4

SPEECH RECOGNITION & FEATURE EXTRACTION APPROACHES

CONTENTS

OBJECTIVE OF THE CHAPTER

The objective of this chapter is to analyze the different feature extraction methods applied to the Bodo language. Since feature extraction is one of the most crucial first steps in any classification and recognition task, this chapter actually provides information on various techniques that can be used for taking this first step towards much more complex speech recognition tasks. With the information presented in this chapter, it is expected that the readers will be able to get a sense of how different feature extraction techniques fare when we particularly consider the Bodo language.

4.1 SPEECH RECOGNITION

Speech recognition may be defined as the ability of computer systems to input the audio signal and then output the same in a text file. In the field of computer it is defined as the ability of computer systems to accept spoken words in audio format – such as wav or raw - and then generate its content in text format. This speech recognition comprises of two types namely isolated word recognition and connected word recognition. A typical speech recognition system starts with a preprocessing stage, which takes a speech waveform as its input, and extracts its feature vectors to perform recognition. Feature Extraction is the process in which sound signal is converted to a form suitable for processing by the system which generally includes extracting parameters like amplitude of the signal, energy of frequencies, etc. Feature extraction phase determines the phonemes location and their waveform characteristics. This stage is efficiently performed by software. The second stage is decoding, which performed using a set of phoneme-level statistical models is called Hidden Markov Models (HMMs). It uses pattern recognition techniques to identify the phonemes, and maps these phonemes into words. Word-level acoustic models are formed by concatenating phone-level models according to a pronunciation dictionary. These word models are then mixed with a language version, which constrains the recognizer to understand only valid word sequences. The decoder stage is computationally expensive.

Development of speech recognition system comprises of two phases: Training and Recognition. In the training phase, one or more acoustic patterns or templates of linguistics units such as the words or phonemes in the recognition vocabulary are derived. In the recognition process, the incoming speech files are matched with the stored acoustic patterns. The stored template pattern with the lowest distance measure from the input pattern is the recognized word. In early days Dynamic Time Wraping (DTW) algorithm is used for identifying the best match between two given sequences. In modern system, the pattern-matching is modeled as a probabilistic process to account for temporal and acoustic variability in the input speech signal. The sequential inconsistency in the speech signal arises due to differences in speaking rate within speaker and across speakers and the impulsive nature of speech. The acoustic variability arises due to linguistic phenomena like co-articulation, speaker characteristics like age, gender, physical attributes, and channel factors like e.g., background noise, transmission network or microphones [36][37] etc. Training a model that is able to account for all these variations is a complex process. In the probabilistic framework, the speech recognition process is represented with the equation (5.1) [35].

$$\mathbf{W^* = argmaxW\ P(O|W)\ P(W)} \qquad \mathbf{(5.1)}$$

Here, W* represents the word sequence that has the maximum a posteriori probability, and O represents the observation that is derived from the speech segment. To estimate the maximum a posteriori probability, the model relies on two probabilities:

- P(O|W) : the probability that a sequence of words W would produce the observation O. This is referred to as the acoustic model and the probabilities or likelihoods are derived from a corpus of recorded audio and corresponding transcriptions.
- P(W): the probability of a sequence of words W in a language. This is referred to as the language model and the probabilities are derived from the corpus of a language.

The observation O is extracted from the speech signal and comprises of a series of vectors representing the acoustic feature of speech. To obtain O, the speech signal is first converted into a digital format and split into short frames of about 10 ms (generally a sliding frame is used). From each frame a set of acoustic features are extracted using digital signal processing techniques [17]. One of the goals during feature extraction is to give importance to perceptually important speaker-independent features and discard the redundant features.

In ASR systems for large vocabulary continuous word recognition the acoustic models are of some units of sound such as phonemes or other linguistic units that make up speech. Larger sequence of phones, such as triphones are more useful as they model a sound in the context in which it is used, i.e., in terms of the sounds preceding it and sounds succeeding it. The process of mapping the relation between the continuous speech signal and the discrete sounds of the words is known as acoustic modeling. In modern general-purpose speech recognition systems acoustic modeling is done using Hidden Markov Model (HMM), in which states are represented by the phones while the observations corresponds to the acoustic features

From an application point, the most important element of the ASR systems is the language model P (W). The language model captures the context in which words are more likely to occur in a language. Here again, for small vocabulary systems the knowledge can be encoded using simple grammar rules (e.g. a context-free grammar). From these rules a finite state network model can be derived that covers all the possible word sequences in the language. For large vocabulary continuous speech recognition, N-gram model are common in use. In N-grams models, probability of a sequence of words is computed as a product of the probabilities of each word [17]. This assumes that the occurrence of each word is determined by the preceding N - 1 word. However, in practice estimating the probability of a word given a large number of preceding words can be computationally expensive (and may lead to overtraining) and therefore N-grams are usually limited to bigram (N=2) or trigram (N=3).

4.1.1 EVALUATION OF SPEECH RECOGNITION

Speech recognition is a process by which a program or a system transcribes an acoustic speech signal to text. Systems generally perform two different types of recognition: single-word and continuous speech recognition. Continuous speech is more difficult to handle because of a variety of effects such as speech rate, co-articulation, etc. Today’s state-of-the-art systems are able to transcribe unrestricted continuous speech from broadcast data with acceptable performance.

The performance of an ASR system is measured in terms of word accuracy or more commonly as word error rate (WER). WER is a minimum edit-distance measure produced by applying a dynamic alignment between the output of the ASR system and a reference transcript. In the alignment process, three different types of errors can be distinguished namely substitutions (*sub*), deletions (*Del*) and insertions (*ins*). Substitution-errors occur due to wrongly transcribed word. For example Word – P is transcribed as Word – Q .Insertions and Deletions concern words which were transcribed in addition to reference words or words omitted during transcription. An insertion error is when a word is recognized when none was spoken. A deletion error is when no word is recognized when one was spoken. And a substitution error is when the wrong word was recognized. WER is calculated using equation 5.2[17, 35].

- Substitution: a reference word is replaced by another word in the best alignment between the reference and the system hypothesis.
- Deletion: a reference word is not present in the system hypothesis in the best alignment.
- Insertion: Some extra words are present in the system hypothesis in the best alignment between the reference and the hypothesis.

Performance of ASR systems is also measured in terms of speed by measuring the processing time and computing the real time factor on a specific hardware configuration. This is an important factor for some applications that may require a real-time processing speed or some devices that are limited in terms of memory or processor speed.

$$\mathbf{WER = (S + D + I) / N} \qquad \mathbf{5.2}$$

In equation (5.2), S, D and I are the number of word substitutions, deletions, and insertions in the reference sequence, respectively. N is the total number of words in the reference string. Word accuracy is than obtained as:

1 - WER.

WER provides an important measure in determining performance on a word-by-word level and is typically applied to measure progress when developing different acoustic and languages models, it only provides one angle of system performance and can be misleading when viewed in isolation.

Although word is the basic unit for assessing ASR systems, the same computation can be made using different granularities (phonemes, syllables, etc.) WER can be greater than 100%, if the number of errors is more important than the number of words. Prior to scoring both hypothesis and reference have to be normalized. The normalization consists of converting the transcription into a more standardized form. This step is language dependent and applies a number of rules for transforming each token into its normalized form. For instance numbers are spelled out, punctuation marks are removed, contractions are expanded, multiple orthographies are converted to a unique form, etc. Although WER is the main metric for assessing ASR system, its major drawback is that all word errors are equally penalized, regardless the importance and meaning of the word, e.g. an empty word has the same importance as a named entity.

4.1.2 CHALLENGES IN BUILDING AN ASR SYSTEM

Training and deploying ASR systems for command-base dialogue systems are a more tractable task than a system for conversational interactions. Systems for supporting conversational speech would need to rely on N-gram based language models for the extended coverage [17]. This leads to two issues, first, a larger vocabulary requires tremendous amount of data for robustness in coverage, but a large vocabulary would also imply a language model with high perplexity, which may affect the recognition accuracy.

In addition to vocabulary size, ASR systems have to deal with other factors such as: speaker independence (training a speaker independent acoustic model), continuous speech (since there is no physical separation in continuous time speech signal it is difficult to determine word boundaries); spontaneous conversational speech (the speech to dialogue systems is spontaneous and unplanned and contain disfluencies, such as hesitations, fillers, restarts and revisions) [17]. Depending on the end goal of the dialogue system, the ASR systems may require training for detection of these linguistic phenomena, which can be vital cues for modelling user state and discourse flow.

4.1.3 SPEECH RECOGNITION MODELLING

This ASR module takes speech as input and returns a string of words. But due to the different pronunciation, background noise, emphasis, pauses common systems have high error-rate. An ASR consists of models, trying to model speech from the real world. In addition, there is a speech engine which contains all the computational logic, making use of the speech models to transcribe speech into text. Two different models are needed for speech, firstly, there is the low level acoustic model, modeling how words sound, and secondly, there is the language model, modeling how words build language [20].

a. ACOUSTIC MODELS: An acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. The model is learned from a set of audio recordings and their corresponding transcripts. An acoustic model represents how words sound. It is commonly created by recording speech with text transcriptions, so that the speech and text can be matched. Acoustic model is sensitive to the environment, such as noise and voice characteristics. To achieve good speech recognition results, the acoustic model must be based on large number of speech corpus. In this subsystem, the connection between the acoustic information and phonetics is established. Speech unit is mapped to its acoustic counterpart using temporal models as speech is a temporal signal. There are many models for this purpose like, Hidden Markov Model (HMM), Artificial Neural Network (ANN), Dynamic Bayesian Network (DBN) etc. ANN is a general pattern recognition model which found its use in ASR in the early years. Rabiner (1991), first suggested the HMM approach leading to substantial performance improvement. Current major ASR systems use HMM for acoustic modeling. Since then researchers have tried to optimize this model for memory and computation requirements [39].

b. LANGUAGE MODELS: A language model is mainly used to restrict word search. It defines which word could follow previously recognized words and helps to significantly restrict the matching process by stripping words that are not probable. The goal of language modeling is to produce accurate value of probability of a word W, Pr (w).It determines the probability of a word occurring after a word sequence. To reach a good accuracy rate, language model must be very successful in search space restriction. This means it should be very good at predicting the next word. A language model usually restricts the vocabulary that is considered to the words it contains. A language model contains the structural constraints available in the language to generate the probabilities. The method and complexity of modeling language would vary with the speech application. This leads to mainly two approaches for language modeling [39]. Several approaches to language models for ASR systems exist. There are mainly two types, the grammar based model and the statistical model:

- **GRAMMAR BASED LANGUAGE MODELS:** In this system a small grammar is given for each point in a human-computer dialog, a simple example is the grammar for digit strings: Grammar based language models describe the domain language, trying to cover all potential utterances that a user may wish to speak. Utterances are not ranked in any way. The ones that are finally recognized by the ASR are chosen entirely based on how well they match the acoustic model.

- **STATISTICAL LANGUAGE MODELS:** Statistical language models (SLM) are usually based on probabilities of word sequences. Probabilities are calculated by counting occurrences of words and word sequences in a corpus, a collection of text. A statistical language model is a probability distribution over sequences of words .It assigns any sentence a probability it's a probability of seeing a sentence according to the LM. Language Models can be context dependent. For example:

p1=P ("Today is Sunday") =0.01
p2=P ("All equation has some solution") =0.0001

Although both sentences are valid, sentence 1 is more likely: for example, because the LM can model some general conversations rather than conversations on a math conference

4.2 REVIEW ON FEATURE EXTRACTION APPROCHES

Feature extraction is the most important part of speech recognition due to the fact that it plays an important function to separate one speech from other. Each speech has different individual characteristics embedded in utterances. These traits can be extracted from a huge range of characteristic extraction techniques proposed and successfully exploited for speech recognition task. Extracted feature must meet some criteria while dealing with the other speech signal such as [40]

a. Easy to measure extracted speech features

b. It has not be prone to mimicry

c. It should to show little fluctuation from one speaking surroundings to another

d. It has to be stable over time

e. It should occur often and naturally in speech

After extracting the features, speech signals are applied to speech recognizers. The most widely used feature extraction techniques are LPC, MFCC etc. Following analysis gives the advantage and disadvantage of different feature extraction techniques.

Principle component analysis (PCA) is a liner map, rapid, nonlinear feature extraction technique based on Eigen vector. It gives good result for Gaussian data. But maximizing direction variance does not always maximize information.

Linear Discriminate Analysis (LDA) is a supervised liner map non liner feature extraction techniques which depend on Eigen vector. It is better than PCA and performance is examined on the randomly generated test data. If the distribution is significantly non-Gaussian the LDA projection will not be able to preserve any complex structure of the data, which may be needed for classification.

Independent Component Analysis (ICA) is a linear map, iterative non Gaussian nonlinear feature extraction technique. It is better than PCA for classification. But the main problem with method is that extracted components are not ordered.

Linear Predictive Coding (LPC) is a 10 to 16 lower sequence coefficient, static feature extraction method. Here spectral analysis is done with a fixed resolution along a subjective frequency scale i.e. Mel frequency scale. Frequencies are weighed equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic.

In Filter Bank Analysis method filter tuned required frequencies. It provides a spectral analysis with any degree of frequency resolution, even with non-linear filter spacing and bandwidths. It always takes more calculation and processing time than discrete Fourier analysis using the FFT

In Mel-Frequency Cestrum Coefficients (MFCC) power spectrum is computed implementing Fourier analysis. MFCC values are not very robust in the presence of additive noises it is common to normalize their values in speech recognition system to reduce the influence of noise.

Kernel based feature extraction is a nonlinear transformations method. Its dimensionally reduction procedure leads to better classification. It is used to remove noisy and redundant features and improves classification error. But similarity calculation speed is slow in this method.

Wavelet techniques replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allows better time resolution at high frequencies than Fourier transform. It required longer compression time.

Cepstral Mean Subtraction is a robust feature extraction technique. It is same as MFCC but working on Mean statically parameter.

RASTA Filtering technique is used in Noisy speech. It increases the dependence of the data on its previous context.

In our work we mainly study the feature extraction techniques mainly LPC and MFCC and combining those features with DTW and HMM for analysis the recognition of Bodo digits which are going to use in our propose spoken dialogue system.

4.3 REVIEW ON SPEECH RECOGNITION APPLICATIONS

In the following section we brief about the various speech recognition applications based on dataset, feature extraction and recognition approach.

Vimala C. and V. Radha [41] developed an isolated speech recognition system for Tamil language. The recognition system was found to be speaker independent. The experiments showed a high accuracy of 88% for trained and tested spoken words. Word Error Rate (WER) was used for the evaluation of the performance of the system and the system gave .88 WER.

Sunitha .K.V.N and Kalyani .N [42] designed a system that can recognize Telugu words. The system uses syllables as the basic units. CIIL Telugu corpus was used for this purpose and a good result was achieved in recognizing the words. The system was able to recognize other words too which were not used in training. 300 words were used for training purpose and testing was done on 300 new words. The system was able to recognize 80% of the words correctly.

Vimal Krishnan [43] developed a speech recognition system for Malayalam language with a small vocabulary using four types of wavelets in feature extraction. Artificial neural network (ANN) is used for the purpose of classification and recognition of the data. Using this method they were able to achieve an accuracy rate of 89%.

Ravinder [44] developed an isolated and connected word recognition system for Punjabi language. The system was speaker dependent and real time system. Technique of Acoustic template matching was used for the development of the recognition system. The system was designed for a medium sized dictionary. AN accuracy of only 61% was achieved with the system.

N. Rajput M. Kumar and A. Verma [45] of IBM research lab developed a continuous speech recognition system for Hindi language. A large vocabulary of about 65000 words was used for the development of the system. The system gave a word accuracy of 75% to 95%.

A small vocabulary speech recognizer has been developed by Anuj Mohamed and K.N Ramachandran Nair [46] using Hidden Markov Models in Malayalam language. The system has produced 94.67% word accuracy.

In 2010, M. P. Sarma and K.K Sarma [47] worked for the development of numeral speech recognition system for Assamese language. Gender and mood variations were given consideration during the recording of speech signals of 10 numeral digits at 8 KHz in mono channel mode. In 2011, M. P. Sarma and K.K Sarma have proposed the design of an optimal feature extraction block and ANN based architecture for speech recognition.

Neural network approach has been proposed by M. R. Hassan [48] for Bengali phoneme recognition. A Bengali speech recognizer is built by training the HTK toolkit that can recognize any word in the dictionary. After acoustic analysis of speech signal, the words are recognized. Technically this work presents training the toolkit and builds a segmented speech recognizer of Bengali.

Mohanty and Swain [49] have made such effort for Oriya language. They have come forward to apply the benefit of automatic speech recognition systems to society by developing an isolated speech recognizer for Oriya language.

A technique for fast bootstrapping of initial phone models of a Gujarati language is presented by Himangshu N. Patel [49]. The training data for the Gujarati language is aligned using an existing speech recognition engine for English language. This aligned data is used to obtain the initial acoustic models for the phones of the Gujarati language. Speech recognition of Gujarati Language is presented by Patel Pravin and Harikrishna Jethva. Neural network was used for developing the system.

In 2003, Sid-Ahmed Selouani, Yousef Ajami Alotaibi [49] were investigating automatic recognition of non-native Arabic speech with large vocabulary speaker independent phonetic/word using MFCC features and HMM classifier on Arabic language. They found that new words make less accuracy for non-native speakers.

In 2004 A.P.Henry,Charles & G.Devaraj [49] performed there research work on Speaker independent Continuous Tamil speech recognition system using MFCC and HMM which offers very high performance.

In 2006, Meysam Mohamad pour, Fardad Farokhi [49] were design an isolated digit recognition system using Discrete Wavelet Transform (DWT) features and Multilayer Perceptron and UTA algorithm as classifier on English with 98% accuracy. In 2006, Corneliu Octavian Dumitru, Inge Gavat did a comparative study on feature extraction methods applied to large vocabulary speaker independent continuous speech recognition in Romanian Language. They used PLP, MFCC, LPC is features extraction techniques and HMM as classifier. They found 90.41% accuracy on MFCC, 63.55% on LPC and 75.78% on PLP.

In 2007, ARathinavelu, G.Anupriya, A.S.Muthanantha Murugavel [49], design a speech recognition model for Tamil Stops using Small vocabulary speaker independent phonemes. They took first five formant values with Feed forward neural networks and got 81% accuracy.

In 2008, M.Chandrasekar, and M.Ponnavaikko [49], designed a medium vocabulary speaker dependent isolated speech recognition system on Tamil language. They reported 80.95% of accuracy using MFCC and back propagation network classifier.

In 2009, Ghulam Muhammad, Yousef A. Alotaibi [49], and Mohammad Nurul Huda, design a small vocabulary speaker independent isolated digit recognition system in Bangla. They used MFCC features and HMM. They got more than 95% for digits (0-5) and less than 90% for digits (6-9)

In 2010, Zhao Lishuang, Han Zhiyan [49], design a large vocabulary speaker independent vowel recognition system integrating MFCC feature and HMM classifier with genetic algorithm on Chinese language. They found this combination is effective and high speed and accuracy. Raji Sukumar, A Firoz Shah, A Babu Anto.P also developed isolated question words recognition from speech queries by using artificial neural networks using DWT and ANN in Malayalam languages. They got 80 % accuracy.

Bassam A. Q. Al-Qatab, Raja N. Ainon [49] builds an Arabic speech recognition using Hidden Markov Model Toolkit (HTK) with MFCC and HMM getting 97.99 % accuracy. Javed Ashraf, Dr Naveed Iqbal, Naveed Sarfraz Khattak, Ather Mohsin Zaidiuild [49], build a speaker independent Urdu speech recognition using MFCC and HMM where they found little variation in WER for new speakers.

N.Uma Maheswari, A.P.Kabilan, R.Venkatesh [49] design a hybrid model of neural network approach for Speaker independent Word Recognition using LPC and hybrid model of radial basis function and the pattern matching method in English with a 91 % accuracy. R.Thangarajan, A.M. Natarajan and M. Selvam [49] designed and Phoneme Based Approach in Medium Vocabulary Continuous Speech Recognition in Tamil language using MFCC and HMM. They found good word accuracy for trained and test sentences read by trained and new speakers

M.K. Deka has proposed an approach for Speech Recognition using LPCC (Linear Predictive Cepstral Coefficient) and MLP (Multilayer Perceptron) based Artificial Neural Network with respect to Assamese and Bodo Language [50]. A new simplified approach has been made for the design and implementation of a noise robust speech recognition using Multilayer Perceptron (MLP) based Artificial Neural Network and LPC-Cepstral Coefficient. Cepstral matrices obtained via Linear Prediction Coefficient are chosen as the eligible features. Here, MLP neural network based transformation method is studied for environmental mismatch compensation.

Utpal Bhattacharjee [4] investigates the problems faced by tonal languages like Bodo during recognition process. The performance of speech recognition system degrades considerably when the recognizers are used to recognize the tonal words. Two approaches have been investigated in this paper for this purpose. In the first approach attempt has been made to develop a feature level solution to the problem of tonal word recognition. In the second approach, a model level solution has been suggested. Experiments were carried out to find the relative merits and demerits of both the methods.

4.4 PROPOSED WORK

After analyzing the various work proposed by different researchers, we found that there has not done much research on these approaches in context of Bodo language. Earlier work does not show any analysis on different feature extraction approaches which gives better accuracy in real time environment and telephony network. So we investigate the LPC, MFCC as feature extraction techniques and DTW and HMM as classifier to find the accuracy with telephone network. Following section gives brief description of our work:

4.4.1 LINEAR PREDICTIVE CODING

The LPC method of speech analysis and synthesis is based upon the principle of linear prediction, which is a major aspect of time series analysis. Linear prediction is an auto-correlation domain analysis and that is why it can be applied from either the time or frequency domain. The least squares error criterion in the time domain translates into spectral matching criteria in the frequency domain. Due to the accuracy and speed of computation, these methods are appreciated very much. The basic idea behind the LPC model is that- a given speech sample can be approximated as a linear combination of the past p speech samples. By minimizing the squared differences between the actual speech samples and the linearly predicted samples, one can determine the predictor coefficients, the weighting coefficients of the linear combinations [51]. Following are some of the reasons why LPC is widely used in a speech-processing system:

a. LPC gives better speech signal modeling.
b. Using low dimension feature vectors LPC gives the spectral envelope.
c. Linear characteristics were given by LPC.
d. Acceptable source-vocal tract separation is obtained by LPC.
e. It is analytically traceable model.
f. LPC is easy to implement in both software and hardware.

4.4.2 MEL FREQUENCY CEPSTRAL COEFFICIENTS

The MFCC is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. Since MFCC actually has filter banks, its transformation doesn’t ensure accurate reconstruction of speech signal. i.e., if we are provide with only the actual features, reconstruction of the original speech signal used to build these features is not possible. The main reason for loosing of the information is the complex computations and not having good robustness. By defining more number of parameters one can improve the accuracy but by paying price of complexity. Even in such cases the robustness difficulties persists [52] [53].

4.4.3 DYNAMIC TIME WARPING

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly while in another they were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics. Indeed, any data which can be turned into a linear representation can be analyzed with DTW [54].

A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of Hidden Markov Models.

4.4.4 HIDDEN MARKOV MODEL

Modern general-purpose speech recognition systems are generally based on Hidden Markov Models. These are statistical models which output a sequence of symbols or quantities. One possible reason why HMMs are used in speech recognition is that a speech signal could be viewed as a piecewise stationary signal or a short-time stationary signal. That is, one could assume in a short-time in the range of 10 milliseconds, speech could be approximated as a stationary process. Speech could thus be thought of as a Markov model for many stochastic processes.

Another reason of popularity of HMM is that they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of *n*-dimensional real-valued vectors (with *n* being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians which will give likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phonemes (so phonemes with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminate analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminate analysis or a global semi tied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques which dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are Maximum Mutual Information (MMI), Minimum Classification Error (MCE) and Minimum Phone Error (MPE). The decoding of the speech is normally done using the Viterbi algorithm to find the best path, and thereafter allowing a choice between dynamically creating a combined Hidden Markov Model which includes both the acoustic and language model information, or combining it statically beforehand (the finite state transducer, or FST, approach)[55].

4.5 SIMULATION RESULTS

In this section we presents the experimental results obtained from the proposed approaches namely LPC, MFCC, Dynamic Time Warping, HMM that was applied to the isolated Bodo digits recognition. The effectiveness of the algorithms is measured through the analysis of the results. The below **Table 4.1** shows the recognition accuracy for all the combination proposed in the research with stored speech sample and with real time speech sample. From analysis we can conclude that, with comparison of all the techniques with stored speech signal is giving 85% accuracy but in case of real time due to noise, accuracy decreased even in noise DTW is giving the best accuracy rate of all the above combination.

Table 4.1: Bodo Digit Recognition Accuracy

WORD	**%Recognition Accuracy**					
	LPC+DTW		**MFCC+DTW**		**MFCC+HMM**	
	off-line	**on-line**	**off-line**	**on-line**	**off-line**	**on-line**
Lathikho	90	70	91	81	94	83
Nai	92	75	87	81	96	84
Se	93	85	88	86	92	89
Tham	87	83	94	82	94	85

Broi	79	72	92	76	89	81
Ba	81	76	96	78	93	88
Da	83	72	91	76	95	84
Sni	85	78	85	80	94	81
Dian	73	69	87	83	95	85
Gu	87	78	96	81	93	87
Average	**85**	**75.8**	**90.7**	**80.4**	**93.5**	**84.7**

4.5.1 RESULT ANALYSIS

a. Speech Recognition (LPC & DTW): This combination of recognition system giving 75% in real time system but in case of stored speech the result obtains is 85%. LPC is based on order as the order increases the recognition time is decreasing but there is no change in recognition accuracy.

b. Speech Recognition (MFCC & DTW): This combination of recognition system having more coefficients i.e. is 39 the performance of each digit is varying and also overall performance is better than LPC and also recognition time is improved.

c. Speech Recognition (MFCC & HMM): This combination of recognition system having more coefficients i.e. 39 the performance of each digit is varying and also overall performance is better than LPC and also recognition time is improved. The difference as compare to above combination is HMM this matching technique is time taking but accuracy of recognition is good.

4.5.2 SEEN, UNSEEN SOUND ANALYSIS

We propose our Bodo word recognition method in HTK [56] environment. The raw speech data is in the form of .wav files. This needs to be converted to MFCC speech vectors. This is a form of spectral analysis of the raw waveform. A prototype HMM model must be created, which must then be re-estimated using the data from the speech files. Silence models must be included. The prototype HMM model used is usually a 3 state left-to-right model with no skips. It has five states, but the first and last ones are dummy states which are used for continuity. The vector size used for the HMM models is usually 39, because it has been found to work well empirically. The final step of model building is to create context dependent triphones from the monophones. The set of triphones is created by cloning the monophones and re-estimating. Similar acoustic states of these triphones are tied to ensure that all state distributions can be robustly estimated. The test data can be recognized and the recognizer's performance can be evaluated. The implemented system is trained for 10 distinct Bodo digits which have been used in various combination modes. The data is recorded with the help of a unidirectional microphone using a recording tool wave surfer in .wav format. The sampling rate used for recording is 44 kHz. Recognition has been tried on three different kinds of sounds.

a. **Seen sound**: The sound files used to train the models.
b. **Unseen speaker**: The user whose voice was not used for training.
c. **Unseen sound seen speaker**: Unused sound files of the speaker who's other sound files were used for training.

Table 4.2: Recognition Accuracy in different set

Type of sound	No. of sound	Correct recognition %	Error rate %
Seen sound	200	89.05	10.95
Unseen sound	200	86.70	13.30
Unseen sound seen speaker	200	89.04	10.96

Overall accuracy of the system is 87.24 in training mode and 82.12 % in testing mode. In the experiment we have used 700 files to train the system and remaining 300 files are used to test the system.

Table 4.3: Word Recognition Accuracy of the System

Mode	Word Accuracy	No. of deletion	No. of substitution	No. of Insertion
Training	87.24	3	2	0
Testing	82.12	7	8	2

By considering the digit recognition analysis, the greatest accuracy was encountered in the case of digit 6, 1 and 2 while the least accuracy is encountered in the case of digit 9. The main cause for such a variation may be attributed to the tokens themselves; 9 is a monosyllabic Bodo digit which is pronounced as” /d/ द “which is short with lower amplitude than the other Bodo digits.

SUMMARY

Speech recognition may be defined as the ability of computer systems to input the audio signal and then output the same in a text file. During this phenomenon several signal processing issues takes place like background noise, pause etc. Therefor speech modeling is necessary. There are two types of model namely acoustic model and statistical model. An acoustic model is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. A language model is mainly used to restrict word search. Before recognizing the speech, feature of the speech signal were extracted using LPC, and MFCC. Out of these two methods MFCC feature extraction technique shown better results, and this features can be applied to recognition techniques like Dynamic time warping and Hidden Markov Models. Different combinations of feature extraction methods and recognition techniques are also implemented to recognize isolated Bodo words through this chapter. Seen and unseen speaker recognition is studied in this chapter which helps to build speaker adaptation module in future. The overall study of this chapter helps to build online recognition module and choose based feature extraction and classifier combination in telephone speech.