memory requirements while maintaining the necessary fidelity of the data.

Vector quantization (VQ) is a classical quantization technique. It divides a large set of vectors into multiple groups. Each group is represented by its own centroid point and codebook is composed of set of all code vectors (centroid) while set of encoding regions form the space partitions. This approach is often applied to ASR [47]. It is useful for speech coders (efficient data reduction). The utility of VQ lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. In isolated word recognition system, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure. In basic VQ, codebooks have no explicit time information since codebook entries are not ordered and can come from any part of the training words.

1.8.3 Artificial Neural Networks (Connectionist Approach)

Connectionist approach of speech recognition is the youngest development in speech recognition and still the subject of much controversy. In connectionist models, knowledge or constraints are not encoded in individual units, rules or procedures, but distributed across many simple computing units. Uncertainty is modeled not as

likelihoods or probability density functions of a single unit, but by the pattern of activity in many units. The computing units in this approach are simple in nature and knowledge is not programmed into any individual unit function; rather, it lies in the connections and interactions between linked processing elements [48]. The simplicity of processing elements makes connectionist models attractive for hardware implementation, which enables the operation of a net to be simulated efficiently. On the other hand, training often requires much iteration over large amounts of training data, and can, in some cases, be prohibitively expensive. As, the style of computation that can be performed by networks of such units bears some resemblance to the style of computation in the nervous system, connectionist models are also referred to as neural networks or artificial neural networks. The basic and main feature of artificial neural network (ANN) is its capability of learning by gaining strength and properties of inter-neuron connections.

ANN in speech recognition is done as the following steps:

- During recognition process, input speech is recorded.
- MFCCs are used to extract the speech signal.
- Get the highest probabilities of feature vector into the neural network.
- Based on that, neural network able to classify the unknown digit.

Generally, neural network has three layers, the input nodes at input layer, hidden nodes at hidden layer and output nodes at output layer. One or more hidden layers of nodes is located between the input and output nodes. The network topology allows input from the input layer to the first hidden layer then from the first hidden layer to the second and until it reach at the last hidden layer to the output layer. The structure of the neural network is usable where it provides adequate training data (input nodes) and hidden nodes.

Neural networks have many similarities with Markov models. Both are stochastic models which are represented as graphs. Where Markov models use probabilities for state transitions, neural networks use connection strengths and functions. Not unlike stochastic models, connectionist models rely critically on the availability of good training or learning strategies. Connectionist learning seeks to optimize or organize a network of processing elements. A key difference is that neural networks are

fundamentally parallel while Markov chains are serial. Also, connectionist models need not make assumptions about the underlying probability distributions.

1.8.4 Support Vector Machine

Support Vector Machine (SVM) is one of the powerful state-of-the art classifiers for pattern recognition which uses a discriminative approach [12]. Optimized margin, between the samples and the classifier border, helps to generalize unseen patterns. SVMs use linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The variable length data has to be transformed to fixed length vectors before SVMs can be used [36].

Conventional statistical and Neural Network methods control model complexity by using a small number of features. SVM controls the model complexity by controlling the VC dimensions of its model. This method is independent of dimensionality and can utilize spaces of very large dimensions spaces, which permits a construction of very large number of non-linear features and then performing adaptive feature selection during training.

1.8.5 Hidden Markov Model

The Hidden Markov Models (HMMs) are most widely used statistical tools in recognition system. It covers from isolated speech recognition to very large vocabulary unconstrained continuous speech recognition and speaker identification fields. Therefore most of the current speech recognitions are conducted based on Hidden Markov Model (HMMs). HMMs have found useful in many areas of signal processing and in particular speech processing.

Template based approach of speech recognition (e.g. dynamic time warping) directly compare the unknown utterance to known examples. Instead HMM creates stochastic models from known utterances and compares the probability that the unknown utterance was generated by each model. HMMs are broad class of doubly stochastic models for non stationary signals that can be inserted into other stochastic models to incorporate information from several hierarchical knowledge sources [11]. One of the examples of the Hidden Markov Model is normally use in pattern application.

1.9 Hidden Markov Model (HMM)

The Hidden Markov Model was first described by Ruslan L. Stratonovich in 1960. It was first used as a basic concept of speech recognition in the mid of 1970's. It is a statistical model in which the system which is being modeled is assumed to be a markov process. The HMM is basically a stochastic finite set of states where each state is associated with a probability distribution. The transitions among these states depend upon a set of probabilities called transition probabilities. In a particular state, an outcome can be generated according to its associated probability distribution. It is only the outcome which is visible to an external observer. The states are not visible to the external observer and hence its name is Hidden Markov Model. HMM based statistical ASR gives better results as compared to the other recognition techniques.

There are many speech recognition based HMM's method system. At the present times, left-to-right models are widely used for the speech model topology. Besides, it also provides a more rigid temporal structure where only transitions from left to right are allowed. Left-to-right HMM model is shown in figure 1.8.

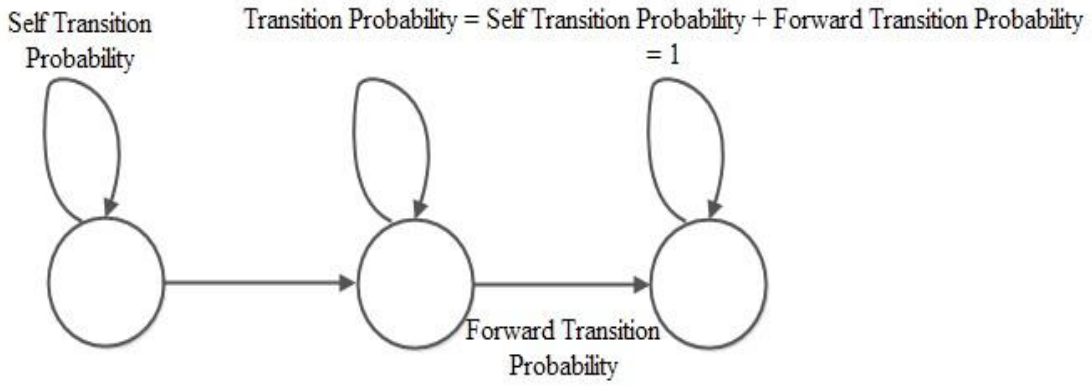


Figure 1.8: Left-to-right HMM Model

For the left to right HMM model topology, the constraint is imposed so that the first frame of speech is allocated to the first state of the model and the last frame of speech is allocated to the last state of the model. There are two probabilities for each frame. The frame can either be allocated to itself (a self transition) or it can be allocated to the next state. To improve the modeling of the temporal structure of speech signal, each state has the probabilities for these two types of transition, known as the transition probabilities. The two transition probabilities always sum to one.

1.9.1 HMM Architecture Type

HMM have several density architectures which are discrete HMM (DHMM) and continuous density HMM (CDHMM). The architecture of these models has its own advantages and disadvantages.

1.9.1.1 Discrete Hidden Markov Model

Discrete Hidden Markov Model (DHMM) is a type of HMM that model speech signals based on VQ technique to produce the speech observations. It is a combination of vector quantization and hidden Markov modeling.

In 1982, Rabiner and his colleagues was successful applied DHMM speech recognizer for a speaker independent isolated digit recognition task. Since then, research progress is made from isolated word recognition until large vocabulary recognition systems such as SPHINX system.

The main advantages of DHMM are its concise concepts and low computational costs. Compare to DTW, DHMM is more efficient and reliable technique. In DHMM, the VQ computation depends on the codebook size and computes the discrete output probability of an observation based on a lookup table. In addition, DHMM can easily model the phonemes and sub words. DHMM uses VQ codebook to represent the speech spectral vectors and creates an inherent spectral distortion in representing the actual analysis vector. The size of codebook can cause the quantization errors. On the other hand, large codebook size cause less training data for each codeword and therefore affects the recognition performance.

1.9.1.2 Continuous Density Hidden Markov Model

Continuous density Hidden Markov Models (CDHMM) models the acoustic observation by directly using estimated continuous probability density function (PDF) which is typically a mixture of Gaussian functions, without VQ.

In this model, parameters must be estimated for each state of each model. Thus, it is able to eliminate the VQ error effect of DHMM and increase the recognition accuracy. Although CDHMM requires longer training and recognition time it is a powerful model of acoustic variability with its Gaussian mixture densities.

1.9.2 Elements of Hidden Markov Model

HMM is characterized by following

1. Number of state N: Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to set of states of the model.

2. Number of distinct observation symbol per state M: If discrete observation densities are used, the parameter M is number of classes or cells that should be used and if continuous observation densities are used, the parameter M is represented by the number of mixtures in every state.

3. State transition probability distribution $A = [a_{ij}]$ where:

$a_{ij} \geq 0$ and $\sum^{N}_{j=1} aij = 1$

4. Observation symbol probability distribution in state j, $B_j (K) = P [V_k \text{ at } t \mid qt = S_j]$

5. The initial state distribution $\pi = \pi_i$ where $\pi_i = P [q_1 = S_i]$ and $1 \leq I \leq N$

A complete specification of an HMM requires two model parameters N and M. The specification of the three sets of probability measures π, A and B are also necessary. Consider λ as the notation for the probability measures, then

$\lambda = (A, B, \pi)$

1.9.3 Three Problems of HMM

There are three basic problems that must be solved for the model to be useful in speech recognition system. These problems are the following:

- **Problem 1: Evaluation**

Given:

Model $\lambda = (A, B, \pi)$ and

Testing observation sequence for $O = O_1, O_2, O_3, \ldots, O_T$

Action:

The probability of the observation sequence given the model, P (O | λ)

This is an evaluation problem which find the probability of producing a given observation O by a given model λ. The solution of problem is allowed to find the best model among multiple solution given the observation for the purpose of classification and recognition.

- **Problem 2: Decoding**

Given:

Model $\lambda = (A, B, \pi)$ and

Testing or training observation sequence $O = O_1, O_2, O_3, ..., O_{T-1}, O_T$

Action:

Track the optimum state sequence $Q = q_1, q_2, q_3, ..., q_{T-1}, q_T$ that most likely produce the given observations, using the given model.

This purpose of the decoding problem is to find the most likely sequence. The decoding procedure allows detecting the state sequence of a given observation.

- **Problem 3: Training**

Given:

Model $\lambda = (A, B, \pi)$ and

Testing observation sequence for $O = O_1, O_2, O_3, ….., O_T$

Action:

Adjust the model parameters $\lambda = (A, B, \pi)$ and maximize $P (O | \lambda)$

The purpose of training problem (estimation problem) is to optimize model parameter so as to best describe as to how given observation sequence comes out. The observation sequence used here are called training sequence since it is used for training HMM. The training problem is the crucial one for most applications of HMMs, since it allows to optimally adapting model parameters to observed training data i.e. to create best models for real phenomena.

1.9.4 Solution to Three HMM Problems

- **Problem 1**

Problem 1 is on evaluating how well a given model matches a given observation sequence. A most straightforward way to determine P (O | λ) to find P (O | I, λ) for fixed state sequence I = i_1, i_2, …, i_T is to multiply it by P (I | λ) and then sum up over all possible I's.

For

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

$P(I \mid \lambda) = \pi_{i1}\, a_{i1i2}\, a_{i2i3} \ldots a_{iT-1iT}$

Therefore

$P(O \mid \lambda) = \Sigma_I\, P(O \mid I, \lambda)\, P(I \mid \lambda)$

$P(O \mid \lambda) = \Sigma_I\, \pi_{i1}\, b_{i1}(O_1)\, a_{i1i2}\, b_{i2}(O_2)\, a_{i2i3} \ldots b_{iT}(O_T)\, a_{iT-1iT}$

where I = i_1, i_2, …, i_T

This calculation is computationally unfeasible, even for small value of N and T. Thus a more efficient procedure is required to solve this problem. The forward algorithm which is a dynamic algorithm is an efficient method to solve this problem.

- **Problem 2**

Problem 2 is on decoding to find the optimal sequence associated with the given observation sequence, I that will maximize P (O, I | λ). The famous algorithm to solve this is called Viterbi algorithm. This algorithm is able to keep possible state sequence for each of the N states as the intermediate state for the desired observation sequence O = O_1, O_2, …, O_T. This is able to find the best path for each of the N states as the last state for the desired observation sequence.

- **Problem 3**

Problem 3 is to adjust parameters (A, B, π) to maximize the probability of the observation sequence given the model. It deals with training the HMM such that it encodes the observation sequence in such a way that it includes many characteristics

similar to the given one be encountered later it should be able to identify it. The famous algorithm to solve this problem is Baum-Welch (BW) algorithm. Baum-Welch algorithm increases P (O | λ) until a maximum value is reached. This optimization criterion is called the maximum likelihood criterion. The function P (O | λ) is called the likelihood function.

1.10 HMM-based Speech Recognition System

Speech recognition is the way of converting spoken speech into the text. Speech recognition system takes an utterance in the form of wave signal as input and converts that into a text sequence similar to the information being conveyed by the input data. There are many methods for recognizing the speech which are described in the previous section. Currently the use of Hidden Markov Models (HMMs) is the most popular statistical approach for automatic speech recognition [28, 49]. Hidden Markov model has been used in some form or another in virtually every start-of-the-art speech and speaker recognition system [46]. Figure 1.9 shows schematic diagram of a HMM-based speech recognition system. The main components of a speech recognition system include front-end processing model, acoustic model, language model and Viterbi decoder.

Speech recorded from microphone is rarely pristine. It contains not only the speech data but also background noise. This noise can interfere with the recognition process, and the speech engine must handle the environment within which the audio is spoken. Its first job is to process the incoming audio signal and convert it into a format best suited for further analysis. This initial stage of speech recognition is referred to as front-end processing. In front-end processing, signal is also passed through a filter which emphasizes higher frequencies. This process is known as preemphasis which will increase the energy of the signal at higher frequency [24]. The Pre-emphasis of the speech signal is realized with this simple filter $H(z) = 1 - a z^{-1}$ where a is from interval [0.9, 1]. Then speech signal is segmented, usually 20-30ms with 10ms overlap between adjacent segments. Each short speech segment is windowed and a Discrete Fourier Transform (DFT) is thereafter applied. The resultant complex spectral coefficients are taken the modulus and logarithm to produce the log power spectrum. Mel frequency averaging is then applied, followed by a Discrete Cosine Transform (DCT) to produce a set of Mel Frequency Cepstral Coefficients (MFCC).

MFCCs are widely used as the acoustic feature vectors in speech recognition. However, they do not account for the underlying speech dynamics which cause speech changes. A simple way to capture speech dynamics is to calculate the first and sometimes the second time derivatives of the static feature vectors and concatenate the static and dynamic feature vectors together. Although the use of dynamic features improves recognition performance, the independence assumption actually becomes even less valid because the observed data for any one frame are used to contribute to a time span of several feature vectors.

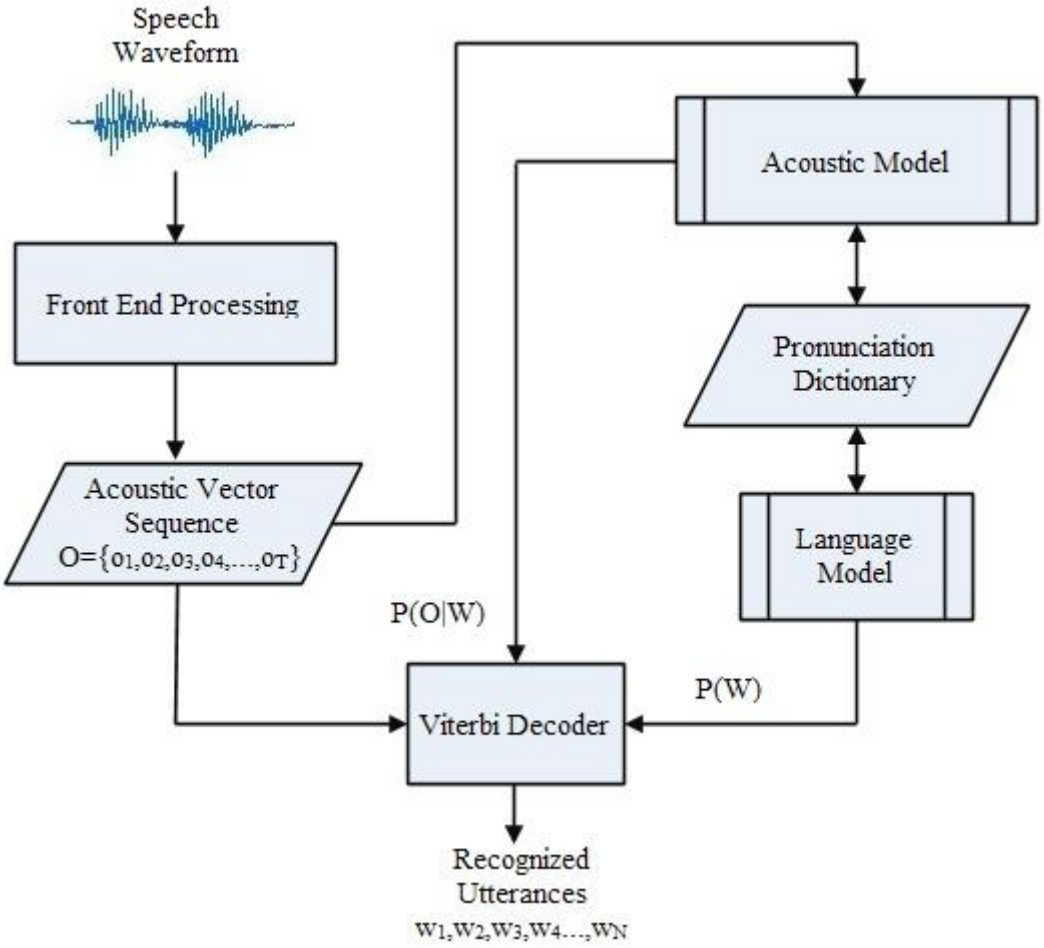


Figure 1.9: Schematic Diagram of HMM-based Speech Recognition System

Given a sequence of acoustic vectors $O = \{o_1, o_2, o_3, o_4, ..., o_T\}$, the speech recognition task is to find the optimal words sequence $W = \{w_1, w_2, w_3, w_4, ..., w_N\}$ which maximizes the probability $P(W \mid O)$ i.e. the probability of the word sequence W given the acoustic feature vectors O. This probability is normally computed by using Baye's theorem:

$$P(W \mid O) = \frac{P(O \mid W)\, P(W)}{P(O)}$$

In the equation above P (W) is the probability of the word sequence W, which is language model probability and computed using a language model; and the probability P (O | W) is referred to as acoustic model probability, which is the probability that the acoustic feature vector O is produced by the word sequence W. P (O | W) is calculated by using acoustic model. P (O) is ignored as it is constant with respect to W.

The pronunciation dictionary specifies the sequence of HMMs for each word in the vocabulary. The probabilities P (O | W) and P (W) are then multiplied together using the Viterbi decoder for all possible word sequences allowed by the language model. The one with the maximum probability is selected as the optimal word sequence.

1.11 Obstacles in Using Speech Recognition System

Most of the history of speech recognition systems has been in making a trade-off between what the user can say or speak, and what the technology interprets, that is of an acceptably high level of accuracy to the end-user. For the past 40 years, speech recognition systems have failed to match the acceptable accuracy. Also, there is a lot of research has been carried out during these 40 years, but we are far from achieving the goal of a robust speech recognizer, which can understand spoken words on any subject by all speakers in any environment. Performance of current speech recognition systems is far below human performance. Following are some of the obstacles in using speech recognition.

- Humans use more than their ears when listening; they use the knowledge they have about the speaker and the subject. But how the speech recognition system can model the world knowledge, the knowledge of the speaker and encyclopedic knowledge; also world knowledge can’t be modeled exhaustively.
- Humans do not communicate only with speech, but also with body signals like with hand waving, eye movements, postures etc. This information is completely missed by ASR.

- Humans also communicate their emotions via speech. We speak differently when we are happy, sad, frustrated, stressed, disappointed, defensive etc. If we are sad, we may drop our voice and speak more slowly, and if we are frustrated we may speak with a more strained voice.
- Normally, speech is uttered in an environment of sounds, a clock ticking, another human speaker in the background etc. This is usually called noise i.e. unwanted information in the speech signal. Generally, ASR systems are made in noise free environment to make the system more accurate but either creating 100% noise free environment is not possible or costlier so adding extraneous cost to the system cost. Also recognition system made in noise free environment is not a real world application. If ASR is implemented in natural environment, it is must to identify and filter out noises from the speech signal.
- Another important obstacle is influences in speech e.g. normal speech is filled with hesitations, repetitions, changes of subject in the middle of an utterance, slips of the tongue etc. A human listener does not even notice these things but this kind of behavior has to be modeled by the ASR system.
- All speakers have their special voices, due to their unique physical body and personality. The voice is not only different between speakers; there are also wide variations within one specific speaker. Difficulties in the vocal tract sizes among individual speakers contribute to the variability of speech.
- Across the world there are several thousand of languages which are spoken. Within those languages, there are wide variations in the syntax and vocabulary used. People in different regions speak the same language but with different accents.
- Men and women have different voices and the main reason to this is that women have in general shorter vocal tract than men. The fundamental tone of women's voices is roughly two times higher than men's voices.
- We speak at different speeds, tones and pitches. We also speak in different speeds if we talk about something known or something unknown.
- Natural language has an inherent ambiguity i.e. we cannot always decide which of a set of words is actually intended. This is, of course, a problem in every computer-related language application. The main ambiguity particular to ASR is homophones. **Homophones also known as sound-alike words** are

words that are pronounced identically although they have different meanings and have different spellings as well. These words are a very common source of confusion. Common examples of sets of homophones include: to, too, and two; they're and their; bee and be; sun and son; which and witch; plain and plane; sail and sale; addition and edition etc.

The obstacles do pose some problems, but do not prevent the design and development of applications that use voice commands.

1.12 Applications of Speech Recognition

There are a number of scenarios where speech recognition is either being delivered, developed for researched or seriously discussed. Speech recognition system has a lot of applications in various fields.

1.12.1 Electronic Health Records

The rush to adopt Electronic Health Records (EHR) is on. Practices around the country are faced with the challenge of getting their physicians to document patient care in the EHR. Now, EHR systems are actually implemented using speech recognition technology. Speech recognition allows physicians to capture anywhere their observations, assessments, and plans for ready access within the EHR. Speech-driven EHR has lot of benefits in health care system.

- **Reduced Transcription Expense**

EHR system driven by speech enables clinicians to dictate substantial sections of the medical record in 'free-text' directly into the EHR, using their own words, without having to rely on transcription. Speech-driven EHR system reduces or eliminates the ongoing cost of transcription by providing physicians greater flexibility to document findings.

- **Dramatically Increased Physician Productivity**

Speech recognition system reduces time-on-documentation by as much as 50%, freeing up the physician to spend more time with patients.

- **Improved Patient Care**

Patient notes created via speech contain deeper and more descriptive information also treatment plans are formulated more rapidly, reducing the chance of adverse medical effects.

- **Increased Cash Flow and Revenue**

There is also the impact of speech recognition on clinical workflow and quality of care and found substantial opportunities to maximize reimbursement per physician using speech recognition with an EHR.

1.12.2 Education

There are number of areas in education sector where speech recognition can be used.

- Speech recognition can be used in making notes of observations during scientific experiments, so the scientist/research can focus on the observation without needing to view the monitor or keyboard.
- Speech recognition enables students who are physically handicapped and unable to use a keyboard to enter text verbally.
- Speech recognition system can be used as restrictive access on a high security computer, where a keyboard or other input device may be used by hackers.
- It can be used in narrative-oriented research, where transcripts are automatically generated. This would remove the time to manually generate the transcript, and human error.
- Capturing the speech of a lecturer or tutor in text format can be done by using speech recognition system.
- Speech recognition system may also be used in an examination.

1.12.3 Aircrafts

Speech recognition system can be widely used in aircraft systems. Speech recognizers have been operated successfully in fighter aircraft with applications including: setting radio frequencies, commanding an autopilot system, setting steer-point coordinates and weapons release parameters, and controlling flight displays.

- Speech recognition has definite potential for reducing pilot workload.

- Achievement of very high recognition accuracy is most critical factor for making the speech recognition system useful, with lower recognition rates, pilots would not use the system.
- More natural vocabulary and grammar, and shorter training times would be useful, but only if very high recognition rates could be maintained.

The Speech recognition system in aircraft is a life critical system. So, its accuracy can't be compensated to anything at any cost.

1.12.4 Telephony Environment

The most widespread application of speech recognition technology is in telephone-based information retrieval systems. This is almost a natural development, as telephones take speech input anyway. At the most basic level, some mobile phones offer the facility to select a phone number from the in-phone directory by saying the name associated with it; the phone then dials the stored number automatically. At a more useful level, speech recognition is increasingly used in automated telephone-based interactive services. For example, it is possible to check the weather forecast, the price of a stock market share, or book a flight using an increasing number of these services. Speech recognition technology is being tentatively used, and researched, in the car industry. There is hands-free use of mobile phone handsets in the car, speech instructions to navigation systems, in-car system interaction, in-car steering systems and many more. Though there are advantages for speech recognition in cars, there are considerable obstacles in terms of in-car noise, and vocal interference from passengers who are inches/feet away.

1.12.5 People with Disabilities

Speech recognition technology helps people with disabilities interact with computers more easily. People with motor limitations, who cannot use a standard keyboard and mouse, can use their voices to navigate the computer and create documents. Some individuals with speech impairments may use speech recognition as a therapeutic tool to improve vocal quality. People with overuse or repetitive stress injuries also benefit from using speech recognition to operate their computers hands free. Speech recognition technology has great potential to provide people with disabilities greater access to computers and a world of opportunities.

1.12.6 Computer

Speech recognition systems can also be used in computers for writing text documents. It can be used for opening, closing and operating various applications in computers. For example, by speaking open my computer opens my computer window. So, computer can be operated by speech input.

1.12.7 Gambling

Online gambling has become a major industry in the last four years. Speech recognition has application in games such as online poker (multiplayer), where vocal commands can be both heard by the other players, and are (where appropriate) interpreted by the host computer in order to deal more cards, adjust the money staked and so forth.

1.12.8 Precision Surgery

There is occasional speculation in various medical for a regarding the use of speech recognition in precision surgery, where a procedure is partially or totally carried out by automated means. For example, in removing a tumor or blockage without damaging surrounding tissue, a command could be given to make an incision of a precise and small length. However, the legal implications of such technology are a formidable barrier to significant developments in this area.

1.12.9 Domestic Applications

Speech recognition is used in domestic appliances such as ovens, refrigerators, dishwashers and washing machines. The main reason of using speech recognition system in these appliances is that it can reduce the number of parts and therefore the cost of production of the appliance. However, removal of the normal buttons and controls would present problems for people who, for physical or learning reasons, cannot use speech recognition systems.

1.12.10 Wearable Computers

Speech recognition can also be used in wearable computers i.e. unobtrusive devices that can be wear like a watch, or are even embedded in clothes. These would allow people to go about their everyday lives, but still store information (thoughts, notes, to-

do lists) verbally, or communicate via email, phone or videophone, through wearable devices. Crucially, this would be done without having to interact with the device, or even remember that it is there; the user would just speak, the device would know what to do with the speech, and would carry out the appropriate task.

In this chapter, the basics of a speech recognition system have been discussed. Along with it, methods for increasing the accuracy and ease of use of speech recognition systems, the types and categories of the speech recognition system, basic model of speech recognition system, fundamental problems and issues of ASR design, various approaches and methods to speech recognition have been introduced. Also, Hidden Markov Model and HMM-based speech recognition system have been elaborated to get basic understanding of model used in this thesis. Finally, the obstacles in using SR system and applications of the proposed system have been discussed. In the next chapter, the review of literature of the proposed system has been discussed.

Chapter 2

Literature Review

This chapter provides an historic perspective on key inventions that have enabled progress in speech recognition and briefly reviews several technology developments as well.

Attempts to develop machines to mimic a human's speech communication capability appear to have started in the 2nd half of the 18th century. The early interest was not on recognizing and understanding speech but instead on creating a speaking machine, perhaps due to the readily available knowledge of acoustic resonance tubes which were used to approximate the human vocal tract. In 1773, the Russian scientist Christian Kratzenstein, a professor of physiology in Copenhagen, succeeded in producing vowel sounds using resonance tubes connected to organ pipes.

2.1 Early Automatic Speech Recognizers

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the phonetic elements of speech and tries to explain how they are acoustically realized in a spoken utterance. These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts.

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

In other early recognition systems of the 1950's, Olson and Belar of RCA Laboratories built a system to recognize 10 syllables of a single talker and at MIT Lincoln Lab, Forgie and Forgie built a speaker-independent 10-vowel recognizer [13].

In the 1960's, several Japanese laboratories demonstrated their capability of building special purpose hardware to perform a speech recognition task. Most notable were the vowel recognizer of Suzuki and Nakata at the Radio Research Lab in Tokyo [21], the

phoneme recognizer of Sakai and Doshita at Kyoto University [13], and the digit recognizer of NEC Laboratories [22]. The work of Sakai and Doshita involved the first use of a speech segmental for analysis and recognition of speech in different portions of the input utterance.

In another early recognition system Fry and Denes, at University College in England, built a phoneme recognizer to recognize 4 vowels and 9 consonants. By incorporating statistical information about allowable phoneme sequences in English, they increased the overall phoneme recognition accuracy for words consisting of two or more phonemes. This work marked the first use of statistical syntax (at the phoneme level) in automatic speech recognition.

In the 1960's Tom Martin at RCA Laboratories and Vintsyuk in the Soviet Union develop a realistic solution to the problems associated with the temporal non-uniformity in repeated speech events and suggested a range of solutions, including detection of utterance endpoints, which greatly enhanced the reliability of the recognizer performance. Since the late 1970's, mainly due to the publication by Sakoe and Chiba has become an indispensable technique in automatic speech recognition [15].

2.2 Research Work since 1970's

In the late 1960's, Atal and Itakura independently formulated the fundamental concepts of Linear Predictive Coding (LPC) [10]. By the mid 1970's, the basic ideas of applying fundamental pattern recognition technology to speech recognition, based on LPC methods, were proposed by Itakura, Rabiner and Levinson and others. Also during this time period, based on his earlier success at aligning speech utterances, Tom Martin founded the first speech recognition commercial company called Threshold Technology, Inc. and developed the first real ASR product called the VIP-100 System.

Among the systems built by the contractors of the Advanced Research Projects Agency (ARPA) program was Carnegie Mellon University's "Harpy" which was shown to be able to recognize speech using a vocabulary of 1,011 words, and with reasonable accuracy. One particular contribution from the Harpy system was the concept of doing a graph search, where the speech recognition language was

represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules.

In parallel to the ARPA-initiated efforts, two broad directions in speech recognition research started to take shape in the 1970's, with IBM and AT&T Bell Laboratories essentially representing two different schools of thought as to the applicability of automatic speech recognition systems for commercial applications.

IBM's effort, led by Fred Jelinek, was aimed at creating a "voice-activated typewriter" (VAT), the main function of which was to convert a spoken sentence into a sequence of letters and words that could be shown on a display or typed on paper. The technical focus was on the size of the recognition vocabulary and the structure of the language model that could appear in the speech signal. This type of speech recognition task is generally referred to as transcription. The set of statistical grammatical or syntactical rules was called a language model, of which the n-gram model, which defined the probability of occurrence of an ordered sequence of n words, was the most frequently used variant. Since their introduction in the 1980's, the use of n-gram language models, and its variants, has become indispensable in large vocabulary speech recognition systems.

At AT&T Bell Laboratories, the goal of the research program was to provide automated telecommunication services to the public, such as voice dialing, and command and control for routing of phone calls. These automated systems were expected to work well for a vast population of talkers without the need for individual speaker training. The focus at Bell Laboratories was in the design of a speaker-independent system that could deal with the acoustic variability intrinsic in the speech signals coming from many different talkers, often with notably different regional accents.

The IBM and AT&T Bell Laboratories approaches to speech recognition both had a profound influence in the evolution of human-machine speech communication technology of the last two decades. One common theme between these efforts, despite the differences, was that mathematical formalism and rigor started to emerge as distinct and important aspects of speech recognition research. While the difference in goals led to different realizations of the technology in various applications, the rapid development of statistical methods in the 1980's, most notably the hidden Markov

model (HMM) framework, caused a certain degree of convergence in the system design. Today, most practical speech recognition systems are based on the statistical framework and results developed in the 1980's, with significant additional improvements in the 1990's.

2.3 Research Work in the 1980's and 1990's

Speech recognition research in the 1980's was characterized by a shift in methodology from the more intuitive template-based approach towards a more rigorous statistical modeling framework. Although the basic idea of the hidden Markov model (HMM) was known and understood early on in a few laboratories, the methodology was not complete until the mid-1980's and it wasn't until after widespread publication of the theory that the hidden Markov model became the preferred method for speech recognition. The popularity and use of the HMM as the main foundation for automatic speech recognition and understanding systems has remained constant over the past two decades, especially because of the steady stream of improvements and refinements of the technology.

The idea of the hidden Markov model appears to have first come out in the late 1960's at the Institute for Defense Analyses (IDA) in Princeton, N.J. Len Baum referred to an HMM as a set of probabilistic functions of a Markov chain, which, by definition, involves two nested distributions, one pertaining to the Markov chain and the other to a set of the probability distributions, each associated with a state of the Markov chain, respectively.

Baum's doubly stochastic process started to find applications in the speech area, initially in speaker identification systems, in the late 1970's [25, 27]. As more people attempted to use the HMM technique, it became clear that the constraint on the form of the density functions imposed a limitation on the performance of the system, particularly for speaker independent tasks where the speech parameter distribution was not sufficiently well modeled by a simple log-concave or an elliptically symmetric density function. In the early 1980's at Bell Laboratories, the theory of HMM was extended to mixture densities which have since proven vitally important in ensuring satisfactory recognition accuracy, particularly for speaker independent, large vocabulary speech recognition tasks.

The merger of the hidden Markov model and the finite state network was an important, although not unexpected, technological development in the mid-1980. A tool, called the FSM (finite-state machine) library, which embodied the finite state network approach in a unified transducer framework (including weighted search) was developed in the mid-1990s and has been a major component of almost all modern speech recognition and understanding systems.

Another technology that was (re)introduced in the late 1980's was the idea of artificial neural networks (ANN). Neural networks were first introduced in the 1950's, but failed to produce notable results initially. The advent, in the 1980's, of a parallel distributed processing (PDP) model, which was a dense interconnection of simple computational elements, and a corresponding "training" method, called error back-propagation, revived interest around the old idea of mimicking the human neural processing mechanism.

In the 1990's, a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error.

The success of statistical methods revived the interest from DARPA at the juncture of the 1980's and the 1990's, leading to several new speech recognition systems including the Sphinx system from CMU, the BYBLOS system from BBN and the DECIPHER system from SRI. CMU's Sphinx system successfully integrated the statistical method of hidden Markov models with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context-dependent phone models in a sophisticated lexical decoding network achieving remarkable results for large-vocabulary continuous speech recognition.

In the 1990's great progress was made in the development of software tools that enabled many individual research programs all over the world. As systems became more sophisticated, a well-structured baseline software system was indispensable for further research and development to incorporate new concepts and algorithms. The system that was made available by the Cambridge University team (led by Steve Young), called the Hidden Markov Model Tool Kit (HTK) [16], was one of the most widely adopted software tools for automatic speech recognition research.

2.4 Research Work since 1990's

In this decade, some techniques were developed to make the speech recognition more robust. The major techniques include Maximum Likelihood Linear Regression (MLLR) technique, the Model decomposition technique, Parallel Model Composition (PMC) technique and the Structural Maximum a Posteriori (SMAP) method. These techniques were developed to overcome the problems coming due to background noises and other disturbances.

2.5 Research Work since 2000's

During the years of this decade, Defense Advanced Research Projects Agency (DARPA) conducted a program, the Effective Affordable Reusable Speech-to-Text (EARS). This program was conducted to develop the Speech-to-Text technology with the aim of achieving substantially richer and much more accurate output than before. The tasks included detection of sentence boundaries, fillers and disfluencies. The program was focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages. The goal of the program was to make it possible for machines to do a much better job of detecting, extracting, summarizing and translating important information, thus enabling humans to understand what was said by reading transcriptions instead of listening to audio signals.

It was noted that although the read speech and similar types of speech e.g. news broadcasts reading a text, could be recognized with accuracy higher than 95% using state-of-the-art speech recognition technology. But the recognition accuracy drastically decreased for spontaneous speech. In order to increase recognition performance for spontaneous speech, several projects were conducted. In Japan, a 5-year national project "Spontaneous Speech: Corpus and Processing Technology" was conducted. The world-largest spontaneous speech corpus, "Corpus of Spontaneous Japanese (CSJ)" consisting of approximately 7 millions of words, corresponding to 700 hours of speech, was built, and various new techniques were investigated. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic speech summarization.

- **Technology Developments on the Timeline**

1936 AT&T's Bell Labs produced the first electronic speech synthesizer called the Voder (Dudley, Riesz and Watkins). This machine was demonstrated in the 1939 World Fairs by experts that used a keyboard and foot pedals to play the machine and emit speech.

1969 John Pierce of Bell Labs said automatic speech recognition will not be a reality for several decades because it requires artificial intelligence.

Early 1970's The Hidden Markov Modeling (HMM) approach to speech recognition was invented by Lenny Baum of Princeton University and shared with several ARPA (Advanced Research Projects Agency) contractors including IBM. HMM is a complex mathematical pattern-matching strategy that eventually was adopted by all the leading speech recognition companies including Dragon Systems, IBM, Philips, AT&T and others.

1971 DARPA (Defense Advanced Research Projects Agency) established the Speech Understanding Research (SUR) program to develop a computer system that could understand continuous speech. Lawrence Roberts, who initiated the program, spent $3 million per year of government funds for 5 years. Major SUR project groups were established at CMU, SRI, MIT's Lincoln Laboratory, Systems Development Corporation (SDC), and Bolt, Beranek, and Newman (BBN). It was the largest speech recognition project ever.

1978 The popular toy "Speak and Spell" by Texas Instruments was introduced. Speak and Spell used a speech chip which led to huge strides in development of more human-like digital synthesis sound.

1982 Covox founded. Company brought digital sound (via The Voice Master, Sound Master and The Speech Thing) to the Commodore 64, Atari 400/800, and finally to the IBM PC in the mid ‘80s and Dragon Systems was founded in 1982 by speech industry pioneers Drs. Jim and Janet Baker. Dragon Systems is well known for its long history of speech and language technology innovations and its large patent portfolio.

1984 Speech Works, the leading provider of over-the-telephone automated speech recognition (ASR) solutions, was founded.

1993 Covox sells its products out to Creative Labs, Inc.

1995 Dragon released discrete word dictation-level speech recognition software. It was the first time dictation speech recognition technology was available to consumers. IBM and Kurzweil followed a few months later.

1996 Charles Schwab is the first company to devote resources towards developing up a speech recognition IVR system with Nuance. The program, Voice Broker, allows for up to 360 simultaneous customers to call in and get quotes on stock and options. It handles up to 50,000 requests each day. The system was found to be 95% accurate and set the stage for other companies such as Sears, Roebuck and Co., and United Parcel Service of America Inc., and E*Trade Securities to follow in their footsteps. BellSouth launches the world's first voice portal, called Val and later Info By Voice.

1997 Dragon introduced "Naturally Speaking", the first "continuous speech" dictation software available.

1998 Lernout & Hauspie bought Kurzweil. Microsoft invested $45 million in Lernout & Hauspie to form a partnership that will eventually allow Microsoft to use their speech recognition technology in their systems.

1999 Microsoft acquired Entropic, giving Microsoft access to what was known as the "most accurate speech recognition system" in the world.

2000 Lernout & Hauspie acquired Dragon Systems for approximately $460 million. TellMe introduces first world-wide voice portal and NetBytel launched the world's first voice enabler, which includes an on-line ordering application with real-time Internet integration for Office Depot.

2001 ScanSoft Closes Acquisition of Lernout & Hauspie Speech and Language Assets.

2002 Rich transcription of meetings, very large vocabulary, limited tasks and controlled environment

2003 ScanSoft Ships Dragon NaturallySpeaking 7 Medical, Lowers Healthcare Costs through Highly Accurate Speech Recognition. ScanSoft Closes Acquisition of Speech Works International, Inc and closes deal to

distribute and support IBM ViaVoice Desktop Products.

2004 Finnish online dictation and almost unlimited vocabulary based on monophones.

2006 Machine translation of broadcast speech.

2007 Difference in acoustic features between spontaneous and read speech using a large scale speech database have been analyzed.

2008 Exploring the application of Conditional Random Field (CRF) to combine local posterior estimates provided by the multilayer perceptions corresponding to the frame level prediction of phone and phonological attributed classes.

2009 Quick adaption of synthesized voice by speech recognition.

2011 Unlimited vocabulary, unlimited tasks, many languages, multilingual systems for multimodal speech enabled devices.

In this chapter, history of speech recognition system has been discussed since 1950's. This chapter, also, briefly reviews several technology developments. In the next chapter, the problem statement has been discussed with the objectives of this thesis work and methodologies to be followed to develop the speech recognition system.

Chapter 3

Problem Statement

Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands and control, data entry, and document preparation. They can also serve as the input to further linguistic processing. Nowadays, a lot of research work is going on for the development of speech recognition system with better accuracy. After years of research and development, the accuracy of automatic speech recognition (ASR) system remains one of the most important research challenges e.g. speaker variability, domain etc. The design of speech recognition system requires careful attention to the challenges or issues such as various types of speech classes, speech representation, feature extraction techniques and performance evaluation. Existing general speech recognition systems are not 100% accurate but the systems developed for particular domains have been very successful.

The core of all speech recognition systems consists of a set of statistical models representing the various sounds of the language to be recognized. Since speech has temporal structure and can be encoded as a sequence of spectral vectors spanning the audio frequency range, the Hidden Markov Model (HMM) provides a natural framework for constructing such models.

In HMM, speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. HMM can be trained automatically and HMM based models are simple and computationally feasible to use. Hence HMM has become the predominant technology in automatic speech recognition (ASR) and most of the state-of-the-art ASR systems are based on HMM. By increasing the number of hidden states present in HMM model or by using a factored hidden-variable representation, HMM's ability increases. A speech recognition system requires a microphone for the person to speak into, speech recognition software, a computer to take and interpret the speech; and a proper and good pronunciation.

3.1 Objectives

With above described context and in order to solve the problem statement, objectives have been framed for this thesis work. The objectives that have been framed for the development of the speech recognition system are as follows.

- To study the architecture of HMM-based speech recognition system
- To train the system with domain specific vocabulary and to recognize the numbers in the range 0-999
- To recognize both connected speech and isolated words
- To make the HMM-based speaker independent speech recognition system
- To test the effect of number of states present in the various HMM models
- To test the accuracy of the speech recognition system in open space and room environment
- To compare the accuracy of triphone and word model based speech recognition system

3.2 Methodology

To achieve all the objectives discussed in above section, the following three methodologies have been adopted.

- Development of HMM-based speech recognition system using word-based acoustic model
- Development of HMM-based speech recognition system using triphone-based acoustic model
- Comparison of triphone and word model based speech recognition system

3.2.1 Methodology for the Development of HMM-based Speech Recognition System Using Word-based Acoustic Model

To develop HMM-based speech recognition system using word-based acoustic model, following steps have been performed.

- ***HMM Tool Kit (HTK)*** version 3.4.1 has been installed on Linux (Ubuntu 11.10). This tool has been used to train the system.
- ***Audacity*** tool has been installed for recording the speech.

- ***Wavesurfer*** tool has been installed for labeling of speech files.
- Data preparation has been done for transforming it to usable form for training.
- Word model as acoustic model has been defined and different topology has been used for different word models.
- Initialization and training of various word models has been performed.
- The set of rules in the form of grammar of the system has been defined.
- Compilation of task grammar has been done so that ***HTK*** can understand it.
- Dictionary has been defined showing the correspondence between the name of the HMM and the name of task grammar variable.
- Speech files for testing have been recorded using Philips microphone.
- System has been tested for all possible inputs in room and open space environment.

3.2.2 Methodology for the Development of HMM-based Speech Recognition System Using Triphone-based Acoustic Model

To develop HMM-based speech recognition system using triphone-based acoustic model, following steps have been performed.

- ***HMM Tool Kit (HTK)*** version 3.4.1 has been installed on Linux (Ubuntu 11.10). This tool has been used to train the system.
- ***Audacity*** tool has been installed for recording the speech.
- Data preparation has been done for transforming it to usable form for training.
- Master Label File (MLF) has been created manually and each word has been replaced with its corresponding phonemes.
- Training with monophones has been performed.
- Training with triphones has been performed for better accuracy.
- The set of rules in the form of grammar of the system has been defined.
- Compilation of task grammar has been done so that ***HTK*** can understand it.
- Dictionary has been defined showing the correspondence between the name of the HMM and the name of task grammar variable.
- Speech files for testing have been recorded using Philips microphone.
- System has been tested for all possible inputs in room and open space environment.

3.2.3 Methodology for the Comparison of Triphone and Word Model Based Speech Recognition System

In order to compare the performance of triphone and word model based speech recognition system, following comparisons are made.

- Performance parameters of word model versus triphone model of isolated word speech recognition system in room environment
- Performance parameters of word model versus triphone model of isolated word speech recognition system in open space environment
- Performance parameters of word model versus triphone model of connected word speech recognition system in room environment
- Performance parameters of word model versus triphone model of connected word speech recognition system in open space environment

The above mentioned methodologies have been followed in the next three chapters to achieve the objectives defined in thesis. In the next chapter, implementation of HMM-based speech recognition system using HTK by varying the number of states in word-based acoustic model has been presented.

Chapter 4

Implementation of HMM-based Speech Recognition System Using HTK by Varying the Number of States in Word-based Acoustic Model

This chapter presents the implementation of HMM-based (Hidden Markov Model-based) speech recognition system using Hidden markov model Tool Kit (HTK).The acoustic model which is used for recognition is word model and the number of states to be used in various word models is variable in nature. For a particular word model, the number of states to be used is decided by the number of phonemes present in the corresponding word and the duration of that word. The system is speaker independent which recognizes the both connected speech and isolated words.

4.1 Introduction

Speech recognition is the process used to recognize speech uttered by a speaker. Speech is the most natural form of communication between the humans; it can be done without any tool or education. Speech recognition is the one of the most important areas of signal processing and has been part of research fields. It can be used in many applications like in security devices, computer, mobile phones, ATM machines, household appliances etc. HMM based statistical speech recognition systems have been popular in the past decade or so and have shown better results as compared to the other recognition techniques. Statistically based Automatic Speech Recognition (ASR) systems are based on the notion that an utterance is represented by some sequence of acoustic feature observations O that derive from the underlying sequence of words W, and the two can be probabilistically related. More specifically, the goal of a statistically based ASR system is to find

$$\arg\max_W P(W|O) \tag{4.1}$$

The equation 4.1 is the word having maximum probability when observation vector O is given. It can be written using Baye's rule as

$$\arg\max_W \frac{P(W)P(O|W)}{P(O)} = \arg\max_W P(W)P(O|W) \quad (4.2)$$

In equation 4.2 P(O) is ignored as it is constant with respect to W. This breaks the problem into two sub problems.

The first is the calculating the probability P(W) which can be calculated by constructing the language model. Also P(W) can be represented as

$$P(W)=P(w_1,w_2,w_3,\text{-----},w_{n-2},w_{n-1},w_n) \quad (4.3)$$

$$P(W)=P(w_1).P(w_2|w_1).P(w_3|w_1w_2).\text{-----}.P(w_n|w_1w_2\text{---}w_{n-1}) \quad (4.4)$$

$$P(W)=\Pi^n_{t=1} P(w_t|w_1w_2w_3w_4\text{---}w_{t-1}) \quad (4.5)$$

The second is calculating the probability P(O|W) which is probability of observed sequence of acoustic feature observation when the word sequence is given; this probability can be calculated by constructing the acoustic model.

This chapter aims to develop and implement an isolated and connected word speech recognition system using Hidden markov model Tool Kit (HTK). HTK is developed in 1989 by Steve Young at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED). This toolkit aims at building and manipulating Hidden Markov Models (HMMs). HTK consists of a set of library modules and tools available in C source form.

Apart from introduction in section 4.1, the chapter is organized as follows. Section 4.2 describes the architecture and functioning of developed speech recognition system. Section 4.3 deals with the implementation work. Section 4.4 shows the experimental results. Section 4.5 concludes the implemented system.

4.2 System Architecture

Figure 4.1 shows architecture of the developed speech recognition system. At the broad level, it includes training phase, testing phase and recognizing phase. First of all, speech files are recorded both for the training purpose and testing purpose. For that, all the vocabulary words are uttered a number of times by many speakers so that speech recognition system obtained is speaker independent. The speech files are represented in more compact and efficient form by extracting the features from them.

After extracting the features, acoustic models are generated from the trained speech files which are used for recognizing the unknown utterances (testing speech files) during recognition phase.

4.2.1 Training Phase

The speech files recorded with the help of microphone goes for acoustical analysis. Acoustical analysis (feature analysis/feature extraction) is the process of parameterization of the speech i.e. representation of speech utterances in terms of feature vectors, which can be used for making the acoustic models. Feature extraction is expected to discard the irrelevant information while keeping the useful one. Basically, the purpose of a feature extractor is to identify, within the data what information is needed to perform accurate classification. The speech signal contains the characteristics information of the speaker and environment in addition to signal message. A feature extractor for speech recognition needs to maximally discard the speaker and environment information and only allow the signal message information to pass. There are various methods for extracting the features which includes MFCC (Mel Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), LPCC (Linear Predictive Cepstral Coefficient), temporal patterns and many more. MFCC has generally obtained a better accuracy and a minor computational complexity with respect to alternative processing as compared to other feature extraction techniques. So, the developed system uses, Mel Frequency Cepstral Coefficients as feature extraction technique. Features are extracted, by using configuration file (.conf) and script file (.scp). Configuration file (.conf) contains the values of various parameters used for acoustic analysis. The script file (.scp) contains the location of the speech files (.wav) and also the location of the acoustic feature files (.mfcc) to be created.

To recognize an unknown utterance, it has to be compared with some reference models called acoustic models. Using these models, most probable sound is identified. There are two kinds of acoustic models-Word model and phoneme model. The developed system, in this chapter, uses word model as acoustic model. Acoustic models can be generated using various approaches such as Hidden Markov Model, Artificial Neural Network, Dynamic Bayesian Network, Support Vector Machine and hybrid approaches. Acoustic models are generated with the help of prototype models and using text file (.lab) obtained by manually labeling of input speech files.

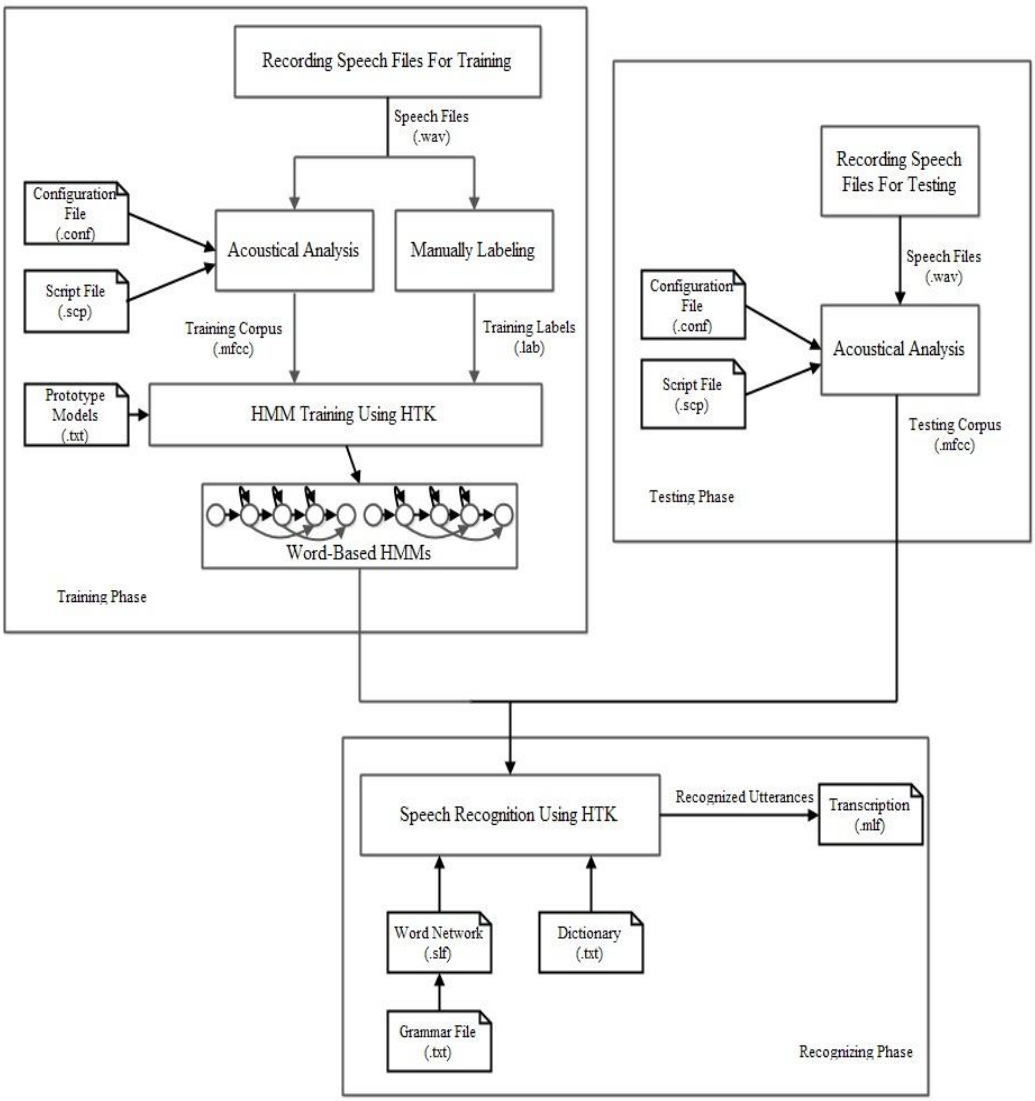


Figure 4.1: Developed Speech Recognition System Architecture

4.2.2 Testing Phase

For the testing purpose, speech files have to be recorded with the help of microphone. As in training phase, speech files (recorded for testing purpose) are analyzed. Features are extracted, by using configuration file (.conf) and script file (.scp), which will be used in recognizing phase.

4.2.3 Recognizing Phase

Recognizing phase recognizes the test samples using Hidden markov model Tool Kit (HTK) based on the acoustic properties which are calculated in the testing phase. Mainly, the job of recognizing phase is to combine information from the acoustic

model, language model (word grammar) that describes how words are concatenated to form valid sentences and the dictionary that describes the ways that each word can be pronounced.

4.3 Implementation

This section gives the implementation details of speech recognition system based upon the architecture described in the previous section.

4.3.1 System Description

The operating system which is used for making the speech recognition system is Linux (Ubuntu 11.10). The system is implemented using Hidden markov model Tool Kit (HTK) version 3.4.1. The system is trained with 29 words and word model is used as acoustic model. In addition to these, different word models are used by varying the number of states present in them.

4.3.2 Database Preparation

In this system, speech files are recorded with the help of Philips microphone. Distance between the mouth of the speaker and the microphone is approximately 5-10 cm. Recording is done at room environment. The speech files are in wave format (.wav). The (.wav) files recorded are saved as HTK transcriptions. The sampling frequency is 16 KHz, sample size is 16 bits and mono channels are used; and the tool used is *Audacity*. Figure 4.2 shows recorded sounds of the word "Zero" using tool *Audacity.*

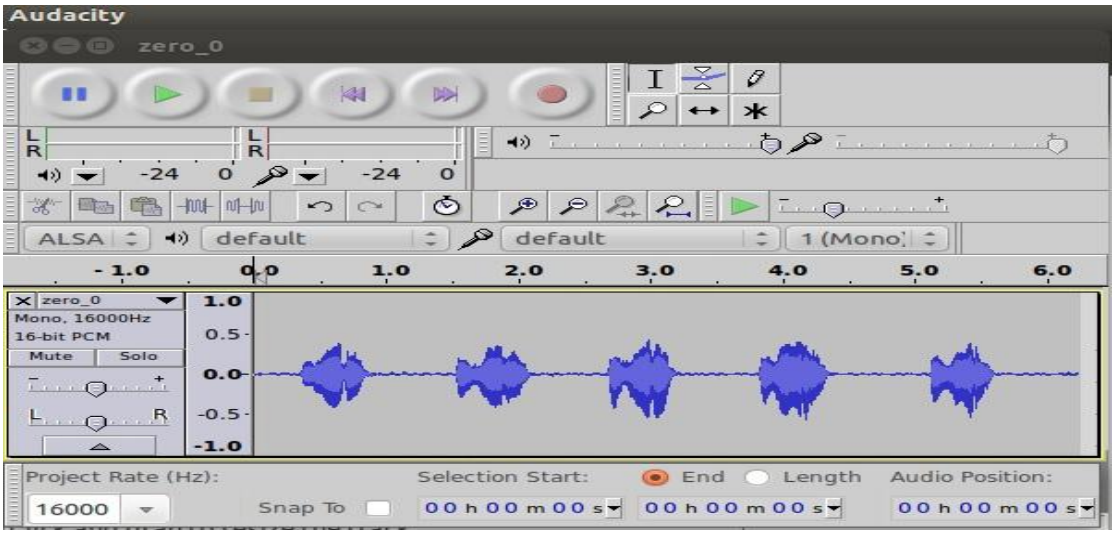


Figure 4.2: Recorded Sounds of the Vocabulary Word "Zero"

Labeling of speech files is done with the help of *Wavesurfer*. The system is trained for 29 words. For the training, speech samples are collected from 10 persons in which 5 men and 5 women are there in the age of 20-25. Each word is uttered 10 times by each speaker. So, the training database includes 2900 (29×10×10) speech files. Figure 4.3 shows HTK transcription of a speech waveform.

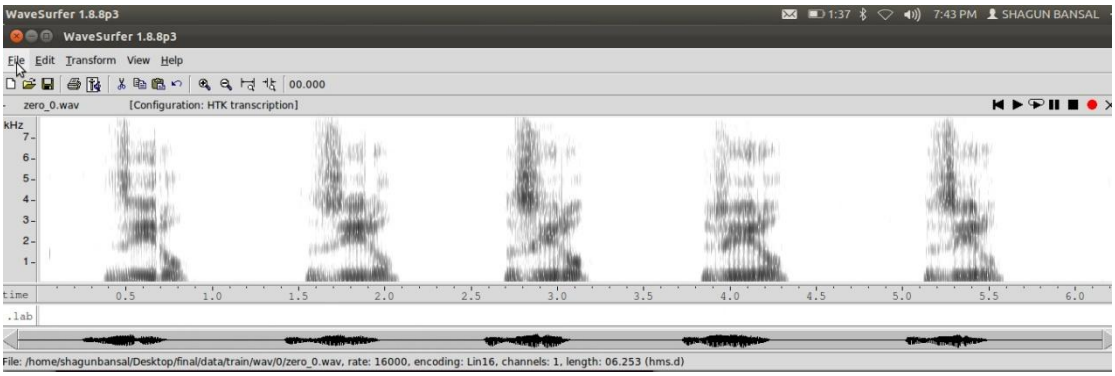


Figure 4.3: HTK Transcription of a Speech Waveform

4.3.3 Acoustic Analysis

In acoustic analysis, the speech files are represented in a more compact and efficient way by parameterized them into a sequence of feature vectors. For this purpose, HTK tool **HCopy** is used. The system uses the Mel Frequency Cepstral Coefficients (MFCCs) to extract feature vectors from the recorded speech files. The recorded speech files are processed at the frame rate of 10 ms with a Hamming window of 25 ms. Acoustic parameters used are 39 MFCCs having 12 MFCC coefficients with log energy and their delta and acceleration coefficients. Also, speech files are parameterized using a coefficient of 0.97 and number of filterbank channels used is 26. The values of various parameters used for acoustic analysis are shown in table 4.1.

Table 4.1: Values of Various Parameters Used for Acoustic Analysis

No.	Parameter	Value
1	SOURCEFORMAT	.wav
2	TARGETKIND	MFCC_0_D_A
3	WINDOWSIZE	25 msec.
4	TARGETRATE	10 msec.
5	NUMCEPS	12
6	USEHAMMING	True
7	PREEMCOEF	0.97
8	NUMCHANS	26
9	CEPLIFTER	22

4.3.4 Acoustic Model

Word model has been used as acoustic model. The used word models are basically defined in terms of statistical Hidden Markov Model (HMM). Different topology has been used for different word models. Word models differ by the number of states present in them. In the system, number of states for a particular word model is decided by the number of phonemes present in the corresponding word and the duration of that word. This system uses 6-18 states HMMs in which first and last are non-emitting states. Initialization of various word models are done using HTK tool **HInit**. Then HMM parameters are re-estimated using HTK tool **HRest**. Re-estimation is done repeatedly until absolute value of convergence factor does not decrease from one **HRest** iteration to another. In this system, re-estimation is done four times. Table 4.2 shows the number of states presents in various word models.

Table 4.2: Number of States Presents in Various Word Models

Word Model	**No. Of States**	**Word Model**	**No. Of States**	**Word Model**	**No. Of States**
ONE	8	ELEVEN	14	THIRTY	10
TWO	6	TWELVE	12	FORTY	12
THREE	8	THIRTEEN	12	FIFTY	12
FOUR	8	FOURTEEN	14	SIXTY	14
FIVE	8	FIFTEEN	14	SEVENTY	16
SIX	10	SIXTEEN	16	EIGHTY	10
SEVEN	12	SEVENTEEN	18	NINETY	12
EIGHT	8	EIGHTEEN	12	HUNDRED	16
NINE	8	NINETEEN	14	ZERO	10
TEN	8	TWENTY	14	SIL	12

4.3.5 Language Model

The developed system uses grammar based approach of language modeling. The grammar is specified in extended Backus-Naur form (EBNF). This form of grammar is not understandable by the HTK. The task grammar has to be compiled so that HTK can understand it. HTK tool **HParse** is used to compile the task grammar. A dictionary is also made which describes the correspondence between the name of the HMM and the name of task grammar variable. Figure 4.4 shows task grammar for the developed speech recognition system.

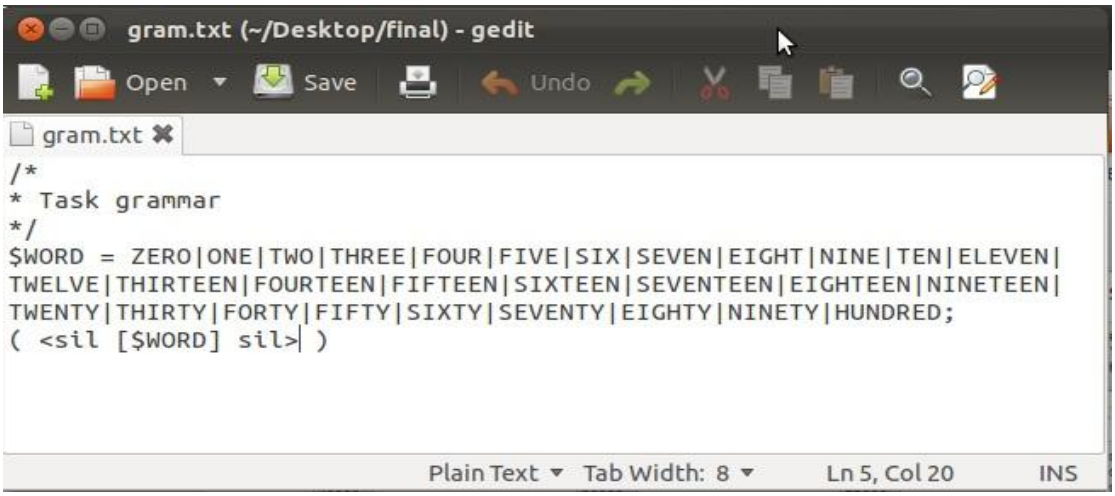


Figure 4.4: Task Grammar for the Developed System

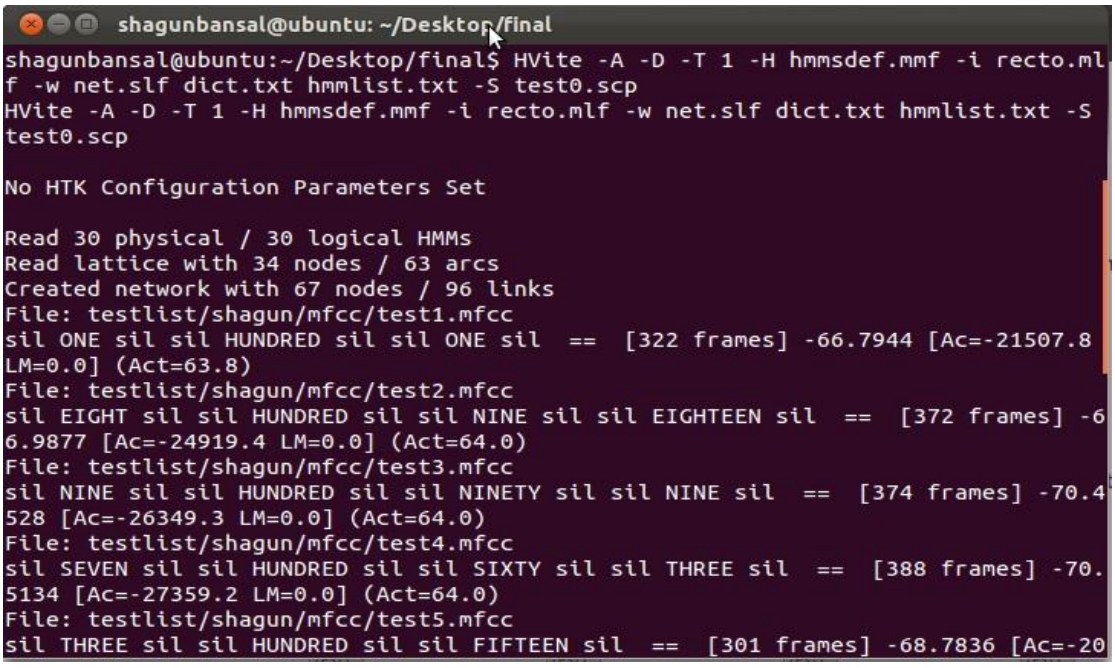


Figure 4.5: Connected Word Recognition

4.3.6 System Testing

During testing, firstly, unknown utterance is converted into the series of acoustic vectors using HTK tool **HCopy**. These acoustic vectors are then processed by Viterbi algorithm. HTK tool **HVite** is used to perform Viterbi based speech recognition. **HVite** takes as input a network describing the allowable word sequences, a dictionary and acoustic model generated during training. Connected word recognition is shown in figure 4.5. Figure 4.6 shows HTK results analysis of recognition of 15 sentences.

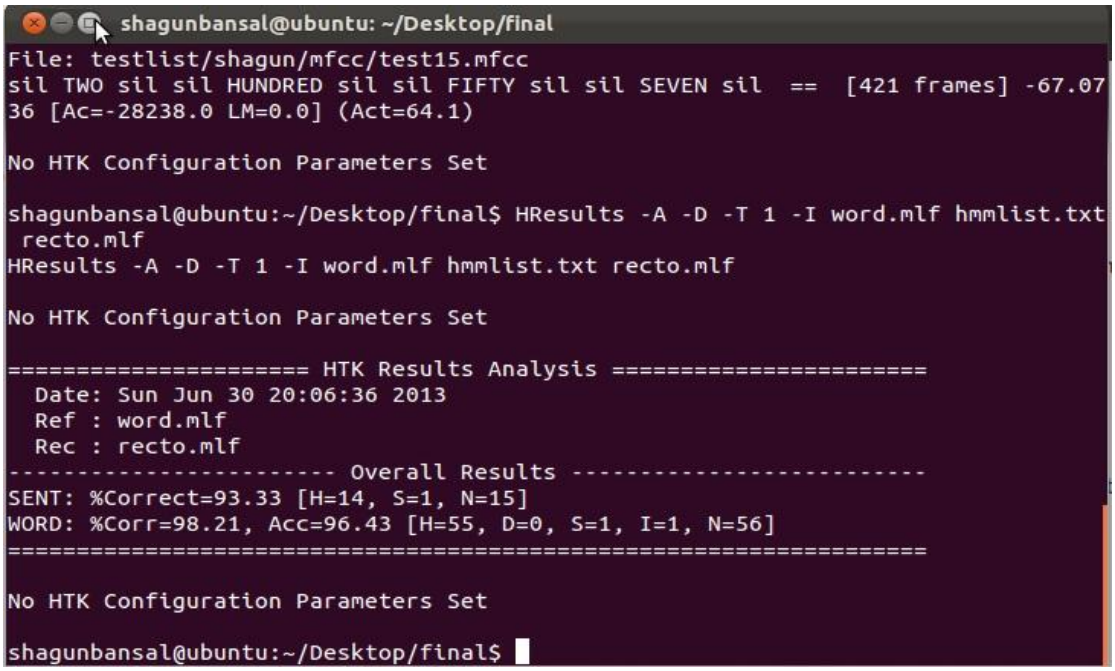


Figure 4.6: HTK Results Analysis

4.4 Experimental Results

First of all, results have been obtained by varying the number of states present in the HMM models. For these results, testing data is used same as that of training data as to find optimal system having maximum recognition accuracy. The results shown in Table 4.3 reveal that optimal system is obtained by using different number of states for different word models. In the system proposed by us, when numbers of states used for each model are 5, 6, 7 and 8 then respective recognition accuracy for connected word recognition is 81.28%, 81.62%, 92.55% and 92.76%. The recognition accuracy is 98.72% when numbers of states are different for different word models in the proposed system. Also, optimal system is used to recognize the isolated and connected word utterances in open space and room environment. To get these results,

two types of sounds are used-sounds spoken by the seen speakers i.e. sounds of the speakers whose other sound files are used for training and sounds spoken by the unseen speakers i.e. sounds of the speakers that does not participate in the training. Table 4.4 to 4.7 shows recognition results. The respective recognition accuracy of optimal system for isolated and connected word recognition is 95% and 94.04% in room environment, and 93.67% and 92.26% in open space environment.

Table 4.3: Experimental Results Obtained by Varying the Number of States in Word-based Acoustic Model

Number Of States in Each Word Model	**Types Of Sound**	**No. Of Spoken Sentences**	**Sentence Correction Rate**	**Word Correction Rate**
5	Seen	580	54.14	81.28
6		580	58.28	81.62
7		580	75.34	92.55
8		580	78.62	92.76
Different No. Of States In Different Word model		580	95.34	98.72

Table 4.4: Isolated Word Recognition in Room Environment

Speaker	**Environment**	**No. Of Spoken Words**	**No. Of Recognized Words**	**No. Of Deletion**	**No. Of Substitution**	**No. Of Insertion**	**Word Correction Rate**	**Word Accuracy Rate**	**Word Error Rate**
Speaker 1	Room	50	50	0	0	0	100	100	0
Speaker 2		50	49	0	1	0	98	98	2
Speaker 3		50	44	0	6	0	88	88	12
Seen Speaker		150	143	0	7	0	95.33%	95.33%	4.67%
Speaker 4	Room	50	45	0	5	1	90	88	12
Speaker 5		50	47	0	3	0	94	94	6
Speaker 6		50	50	0	0	0	100	100	0
Unseen Speaker		150	142	0	8	1	94.67%	94%	6%
Average System Performance:							95%	94.66%	5.34%

Table 4.5: Isolated Word Recognition in Open Space Environment

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Open Space	50	46	0	4	3	92	86	14
Speaker 2		50	47	0	3	0	94	94	6
Speaker 3		50	49	0	1	0	98	98	2
Seen Speaker		150	142	0	8	3	94.67%	92.67%	7.33%
Speaker 4	Open Space	50	46	0	4	4	92	84	16
Speaker 5		50	48	0	2	1	96	94	6
Speaker 6		50	45	0	5	2	90	86	14
Unseen Speaker		150	139	0	11	7	92.67%	88%	12%
Average System Performance:							93.67%	90.33%	9.67%

Table 4.6: Connected Word Recognition in Room Environment

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Room	56	56	0	0	0	100	100	0
Speaker 2		56	55	0	1	0	98.21	98.21	1.79
Speaker 3		56	48	0	8	1	85.71	83.93	16.07
Seen Speaker		168	159	0	9	1	94.64%	94.04%	5.96%
Speaker 4	Room	56	51	0	5	4	91.07	83.93	16.07
Speaker 5		56	55	0	1	0	98.21	98.21	1.79
Speaker 6		56	51	0	5	2	91.07	87.50	12.5
Unseen Speaker		168	157	0	11	6	93.45%	89.88%	10.12%
Average System Performance:							94.04%	91.96%	8.04%

Table 4.7: Connected Word Recognition in Open Space Environment

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Open Space	56	55	0	1	1	98.21	96.43	3.57
Speaker 2		56	55	0	1	0	98.21	98.21	1.79
Speaker 3		56	48	0	8	2	85.71	82.14	17.86
Seen Speaker		168	158	0	10	3	94.04%	92.26%	7.74%
Speaker 4	Open Space	56	49	0	7	5	87.50	78.57	21.43
Speaker 5		56	53	0	3	0	94.64	94.64	5.36
Speaker 6		56	50	0	6	4	89.29	82.14	17.86
Unseen Speaker		168	152	0	16	9	90.48%	85.12%	14.88%
Average System Performance:							92.26%	88.69%	11.31%

4.5 Conclusion

The speech recognition using word model gives optimal results when the topology used in word models differ by the number of states used in them. The number of states present in a particular word model is decided based upon the number of phonemes present in the corresponding word and the duration of that word. Also, the results conclude that accuracy of the system is sensitive to the changing environment. In the next chapter, speech recognition system is implemented using triphone based (context-dependent phoneme based) acoustic model.

Chapter 5

Implementation of HMM-based Speech Recognition System Using Triphone-based Acoustic Model

Previous chapter presents the implementation of speech recognition system using word based acoustic model. This chapter presents the HMM-based (Hidden Markov Model-based) speech recognition system using triphone based (context-dependent phoneme based) acoustic model. Again, the system can recognize both connected speech and isolated words; also the system is speaker independent. The system is trained to recognize any sequence of words selected from the vocabulary of 29 distinct English words. Basically, these 29 words are chosen so uniquely that the system can recognize the numbers in the range 0-999 like the system can recognize the number 857 if uttering "Eight Hundred Fifty Seven" in the microphone.

5.1 Introduction

Speech recognition is the process which transforms the spoken utterances into its equivalent text form. The research work on speech recognition has been done since last many years and has been done all over the world. The improvement in accuracy of speech recognition system is increasing rapidly day by day. There are many applications where speech recognition can be used like in security devices, mobile phones, computer, ATM machines, household appliances etc. Mainly, the researchers are showing great interests towards speech recognition in hand-held devices like in mobile phones, iPods and iPhones, and lot of work has been done to improve the accuracy of recognizing the speech in these devices. There are many issues behind the speech recognition like in which environment recognition to be done, recognizing the speech is speaker independent or speaker dependent, size of vocabulary to be used by the speech recognition system, ability to recognize isolated words and/or continuous words and many more. In the present era, speech recognition is mainly done by the Hidden Markov Model-based (HMM-based) statistical approach. HMM is a doubly embedded stochastic process with an underlying stochastic process that is not directly observable but can be observed only through another set of stochastic processes that

produce the sequence of observations. When using HMM as speech recognition technique, there are many issues regarding the basic model (acoustic model) to be use. The main issue is what is the kind of speech units to model? Speech units to be model can be a whole word or sub word.

The system, in this chapter, uses context-dependent phonemes particularly triphones as sub word of speech units for modeling purpose. Context-dependent phoneme model assumes that while modeling phone, their neighbors are also considered. Context-dependent phoneme model is the one which is dependent on the left and/or right neighboring phone. The model which considers either the left (preceding) or right (succeeding) phone is biphone model and the model which considers both neighboring phones is triphone model.

Apart from introduction in section 5.1, the chapter is organized as follows. Section 5.2 describes the architecture and functioning of HMM-based speech recognition system. Section 5.3 deals with the implementation work. Section 5.4 shows the experimental results. Section 5.5 concludes the implemented system.

5.2 HMM-based Speech Recognition System

Figure 5.1 shows Hidden Markov Model-based (HMM-based) speech recognition system. At the broad level, it includes training phase, testing phase and recognizing phase.

5.2.1 Training Phase

The speech files recorded with the help of microphone goes for acoustical analysis. Acoustical analysis (feature analysis/feature extraction) is the process which transforms the input data into a set of acoustic features. As input speech signal is complex in nature, acoustic analysis helps in removing irrelevant information consisting of background noise, recording device characteristics so that compact form of acoustic information is obtained which can be used with various acoustic models. System in this chapter uses Mel Frequency Cepstral Coefficients for acoustic analysis. After the acoustical analysis of input speech files, acoustic models are generated with the help of prototype model and phone level transcriptions. Configuration file (.conf) and script file (.scp) are used during acoustical analysis. Configuration file (.conf)

contains the values of various parameters used for acoustic analysis. The script file (.scp) contains the location of the speech files (.wav) and also the location of the acoustic feature files (.mfcc) to be created.

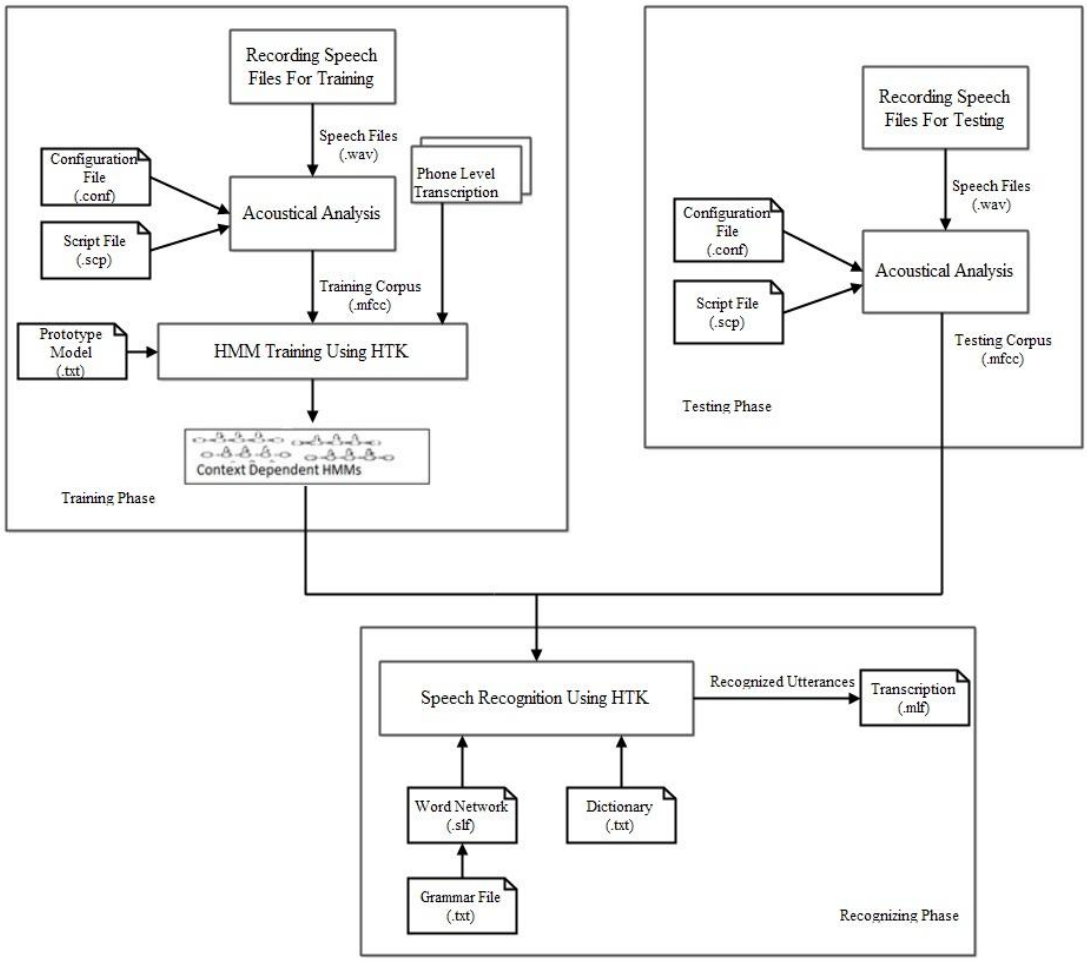


Figure 5.1: HMM-based Speech Recognition System

5.2.2 Testing Phase

For the testing purpose, speech files have to be recorded with the help of microphone. As in training phase, speech files (recorded for testing purpose) are analyzed. Features are extracted by using configuration file (.conf) and script file (.scp).

5.2.3 Recognizing Phase

Recognizing phase recognizes the test samples using Hidden markov model Tool Kit (HTK) based on the acoustic properties which are calculated in the testing phase. The job of recognizing phase is to combine information from the acoustic model, language

model (word grammar) that describes how words are concatenated to form valid sentences and the dictionary that describes the ways that each word can be pronounced.

5.3 Implementation

This section gives the implementation details of speech recognition system based upon the architecture described in the previous section.

5.3.1 System Description

The isolated and connected word speech recognition system based on MFCC is developed using Hidden markov model Tool Kit (HTK) version 3.4.1. The system is developed in the Linux environment (Ubuntu 11.10). The system is designed to recognize both connected speech and isolated words from the set of 29 distinct English words. The vocabulary of 29 words is chosen so uniquely that the system can recognize the numbers in the range 0-999.

5.3.2 Training Data Preparation

This phase consists of recording and labeling of speech signals. In this system, recording is done with the help of Philips microphone. Distance between the mouth of the speaker and the microphone is approximately 5-10 cm. The speech files recorded in wave format (.wav) are saved as HTK transcriptions. Recording is done at room environment using tool *Audacity* at sampling rate of 16000 Hz on the mono channel and 16 bits are used as sample size. Voices of 10 people are used to train the system in which 5 men and 5 women are there in the age of 20-25. Each one is asked to utter each word 10 times making a total of 2900 (29×10×10) speech files as system is trained with 29 distinct words. After recording of speech files, a Master Label File (MLF) is manually created that contains a label entry for each speech file and then HTK tool **HLEd** is used to expand word level transcriptions to phoneme level transcriptions i.e. replacing each word with its phonemes.

5.3.3 Acoustical Analysis

In acoustical analysis, recorded speech files are parameterized into a set of acoustic features. For this purpose, HTK tool **HCopy** is used. The technique used for parameterization of the speech is Mel Frequency Cepstral Coefficients (MFCCs). The

recorded speech files are processed at the frame rate of 10 ms with a Hamming window of 25 ms. Table 5.1 shows the properties of the speech files and the parameters used for acoustic analysis.

Table 5.1: Values of Various Parameters Used for Acoustic Analysis

No.	Parameter	Value
1	SOURCEFORMAT (Input File Format)	.wav
2	TARGETKIND (Coefficients To Use)	MFCC_0_D_A(MFCC with energy, delta and acceleration coefficients)
3	WINDOWSIZE (Length Of Time Frame)	25 msec.
4	TARGETRATE (Frame Periodicity)	10 msec.
5	NUMCEPS (Number Of MFCC Coefficients)	12
6	USEHAMMING (Hamming Function For Window Frame)	True
7	PREEMCOEF (Pre-emphasis Coefficient)	0.97
8	NUMCHANS (Number Of Filterbank Channels)	26
9	CEPLIFTER (Length Of Cepstral Liftering)	22

5.3.4 Acoustic Modeling

The acoustic model used is triphone model. The used acoustic models are basically defined in terms of statistical Hidden Markov Model (HMM). For generating triphone model (context-dependent phoneme model), first of all, monophone models (context-independent phoneme models) are generated. For generating monophone models, a prototype model is defined. Then global mean and variance are computed using HTK tool **HCompV**. After that, a Master Macro File (MMF) called hmmdefs containing a

copy for each of the required monophone HMMs is constructed by manually copying the prototype and relabeling it for each required monophone including 'sil'. Then flat start monophones are re-estimated using the embedded re-estimation HTK tool **HERest**. Now triphone-based HMMs are generated using HTK tool **HLEd** which will convert the monophone transcriptions (context-independent phoneme transcriptions) to an equivalent set of triphone transcriptions (context-dependent phoneme transcriptions). Then cloning of models is done using HTK tool **HHEd** and re-estimation of triphone set is done using HTK tool **HERest**. After re-estimation, context-dependent HMMs are generated.

Table 5.2: Phoneme Representation of Vocabulary Words

Vocabulary Word	**Phoneme Representation**	**Vocabulary Word**	**Phoneme Representation**
ONE	w ah n	SIXTEEN	s ih k s t iy n
TWO	t uw	SEVENTEEN	s eh v ih n t iy n
THREE	th r iy	EIGHTEEN	ey t iy n
FOUR	f ow r	NINETEEN	n ay n t iy n
FIVE	f ay v	TWENTY	t w eh n t iy
SIX	s ih k s	THIRTY	th er t iy
SEVEN	s eh v ih n	FORTY	f ow r t iy
EIGHT	ey t	FIFTY	f ih f t iy
NINE	n ay n	SIXTY	s ih k s t iy
TEN	t eh n	SEVENTY	s eh v ih n t iy
ELEVEN	ih l eh v ih n	EIGHTY	ey t iy
TWELVE	t w eh l v	NINETY	n ay n t iy
THIRTEEN	th er t iy n	HUNDRED	hh ah n d r ih d
FOURTEEN	f ow r t iy n	ZERO	z iy r ow
FIFTEEN	f ih f t iy n	SENT-START SENT-END	sil

5.3.5 Language Modeling - Task Definition

The developed system uses grammar based approach of language modeling. The grammar is specified using extended Backus-Naur form (EBNF). The task grammar is compiled using HTK tool **HParse** which generates a task network describing the sequence of vocabulary words that can be recognized by the system. A dictionary is also made describing the correspondence between the name of the HMM and the name of task grammar variable. As triphone model is used as acoustic model, the name of the HMM of a vocabulary word is the sequence of phonemes of that word as HMM of a certain word is made up of the concatenation of the HMMs of the phonemes that constitute the word. Table 5.2 shows phoneme representation of all the vocabulary words.

5.3.6 System Testing

For testing, firstly, recording of unknown utterances (speech signals) is done. Unknown utterances are converted into the series of acoustic features using HTK tool **HCopy**, same as done of training speech signals during acoustical analysis phase. These acoustic features along with dictionary and task network (generated during language modeling) and HMMs definition (generated during acoustic modeling) are taken as input by HTK tool **HVite**. It performs Viterbi based speech recognition which generates as output the recognized utterances in the form of transcription file (.mlf).

5.4 Experimental Results

The performance of the system is tested by recognizing the connected speech and isolated words in open space and room environment. To get these results, two types of sounds are used-sounds spoken by the seen speakers i.e. sounds of the speakers whose other sound files are used for training and sounds spoken by the unseen speakers i.e. sounds of the speakers that does not participate in the training. The respective recognition accuracy of developed system for isolated and connected word recognition is 96% and 94.94% in room environment, and 94% and 93.75% in open space environment. Table 5.3 to 5.6 shows recognition results of the developed system.

Table 5.3: Isolated Word Recognition in Room Environment Using Triphone-Based HMMs

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Room	50	50	0	0	0	100	100	0
Speaker 2		50	49	0	1	0	98	98	2
Speaker 3		50	48	0	2	0	96	96	4
Seen Speaker		150	147	0	3	0	98%	98%	2%
Speaker 4	Room	50	43	0	7	1	86	84	16
Speaker 5		50	49	0	1	0	98	98	2
Speaker 6		50	49	0	1	0	98	98	2
Unseen Speaker		150	141	0	9	1	94%	93.33%	6.67%
Average System Performance:							96%	95.66%	4.34%

Table 5.4: Isolated Word Recognition in Open Environment Using Triphone-based HMMs

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Open Space	50	50	0	0	0	100	100	0
Speaker 2		50	46	0	4	0	92	92	8
Speaker 3		50	48	0	2	1	96	94	6
Seen Speaker		150	144	0	6	1	96%	95.33%	4.67%
Speaker 4	Open Space	50	47	0	3	5	94	84	16
Speaker 5		50	46	0	4	0	92	92	8
Speaker 6		50	45	0	5	1	90	88	12
Unseen Speaker		150	138	0	12	6	92%	88%	12%
Average System Performance:							94%	91.66%	8.34%

Table 5.5: Connected Word Recognition in Room Environment Using Triphone-based HMMs

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Room	56	56	0	0	0	100	100	0
Speaker 2		56	54	0	2	0	96.43	96.43	3.57
Speaker 3		56	52	0	4	0	92.86	92.86	7.14
Seen Speaker		168	162	0	6	0	96.43%	96.43%	3.57%
Speaker 4	Room	56	49	0	7	1	87.5	85.71	14.29
Speaker 5		56	55	0	1	0	98.21	98.21	1.79
Speaker 6		56	53	0	3	0	94.64	94.64	5.36
Unseen Speaker		168	157	0	11	1	93.45%	92.86%	7.14%
Average System Performance:							94.94%	94.64%	5.36%

Table 5.6: Connected Word Recognition in Open Environment Using Triphone-based HMMs

Speaker	Environment	No. Of Spoken Words	No. Of Recognized Words	No. Of Deletion	No. Of Substitution	No. Of Insertion	Word Correction Rate	Word Accuracy Rate	Word Error Rate
Speaker 1	Open Space	56	55	1	0	0	98.21	98.21	1.79
Speaker 2		56	53	0	3	0	94.64	94.64	5.36
Speaker 3		56	53	0	3	0	94.64	94.64	5.36
Seen Speaker		168	161	1	6	0	95.83%	95.83%	4.17%
Speaker 4	Open Space	56	52	0	4	0	92.86	92.86	7.14
Speaker 5		56	55	0	1	0	98.21	98.21	1.79
Speaker 6		56	47	0	9	0	83.93	83.93	16.07
Unseen Speaker		168	154	0	14	0	91.67%	91.67%	8.33%
Average System Performance:							93.75%	93.75%	6.25%

5.5 Conclusion

This chapter presents a speech recognition system using HMM-based approach. The presented system can recognize both connected speech and isolated words; also the system is speaker independent. The system uses triphone-based acoustic model and trained to recognize any sequence of words selected from the vocabulary of 29 distinct English words. To evaluate the system performance, system is tested in the room environment and open space environment and experimentally it has been observed that the respective recognition accuracy of presented system for isolated and connected word recognition is 96% and 94.94% in room environment, and 94% and 93.75% in open space environment. In the next chapter, experimental results of 4^{th} chapter and 5^{th} chapter are compared i.e. the speech recognition system based on word model presented in chapter 4 and speech recognition system based on triphone model presented in chapter 5 are compared.

Chapter 6

Comparing Triphone and Word Model Based Speech Recognition for Small Vocabulary System

Previous two chapters present the implementation of two speech recognition systems, using word-based and triphone-based acoustic models. This chapter compares both of the speech recognition system. Both of the systems are trained to recognize any sequence of words selected from the vocabulary of 29 distinct English words.

6.1 Introduction

To recognize an unknown word, features extracted from raw speech signal have to be compared with some reference models to identify the sound that was produced as the word was spoken. This reference models are called as acoustic models. Basically, there are two kinds of acoustic model-Word model and phoneme model.

6.1.1 Word Model

System using word model as acoustic model defines one HMM for each word. In this model, words are modeled as a whole. Word model for English word “six” is shown in figure 6.1. The advantage of modeling speech at the word level directly is that such models allow the system designer to be oblivious to the internal structure of the words in the vocabulary, since the word models must capture co-articulation effects, etc, themselves. However, adding a new word in the vocabulary requires the system to train for the new word.

6.1.2 Phoneme Model

Every word is made up of a number of phonemes and the system using phoneme model as acoustic model defines one HMM for each phoneme of the word. HMM of a certain word is made up of the concatenation of the HMMs of the phonemes that constitute the word. The advantage of modeling speech at the phoneme level is that such models allow the representation of speech signal to be less redundant than the original acoustic representation, and consequently storage requirements would be

reduced. Also, adding a new word in the vocabulary is easy as the sounds of phones corresponding to the newly added word may be already known. In practice, the realization of one and the same phone differs a lot depending on its neighboring phones (the phone 'context'). Based on the context dependency, phoneme model can be further categorized into the context-independent and context-dependent phoneme model.

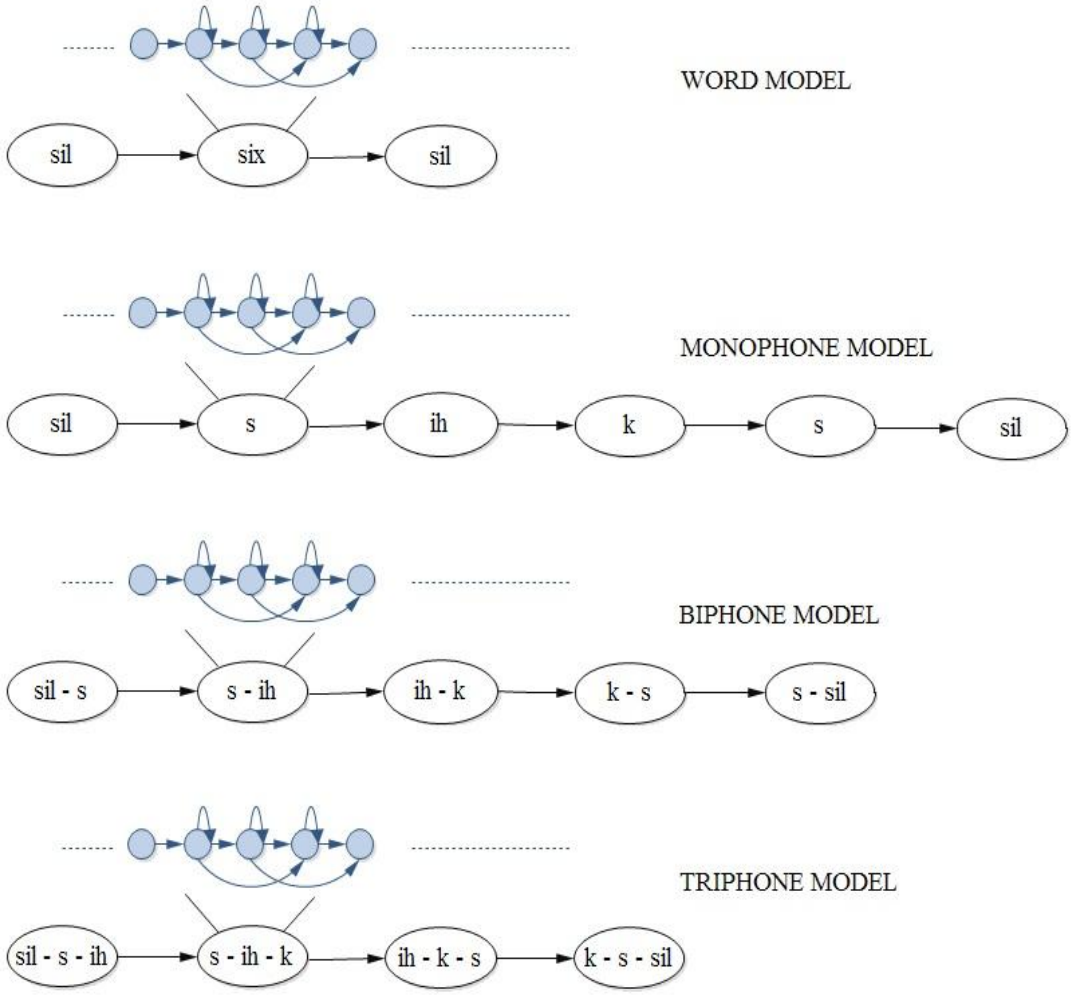


Figure 6.1: Word, monophone, biphone and triphone HMMs for the English word "six" [s ih k s]. 'sil' stands for silence at the beginning and end of the utterance, which is modeled as a 'phone', too.

- **Context-Independent Phoneme Model (Monophone Model)**

Context-independent phoneme model assumes that speech is produced as a sequence of concatenated phones which are unaffected by the context. So, in this model individual phones are modeled. The good thing about this model is that the number of

phones that has to be modeled is small, but generally, accuracy of system obtained by monophone model is less as compared to context-dependent phoneme model. Monophone model for English word “six” is shown in figure 6.1.

- **Context-Dependent Phoneme Model**

Context-dependent phoneme model assumes that while modeling phone, their neighbors are also considered. This model is the one which is dependent on the left and/or right neighboring phone. The model which considers either the left (preceding) or right (succeeding) phone is biphone model and the model which considers both neighboring phones is triphone model. In triphone model, for each phone different models are used for a different context. High phone recognition accuracies can be obtained using context-dependent phoneme models. Biphone model and triphone model for English word “six” is shown in figure 6.1.

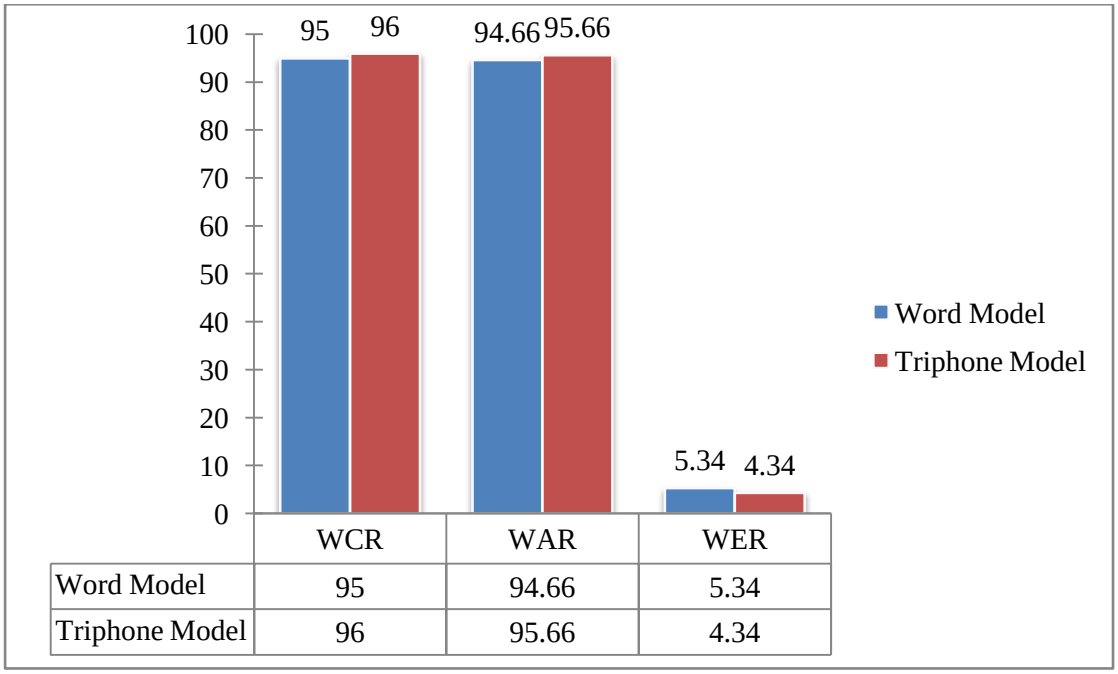
In this chapter, two speech recognition systems are compared. One is using word model as acoustic model and another one is using triphone model (context-dependent phoneme model) as acoustic model for recognition purpose.

Apart from introduction in section 6.1, the chapter is organized as follows. Section 6.2 presents the comparative analysis. Section 6.3 concludes the chapter.

6.2 Comparative Analysis

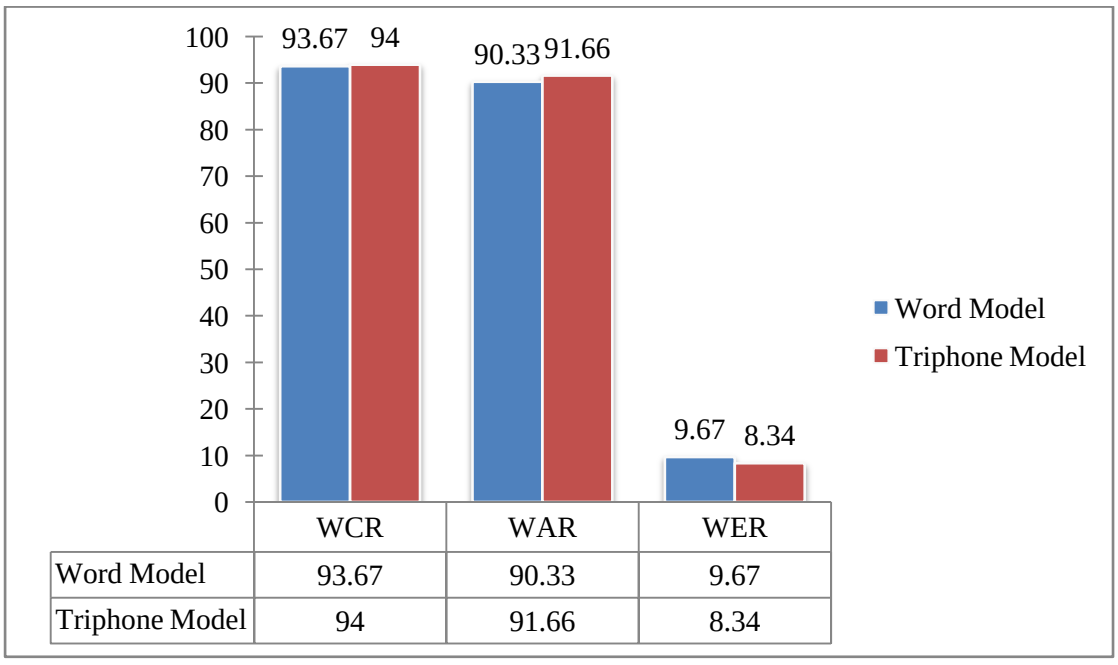
The system presented in 4th chapter uses word model as acoustic model and the number of states to be used in various word models is variable in nature. In the system, number of states for a particular word model is decided by the number of phonemes present in the corresponding word and the duration of that word. This system uses 6-18 states HMMs in which first and last are non-emitting states. Initialization of various word models are done using HTK tool **HInit** and HMM parameters are re-estimated using HTK tool **HRest**. The system presented in 5th chapter uses triphone model as acoustic model. Triphone-based HMMs are generated using HTK tool **HLEd** which will convert the monophone transcriptions (generated using HTK tools **HCompV** and **HERest**) to an equivalent set of triphone transcriptions. Re-estimation of triphone set is done using HTK tool **HERest**. After re-estimation, context-dependent HMMs are generated. Figure 6.2 to 6.5 shows

performance parameters of word model versus triphone model of isolated and connected word speech recognition system in room and open space environment.



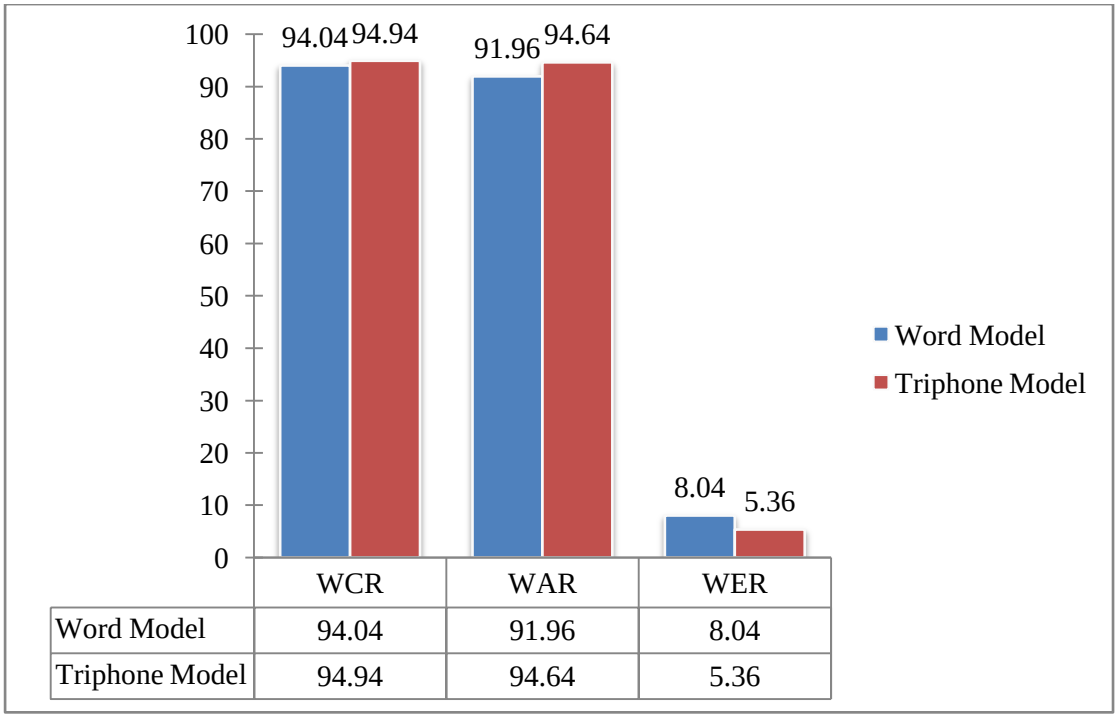
	WCR	WAR	WER
Word Model	95	94.66	5.34
Triphone Model	96	95.66	4.34

Figure 6.2: Performance Parameters of Word Model versus Triphone Model of Isolated Word Speech Recognition System in Room Environment



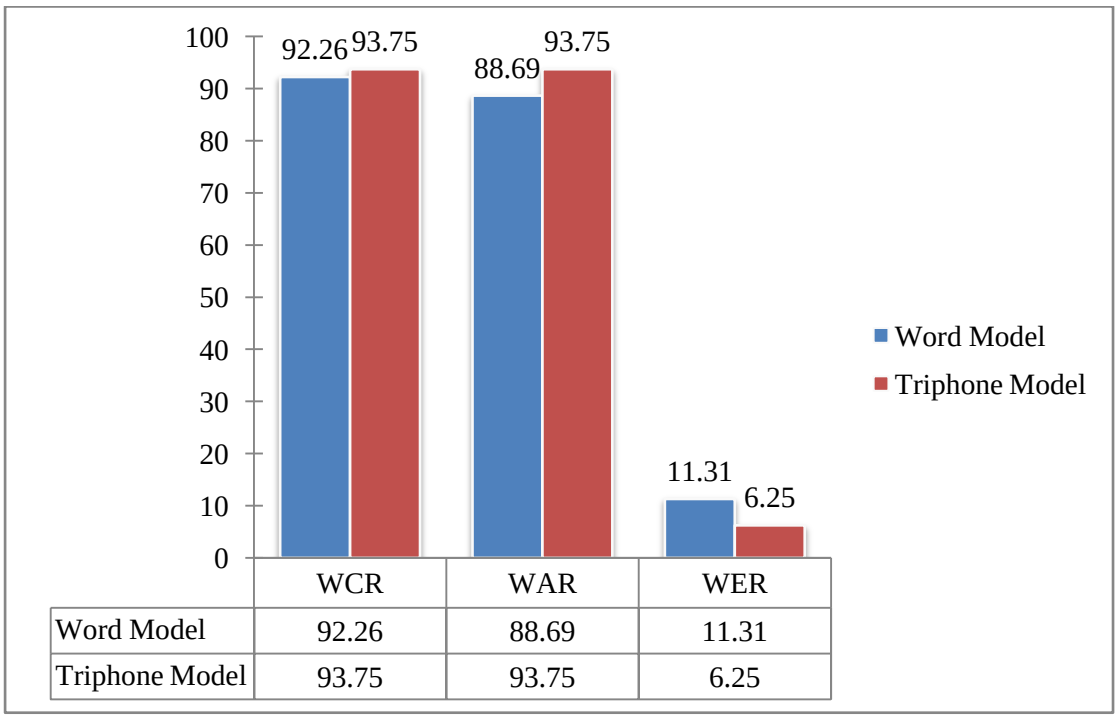
	WCR	WAR	WER
Word Model	93.67	90.33	9.67
Triphone Model	94	91.66	8.34

Figure 6.3: Performance Parameters of Word Model versus Triphone Model of Isolated Word Speech Recognition System in Open Space Environment



	WCR	WAR	WER
Word Model	94.04	91.96	8.04
Triphone Model	94.94	94.64	5.36

Figure 6.4: Performance Parameters of Word Model versus Triphone Model of Connected Word Speech Recognition System in Room Environment



	WCR	WAR	WER
Word Model	92.26	88.69	11.31
Triphone Model	93.75	93.75	6.25

Figure 6.5: Performance Parameters of Word Model versus Triphone Model of Connected Word Speech Recognition System in Open Space Environment

6.3 Conclusion

This chapter presents the comparison of triphone and word model based speech recognition systems. The respective recognition accuracy of triphone-based and word-based system for isolated word recognition is 96% and 95% in room environment, and 94% and 93.67% in open space environment; and respective recognition accuracy of triphone-based and word-based system for connected word recognition is 94.94% and 94.04% in room environment, and 93.75% and 92.26% in open space environment. So, it can be concluded here that speech recognition system using triphone-based acoustic approach gives better results (word correction rate, word accuracy rate and word error rate) as compared to the word-based acoustic approach.

Chapter 7

Conclusion and Future Scope

This chapter has been organized as follows. Section 7.1 gives the concluding remarks over the work performed in the thesis. Some future works have been discussed in the section 7.2.

7.1 Concluding Remarks

Thesis provides the brief introduction related to the basic concepts of speech recognition which include methods for increasing the accuracy and ease of use of speech recognition systems, the types and categories of the speech recognition system, basic model of speech recognition system, fundamental problems and issues of ASR design, various approaches and methods to speech recognition. Also, Hidden Markov Model and HMM-based speech recognition system have been described. The obstacles in using SR system and applications of the proposed system also have been discussed. The review of literature of the proposed system has been discussed.

This thesis mainly deals with the implementation of HMM-based speech recognition system. Two different approaches are used for modeling the system. One is word-based approach as acoustic model and another one is triphone-based approach as acoustic model. The presented system can recognize both connected speech and isolated words; also the system is speaker independent.

It has been observed from the performed experiments that optimal system is obtained by using different number of states for different word models. In the system proposed by us, when numbers of states used for each model are 5, 6, 7 and 8 then respective recognition accuracy for connected word recognition is 81.28%, 81.62%, 92.55% and 92.76%. The recognition accuracy is 98.72% when numbers of states are different for different word models in the proposed system. The number of states present in a particular word model is decided based upon the number of phonemes present in the corresponding word and the duration of that word.

The respective recognition accuracy of triphone-based and word-based system for isolated word recognition is 96% and 95% in room environment, and 94% and 93.67% in open space environment; and respective recognition accuracy of triphone-based and word-based system for connected word recognition is 94.94% and 94.04% in room environment, and 93.75% and 92.26% in open space environment. So, it can be concluded here that speech recognition system using triphone-based acoustic approach gives better results as compared to the word-based acoustic approach. Also, the results conclude that accuracy of the system is sensitive to the changing environment.

7.2 Future Scopes

The work can be further extended in many directions.

- The thesis deals with the implementation of HMM-based speech recognition system for small vocabulary size. So, the work can be further extended to large vocabulary size.
- The system is implemented for recognizing the isolated and connected words. So, the work can be extended from connected words to continuous words or spontaneous speech.
- The proposed system has been developed for English language. So, it can be trained for other languages also.
- The results conclude that accuracy of the system is sensitive to the changing environment. So, noise compensation/speech enhancement techniques can be used so that system obtained is more accurate and efficient in noisy environments.

REFERENCES

[1] A. Srinivasan, "Speech Recognition Using Hidden Markov Model", Applied Mathematical Sciences, Vol. 5, No. 79, pp. 3943 – 3948, 2011.

[2] A.P.Varga, and R.K.Moore, "Hidden Markov Model Decomposition of Speech and Noise, Proc. ICASSp, pp.845-848, 1990.

[3] B. Gold, "A Neural Network for Isolated Word Recognition", IEEE International Conference on Acoustics, Speech and signal processing, 1988.

[4] C.H.Lee, and L.R.Rabiner, "A Frame Synchronous Network Search Algorithm for Connected Word Recognition", IEEE Trans. Acoustics, Speech, Signal Proc., Vol. 37, No. 11, pp. 1649-1658, November 1989.

[5] C.S.Myers, and L.R.Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. Acoustics, Speech Signal Proc., ASSP-29, pp. 284-297, April 1981.

[6] D. Raj Reddy, "Speech Recognition by Machine: A Review", Proc. of the IEEE, April 1976.

[7] D.R. Reddy, An Approach to Computer speech Recognition by direct analysis of the speech wave,Tech.Report No.C549,Computer Science Department, Stanford University, Sept. 1996.

[8] D.Shakina Deiv, Gaurav, and Mahua Bhattacharya, "Automatic Gender Identification for Hindi Speech Recognition", International Journal of Computer Applications, Vol. 31, No. 5, October 2011.

[9] Dat Tat Tran, "Fuzzy Approaches to Speech and Speaker Recognition", A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra.

[10] F. Itakura, "Minimum Prediction Residual Applied to Speech recognition", IEEE Trans Acoustics, Speech, Signal Proc., ASSP-23(1), pp. 67-72, 1975

[11] G. 2003 Lalit R .Bahl et.al., "Estimating Hidden Markov Model Parameters so as to maximize speech recognition Accuracy", IEEE Transaction on Audio, Speech and Language Processing, Vol.1 No.1 , Jan.1993.

[12] G. Guo, and S. Z. Li, "Content Based Audio Classification and Retrieval by Support Vector Machines", IEEE Transactions on Neural Networks, Vol. 14 No. 1, pp. 209-215, 2003.

[13] H. F. Olson, and H. Belar, “Phonetic Typewriter”, J. Acoust. Soc. Am., Vol. 28, No. 6, pp. 1072-1081, 1956.

[14] H. Hermansky, “Perceptually linear Predictive (PLP) analusis of speech”, Journal of Acoustic society of America, Vol. 87, No. 4, pp.1738-1752, 1990.

[15] H.Sakoe, and S.Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1).pp.43- 49, 1978.

[16] HTK “Hidden Markov Model Toolkit”. Obtained through the Internet http://htk.eng.cam.ac.uk, 2013, [accessed on Jan. 5, 2013].

[17] J D Ferguson, “Hidden Markov analysis: an introduction” In Hidden Markov models for speech. Princeton: Institute for Defense Analyses 1980(a).

[18] J D Ferguson, “Variable duration models of speech” Application of Markov models to text and speech, Princeton, Institute for Defense Analyses 1980(b).

[19] J. D. Markel, and A. H. Gray Linear Prediction of Speech, New York: Springer-Verlag.

[20] J. Hai, and E. M. Joo, “Improved Linear Predictive Coding method for Speech Recognition, Information, Communication and signal processing, 2003.

[21] J.Suzuki, and K.Nakata, “Recognition of Japanese Vowels Preliminary to the Recognition of Speech”, J.Radio Res.Lab, Vol. 37, No. 8, pp. 193-212, 1961.

[22] K. Nagata, Y. Kato, and S. Chiba, “Spoken Digit Recognizer for Japanese Language”, NEC Res. Develop., No. 6, 1963.

[23] K. S. Rao, and B. Yegnanarayana, “Modeling Syllable Duration in Indian Language using Neural Networks”, Acou. Speech Signal Processing, Canada, pp. 313-316, 2004.

[24] Kuldeep Kumar, R. K. Aggarwal, and Ankita Jain, “A Hindi speech recognition system for connected words using HTK”, Int. J. Computational Systems Engineering, Vol. 1, No. 1, pp.25–32, 2012.

[25] L. E. Baum, “An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes, Inequalities”, Vol. 3, pp. 1-8, 1972.

[26] L. R. Bahl, P. F. Brown, P. V. deSouza, and L. R. Mercer, “Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition”, Proc. ICASSP 86, Tokyo, Japan, pp. 49-52, April 1986.

[27] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”, Annals of Mathematical Statistics, Vol. 41, No. 1, pp.164-171, 1970.

[28] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proc. of the IEEE, Vol. 77, Issue 2, pp. 257–286, 1989.

[29] L.R.Rabiner, and B.H.Juang, “Fundamentals of Speech Recognition”, Prentice-Hall, Englewood Cliff, New Jersy, 1993.

[30] L.R.Rabiner, and S. E. Levinson, “Isolated and Connected Word Recognition Theory and Selected Applications’, IEEE Transactions on Communications, Vol. 29, No. 5, pp. 621-659, 1981.

[31] L.R.Rabiner,, and J. G. Wilpon, “A simplified, robust training procedure for speaker trained, isolated word recognition systems”,Acou. Society of America, Vol. 68, No. 5, 1980.

[32] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer, “Estimating Hidden Markov Model Parameters So As To Maximize Speech Recognition Accuracy”, IEEE transactions on speech and audio processing, Vol. 1, No. 1, January 1993.

[33] Lalit R.Bahl et.al, “Estimating Hidden Markov Model Parameters So as to maximize speech recognition Accuracy”, IEEE Transactions on Audio, Speech and Language processing Vol.1, No.1, Jan.1993.

[34] Li. Deng, D. O’Shaughnessy, Speech Processing- A Dynamic and Optimization- Oriented Approach. Chapter 12, Marcel Dekker Inc. New York,2003.

[35] Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon, “An Improved Endpoint Detector for Isolated Word Recognition”, IEEE transactions on acoustics, Speech, and signal processing, Vol. 29, No. 4, August 1981.

[36] M.A.Anusuya, and S.K.Katti, “Speech Recognition by Machine: A Review”, International Journal of Computer Science and Information Security, Vol. 6, No.3, 2009.

[37] M.Weintraub et.al, “linguistic constraints in hidden markov Model based speech recognition”, Proc.ICASSP, pp. 699-702, 1989.

[38] Marvin R. Sambur, and Lawrence R. Rabiner, "A Statistical Decision Approach to the Recognition of Connected Digits", IEEE transactions on acoustics, Speech, and signal processing, Vol. 24, No. 6, December 1976.

[39] Mohit Dua, R.K.Aggarwal, Virender Kadyan, and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No. 1, July 2012.

[40] N Rai Isolated word speaker Independent Speech recognition for Indian Languages, Department of Computer science and Engineering, Indian Institute of Technology, Kanpur, India, 2005.

[41] Nigel Ward, "Artificial Intelligence and Other Approaches to Speech Understanding: Reflections on Methodology", Journal of experimental and theoretical artificial intelligence, Vol. 10, pp. 487-493, 1998.

[42] Pinki Roy, and Pradip K. Das, "A hybrid VQ-GMM approach for identifying Indian languages", Int J Speech Technology, 2013.

[43] R K Aggarwal, and M. Dave, "Markov Modeling in Hindi Speech Recognition System: A Review", CSI Journal of Computing, Vol. 1, No.1, pp. 38-47, 2012.

[44] R. Gupta, Speech Recognition for Hindi, Master's Project Report, Department of Computer science and Engineering, Indian Institute of Technology, Bombay, Mumbai, India, 2006.

[45] R. K. Aggarwal, and M. Dave, "Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I)" International journal Speech Technology, Springer, Vol.14, Issue 2, 2011.

[46] R. K. Aggarwal, and M. Dave, "Fitness Evaluation of Gaussian Mixtures in Hindi Speech Recognition System", First International Conference on Integrated Intelligent Computing, SJB Institute of Technology, Bangalore, 2010.

[47] R. K. Moore, Twenty things we still don t know about speech, Proc. CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology , 1994.

[48] R.P.Lippmann, "An introduction to computing with neural nets", IEEE ASSP Mag., Vol. 4, No. 2, pp. 4-22, April 1987.

[49] Rabiner, L.R. and Huang, "An introduction to hidden Markov models", IEEE Acoust., Speech Signal Processing, Vol. 4, No. 16, pp.4–16, 1986.

[50] Ravinder Kumar, "Comparison of HMM and DTW for Isolated Word Recognition of Punjabi Language" In Proceedings of Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Sao Paulo, Brazil. Vol. 6419 of Lecture Notes in Computer Science (LNCS), pp. 244–252, Springer Verlag, November 8-11, 2010.

[51] Ravinder Kumar, and Mohanjit Singh, "Spoken isolated Word Recognition of Punjabi Language Using dynamic time Warping Technique", Communication in Computer and Information Science (CCIS), Vol. 139, Page 301, Springer Verlag, 2011.

[52] Rupayan Das, and Pradip K. Das, "Design and Implementation of Monophones and Triphones-based Speech Recognition Systems for Voice Activated Telephony", BIJIT, Vol. 5, No. 1, ISSN 0973 – 5658, 2013.

[53] S. Davis, and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing Vol. 28, No. 4, pp. 357–366,1980.

[54] Santosh K.Gaikwad, Bharti W.Gawali, and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications, Volume 10, No.3, November 2010.

[55] Wiqas Ghai, and Navdeep Singh, "Literature Review on Automatic Speech Recognition", International Journal of Computer Applications, Vol. 41, No.8, pp. 42-50, March 2012.

[56] X D Hauang,Y Ariki, and M A Jack, "Hidden Markov Models for speech recognition", Edinburg University Press, 1990.