IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 5, May 2016*

# Identification of Isolated Upper and Lower Assamese words: A Survey

**Mr. Adarsh Pradhan[1], Jyoti Rekha Saikia[2]**

Assistant Professor, CSE, GIMT, Guwahati, India[1]

M.Tech., 4th Sem, CSE, GIMT, Guwahati, India[2]

**Abstract**: Communication is the fundamental purpose of speech, i.e., the transmission of messages. Speech is the most common form of interaction between human to human which is very natural and efficient. It conveys the information about words, speaker identity, expression, emotion, gender of the speaker. Now-a-days, both speech and speaker recognition are the two most important research areas. In this paper we try to recognize Assamese spoken words. This paper describes feature extraction technique Linear Predictive Coding (LPC) , Mel Frequency Cepstral Co-efficient (MFCC) and for classification Artificial Neural Network(ANN) are used to recognize Assamese speech samples (words).We create a small database for seven Assamese isolated words which consist of 16 speakers of equal numbers of male and female. Each word is uttered by seventy five times by each speaker.

**Keywords**: Speech Recognition, LPC, MFCC, DTW, HMM, Neural Network.

## I. INTRODUCTION

Speech recognition is also known as Automatic Speech Recognition (ASR) or computer speech recognition. It is a process through which a speech signal is converted to a sequence of words, by means of an algorithm implemented as a computer program. Speech recognition technology has made it possible for computer to follow human voice commands and then understand human languages. To develop techniques and systems for speech input to machine is the main goal of speech recognition area [11].

System recognition systems can be divided into different classes based on the type of speech utterance (Isolated, continuous or spontaneous),type of speaker model (speaker dependent or speaker independent), and the type of vocabulary (small ,medium, large, very large, out of vocabulary) [9].

The Assamese (IPA: oxomija) is a major language in the north-eastern part of India whose origin root is Indo European family of languages [1].In Assamese language there are thirty two phonemes out of which eight are vowel phonemes and twenty four are consonant phonemes. Assamese script derived from Devanagari scripts consists of thirty nine consonant and eleven vowel symbols [3].

## II. SPEECH RECOGNITION PROCESS

Speech is the most widely used and natural form of communication between humans in our live. Hence the development of robust speech technology is necessary which is a complex and challenging task.

Main task involved in speech recognition are pre-processing of speech signal, feature extraction and then pattern matching. The following fig 1 involved the steps of speech recognition process:

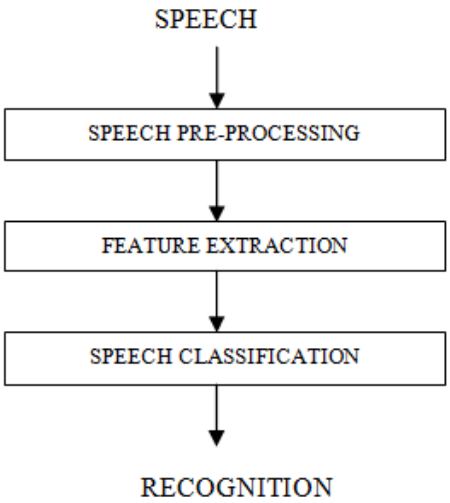


Fig. 1. Speech Recognition Process

A. For speech data, the speech of a person is received in a waveform which is recorded through many types of software and then transformed it into discrete data.

B. Pre-processing is considered in the development of robust and efficient speech or speaker recognition system. It also improves the readability accuracy of automatic speech recognition.

C. As the physiological structure of a vocal tract is different for every person, so for every individual person speech differs from each other. Feature extraction is used to separate one speech from the other. LDA, PCA, LPC, MFCC are some of the feature extraction technique.

D. Classification or recognition phase include different methods in identification of the speakers.ANN, HMM, DTW are techniques used for speech classification.

## III. LITERATURE REVIEW

Bhargab Medhi et.al [3] proposes an approach to recognise Assamese speaking person using Artificial Neural Network. LPC and MFCC are used to create the feature vector of the Assamese speech samples. The main database consists of ten speakers with equal number of male and female speakers where each word is uttered by twenty times by each speaker. The system contains the training phase, testing phase and recognition phase.

In 2015 Bhargab Medhi et.al [4] describes the features parameters Zero Crossing Rate (ZCR), Short-time Energy (STE), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) are analyzed which are extracted from Assamese vowel phonemes, uttered by Assamese Speaker. The authors create a small database of all eight Assamese vowel Phonemes, repeated each vowel phoneme in ten times, of total number of ten speakers with equal number of male and female speaker. Thus, their database consists of eight hundred phonemes. These parameters, extracted from each vowel phonemes, are useful to design an Automatic Assamese Speech and Speaker Recognition system.

In 2013 Utpal Bhattacharjee [5] describes LPC, MFCC and their performances have been evaluated for the recognition of Assamese phonemes. A multilayer perceptron based baseline phoneme recognizer has been built and all the experiments have been carried out using that recognizer.

Deshmukh Aniket Anand et.al [6] examined isolated speech recognition viz., pre-processing, feature extraction, pattern matching and their implementation. Different existing methods, e.g., the Dynamic Time Warping (DTW), Hidden Markov Model (HMM) for implementing speech recognition are discussed and implemented for isolated digit recognition.

A new method, which combines Linear Prediction Coefficients and Mel-Frequency Cepstral Coefficients features, has been proposed and recognition has been done using DTW. The performance of the proposed method is found to be better than existing methods.

In 2013 Akalpita Das et.al [7] tries to propose an approach to recognize isolated Bodo spoken words. They deal with a speech feature extraction using MFCC and along with K-mean clustering. In real time that system was proposed, speaker dependent and independent word recognition systems for limited number of words.

In 2010 Wouter Gevaert, Georgi et.al [8] presented an investigation of the speech recognition classification performance. This investigation on the speech recognition classification performance is performed using two standard neural networks structures as the classifier. The utilized standard neural network types include Feed-forward Neural Network (NN) with back propagation algorithm and a Radial Basis Functions Neural Networks.

Nidhi Srivastava [9] describes about Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ), etc. In this paper she uses Neural Network (NN) along with Mel Frequency Cepstrum Coefficients (MFCC) for speech recognition. MFCC has been used for the feature extraction of speech. This gives the feature of the waveform. For pattern matching Feed Forward Neural Network with Back propagation algorithm has been applied. The paper analyzes the various training algorithms present for training the Neural Network and uses train scg for the experiment. The work has been done on MATLAB and experimental results show that system is able to recognize words at sufficiently high accuracy.

Mayur R Gamit et.al [10] presents the use of an Artificial NeuralNetwork (ANN) for isolated word recognition. The Pre-processing is done and voiced speech is detected based on energy and zero crossing rates (ZCR). The proposed approach used in speech recognition is Mel Frequency Cepstral Coefficients (MFCC) and combine features of both MFCC and Linear Predictive Coding (LPC).

Divyesh S.Mistry et.a l[11] have used MFCC and Neural Network for speech recognition. The whole paper demonstrates how to use the mel-frequency cepsral coefficients and the neural network in speech recognition technology. And also demonstrates approach, application for speech recognition.

In 2014 Safdar Tanweer et.al [12] make a system to recognize the English word corresponding to digit (0-9) spoken by 2 different speakers is captured in noise free environment. For feature extraction, speech Mel frequency cepstral coefficients (MFCC) has been used which gives a set of feature vectors from recorded speech samples. Neural network model is used to enhance the recognition performance. Feed forward neural network with back propagation algorithm model is used.

## IV. FEATURE EXTRACTION

As speech of every individual person differs from each other, feature extraction is responsible for extracting relevant information from the speech frames, as feature parameters or vectors. It is usually performed in three main stages. The first stage is called the speech analysis or the acoustic front-end. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors. There are different methods for feature extraction. Some feature extraction techniques are LPC, MFCC, and DTW.

### A. Linear Predictive Coding (LPC)

It is desirable to compress signal for efficient transmission and storage. Digital signal is compressed before transmission for efficient utilization of channels on wireless media. For medium or low bit rate coder LPC is widely used. LPC (Linear Predictive Coding) is another method for feature extraction which is useful for encoding

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 5, May 2016*

good quality speech at a low bit rate. According to Robert M. Gray of Stanford University, the first ideas leading to LPC started in 1966. LPC is based on the assumption that the shape of the vocal tract governs the nature of the sound being produced [5]. LPC methods are the most widely used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage. LPC is based on the idea that each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding. Due to a very small error can distort the whole spectrum; LPC has to be tolerant of transmission errors [4]. The block diagram of LPC shown in fig: 2

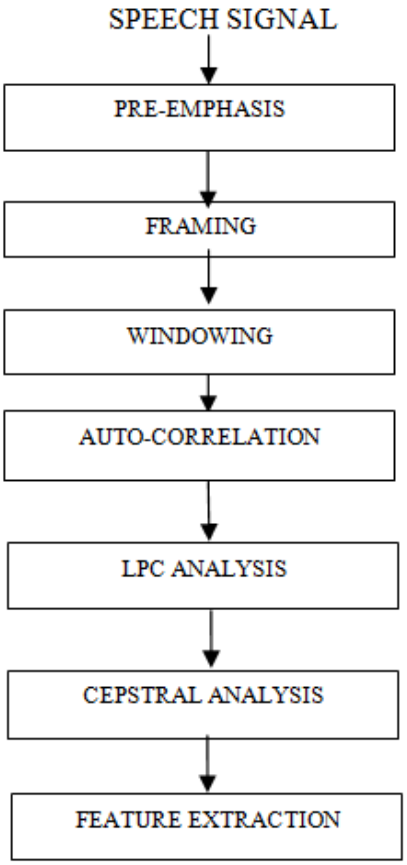


Fig. 2. Block Diagram of LPC

1. Firstly, Pre-emphasis is applied to the speech signal to make it less susceptible to finite precision effects in the processing of speech. The most widely used pre-emphasis is the fixed first-order system. The calculation of pre-emphasis is shown as follows.

$$H(z) = 1 - az^{-1} \quad ; 0.9 \leq a \leq 1.0$$

The most common value for 'a' is 0.95. A Pre-Emphasis can be expressed as follows

$$\hat{S}(n) = s(n) - 0.95s(n-1)$$

2. According to Rabiner (1993), the speech signal is assumed to be stationary when it is examined over a short period of time is called frame. As Speech signal is a kind of unstable signal, framing is applied to divide the speech signal in to frames of N samples, with adjacent being separated by M samples.

3. Next, windowing is applied to each frame in order to minimize the signal discontinuities or the signal is narrowed to zero at the starting and ending of each frame.

4. Auto- correlation method is applied to find the co-relation between the signal and a delayed version of itself.

5. The next process LPC analysis is applied to convert each frame of autocorrelation coefficients into the LPC parameters. The LPC parameters can be the LPC coefficients.

**B. Mel Frequency Cepstral Coefficient (MFCC)**

In speech recognition, MFCC have widely been used and have managed to handle the dynamic features as they extract both linear and non-linear properties of the signal. MFCC was first mentioned by Bridle and Brown in 1974 and further developed by Mermelstein in 1976 [8]. MFCC can be a useful tool of feature extraction in vibration signals as vibrations contain both linear and non-linear features. The technique MFCC contain both time and frequency information of the signal which makes more useful for feature extraction. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech.

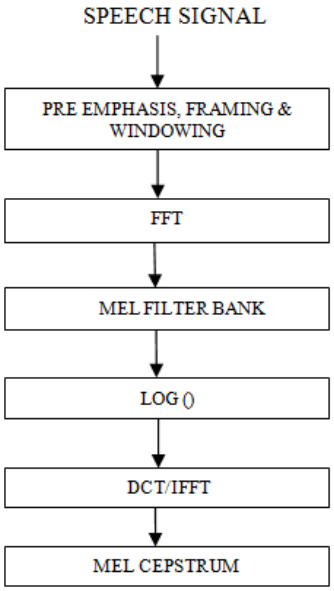


Fig. 3. MFCC Block Diagram

**1.** Firstly, a signal preprocessing is applied on a speech signal which consists of pre-emphasis filter to equalize the accurate size.

**A. Pre-Emphasis Filter**: In speech signal the higher frequency components get weakened, when transmitted via air. To overcome this issue, pre-emphasis filter has been used which boosts the energy in the higher frequency components.

Pre-emphasis filter has been implemented as given below…

$y(n) = x(n) - 0.95 * x(n-1)$

where * represents multiplication, x(n) represents the original signal and y(n), the pre-emphasized signal.

**B. Division Into Frames**: Though speech signal is not a stationary signal, it can be safely assumed to be stationary over short time intervals. Typically,a frame size of 10-25 ms, with an overlap of 50% between two adjacent frames is used. Framing divides the speech signal to get piecewise stationary.

**C. Windowing**:After framing the frames are passed through a hamming filter to eliminate the discontinuity within the frames and also at both the ends of each frame.Hamming window can be computed using the following equation:

$W(n) = 0.54 - 0.46 * \cos(2\pi n/N)\ ,\ 0 \leq n \leq N$

This returns an (N+1) point symmetric hamming window.

$S(n) = y(n) * w(n)$

Where y(n) represents the pre-emphasized signal and w(n) is the hamming window.

**D. Fast Fourier Transformation (FFT):** Fast Fourier Transformation (FFT) is calculated after completing the windowing for each frame to extract frequency components of a signal in the time domain. FFT is used to speed up the processing.

**2. Mel Filter Bank**: The Mel filter bank is a set of overlapping band pass filters adjusted according to the mel frequency scale. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies . The relation between frequency of speech and Mel scale can be established as

Frequency (Mel Scaled) = [2595log (1+f (Hz)/700)]
Here, f denotes the frequency of the speech signal.

**3. DCT**: The last step in MFCC is to calculate Discrete Cosine Transformation (DCT) of the outputs from the filter bank. DCT ranges coefficients according to significance, whereby the 0th coefficient is excluded since it is unreliable. It minimizes the distortion in frequency domain [12].The result of DCT is Mel Frequency Cepstral Coefficient

## IV. CLASSIFICATION TECHNIQUES

### A. Artificial' Neural Network (ANN)

The artificial' neural network (ANN), is provided by the inventor of one of the first neuron computers, Dr. Robert Hecht-Nielsen. He defines a neural network as: "a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs[11].

In this system we train a neural network to carry out the function by adjusting the values of the connections between elements .Feature vector data set i.e. MFCC co-efficient is divided into three data sets which are training data set, validation data set and testing data set. The weights are adjusted as per requirement of the system and hidden layer is linked to the output layer where the result is displayed.

### B. Hidden Markov Model (HMM)

It is one of the most successful and most used pattern recognition method for speech recognition which is a mathematical model derived from a Markov Model.
A HMM is characterized by the following
i) N - number of states in the model
ii) M - number of distinct observations per state
iii) A - state transition probability distribution
iv) B - observation symbol probability distribution
v) $\mu$ - Initial state distribution.

Therefore, to specify a HMM model one needs to specify two model parameters “N” and “M”, the observation symbols, probability majors “A”, “B” and initial state distribution [6]. As every sound entity is treated separately, HMM seems to perform quite well in noisy environments. When a sound entity is lost in the noise, the model might be able to guess that entity based on the probability of going from one sound entity to another.

### C. Dynamic Time Warping (DTW)

DTW is one of the techniques which is used to pattern matching for isolated speech recognition. It compares words with reference words. As every reference word has a set of spectra; but there is no distinction between separate sounds in the word. Since a word can be pronounced at different speeds, time normalization will be necessary [8]. DTW is a programming technique through which the time dimension of the unknown word is changed (stretched and shrinked) until there is a similarity with a reference word.

## VI. CONCLUSION

This paper describes about speech recognition process. It also explained about different feature extraction techniques like LPC, MFCC and also describes about different classification technique like ANN, HMM and DTW.

## ACKNOWLEDGEMENT

We sincerely thank all the entire researchers for their wide contribution to understand in the area of speech recognition technology. We are also really thankful to all other people who helped us in recording the voices.

## REFERENCES

[1]. Banikanta Kakati, “Assamese, its Formation and Development”, 5th ed.,Guwahati, India ,LBS publication, 2007.

[2]. Lawrence R. Rabiner1 and Ronald W. Schafer," Foundations and Trends in Signal Processing", Vol. 1, Nos. 1-2(2007) 1-194. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[3]. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.
[4]. Bhargab Medhi1, P.H.Talukdar,"Assamese Speaker Recognition using Artificial Neural Network",(IJARCCE) International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021 ,Vol. 4, Issue 3, March 2015.
[5]. Bhargab Medhi1, P.H.Talukdar," Different acoustic feature parameters ZCR, STE, LPC and MFCC analysis of Assamese vowel phonemes", International Conference on Frontiers in Mathematics 2015.
[6]. Utpal Bhattacharjee," A Comparative Study of LPCC And MFCC Features For The Recognition Of Assamese Phonemes", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 1, January- 2013.
[7]. Deshmukh Aniket Anand, Pramod Kumar Meher," Combined LPC and MFCC features based technique for Isolated Speech Recognition", Indian Institute of Technology Hyderabad ,Hyderabad, India.
[8]. Akalpita Das," Isolated BODO Spoken Word Identification using Mel-frequency Cepstral Coefficients and K-means clustering",International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 11, November 2013.
[9]. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov," Neural Networks used for Speech Recognition", Journal Of Automatic Control, University Of Belgrade, Senior Member, IEEE Vol. 20:1-7, 2010.
[10]. Nidhi Srivastava," Speech Recognition Using Artificial Neural Network", International Journal of Engineering Science and Innovative Technology (IJESIT), Volume 3, Issue 3, May 2014.
[11]. Mayur R Gamit1, Kinnal Dhameliya," Isolated Words Recognition Using MFCC, LPC and Neural Network", International Journal of Research in Engineering and Technology (IJRET), eISSN: 2319-1163 | pISSN: 2321-7308.
[12]. Divyesh S.Mistry, Prof.A.V.Kulkarni, Dr. D. Y. Patil ,"Overview: Speech Recognition Technology, Melfrequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN)", International Journal of Engineering Research & Technology (IJERT) ,Vol. 2 Issue 10, October – 2013.
[13]. Safdar Tanweer, Abdul Mobin, Afshar Alam," Analysis of Combined Use of NN and MFCC for Speech Recognition, International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:8, No:9, 2014.
[14]. B Medhi, P.H. Talukdar, "Assamese vowel phoneme recognition using Zero Crossing Rate and Short Time Energy", IJARCSSE Vol.4, 2014.
[15]. Bhargab Medhi1, P.H.Talukdar," LPC and MFCC Analysis of Assamese Vowel Phoneme", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 5, Issue 1, January 2015.
[16]. Mousmita Sarma, Krishna Dutta and Kandarpa Kumar Sarma,"Assamese Numeral Corpus for Speech Recognition using Cooperative ANN Architecture", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:3, No:4, 2009.


DOI 10.17148/IJARCCE.2016.5533