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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Assamese Speaker Recognition Using Artificial Neural Network

Bhargab Medhi[1], Prof. P.H.Talukdar[2]

Research Scholar, Department of Instrumentation & USIC, Gauhati University, Guwahati, India [1]

Professor, Department of Instrumentation & USIC, Gauhati University, Guwahati, India [2]

Abstract: This paper proposes an approach to recognise Assamese speaking person using Artificial Neural Network Model. Speaker recognition is the process of Identification of the person who is speaking depending on the characteristics of his/her voices. The features Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficient (MFCC) are used to create the feature vector of the Assamese speech samples (words). Our 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.

Keywords: Speaker Recognition, LPC, MFCC, Neural Network

I. INTRODUCTION

Speech is the most effective way to communicate with each other between human beings. Speech conveys the linguistic information, the speaker's vocal tract characteristics and the speaker's emotion. Recent development has made possible to use the speech in different security systems [4,5,6]. Automatic speaker recognition technology is flattering increasingly widespread in many applications such as biometric personal identification, physical access control, computer data access control, and so on.

Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is a method to determine which one of a group of known voices best matches the input voice sample. On the other hand, speaker verification is a method to determine from a voice sample if a person is whom he/she claims to be. For both the methods, the utterances have two types i.e. text dependent and text independent. In text dependent, the utterances have a finite set of sentences where as in text independent the utterances are totally unconstrained [10].

Speaker recognition system consists of two main modules: feature extraction and feature matching. In feature extraction part, different features are extracted from the voice sample to form the feature vector [7,8]. On the other hand, feature matching involves the actual procedure to identify the unknown speaker by comparing extracted feature from his/her speech input with the one from a set of known speakers.

The Assamese (IPA: ɔxɔmija) is a major language in the north-eastern part of India whose origin root is Indo-European family of languages [1]. There are thirty two essential phonemes (speech unit) in Assamese language out of which eight are vowel phonemes and twenty four are consonant phonemes. Assamese scripts, derived from Devanagari scripts consists of thirty nine consonant and eleven vowel symbols [8].

In this proposed method, a neural network (Multi Layer Perceptron: MLP) is designed through Matlab R2010b. The features 12 LPC and 12 MFCC are extracted from each speech sample. 80% data of speech sample is used to train the neural network.20% data of speech sample is used for testing process to recognise the speakers. Five isolated Assamese words are taken as speech samples uttering each sample twenty times resulting one thousand speech samples in the database.

II. FEATURE EXTRACTION

Feature extraction is the most important part of speaker recognition as it distinguishes one speaker from the other. The feature extraction gives one feature vector from one speech sample [2, 3]. Extracted feature must meet some criteria such as:

- Easy to measure extracted speech feature.
- It should not be receptive to imitation.
- It should confirm less variation from one speaking environment to another.
- It should be balanced over time period.
- It should produce normally and naturally in speech.

In this method, two features LPC and MFCC are used to make feature vector of the speech sample.

A. *Linear Predictive Coding(LPC)*

LPC is one of the most powerful signal analysis techniques for linear prediction.LPC is a way of encoding the information in a speech signal into a smaller space for transmission over a restricted channel. LPC encodes a signal by finding a set of weights on earlier values that can predict the next signal value. The output of LPC analysis is a set of co-efficient a[1,….,k] and an error signal e(n), the error signal will be as small as possible and represents the difference between the predicted signal and the original [6,7]. The mathematical model of speech Production is often called LPC model. The block diagram of LPC computation can be shown by the figure fig.1.

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

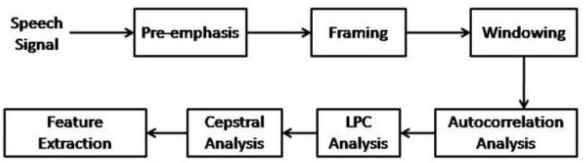

Fig. 1 Block diagram of LPC

a) Pre-emphasis

The speech signal which is digitized has a high dynamic range and it contains additive noise. So, pre-emphasis is applied to spectrally level the signal so as 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 computation of pre-emphasis is shown as follows.

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

The most common value for a is 0.95 (Deller et at; 1993). A Pre-Emphasis can be expressed as:

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

b) Frame blocking

Generally the speech signal is dynamic or time-variant in nature. According to Rabiner (1993), the speech signal is assumed to be stationary when it is examined over a short period of time called frame. In order to analyse the speech signal, it is divided into frames of N samples, with adjacent frames being separated by M samples. If M ≤ N, then LPC spectral estimates from frame to frame will be quite smooth. On the other hand if M > N there will be no overlap between adjacent frames.

c) Windowing

Each frame is windowed in order to minimize the signal discontinuities or the signal is narrowed to zero at the starting and ending of each frame [2,3,4]. If window is defined as w(n), then the windowed signal is

$$\tilde{x}(n) = x(n)w(n), 0 \le n \le N-1$$

A typical Hamming window is used which has the form

$$w(n) = 0.54 - 0.46\cos(\frac{2n}{N-1}), 0 \le n \le N-1$$

The value of the analysis frame length N must be long enough so that tapering effects of the window do not seriously affect the result.

d) Autocorrelation analysis

The technique relies on finding the co-relation between the signal and a delayed version of itself [5.6]. Each frame of windowed signal is next auto correlated to give

$$R(n) = \sum_{x=0}^{N-1-m} \tilde{x}(n)\tilde{x}(n+m), m = 0,1,2,....., p$$

Where, the highest autocorrelation value, P is the order of the LPC analysis. The selection of p depends primarily on the sampling rate.

e) LPC analysis

The next processing step is the LPC analysis which converts each frame of autocorrelation coefficients R into the LPC parameters. The LPC parameters can be the LPC coefficients [6,7]. This method of converting autocorrelation coefficients to LPC coefficients is called as Durbin's method. Levinson-Durbin recursive algorithm is used for LPC analysis.

$$E_0 = R(0)$$

$$k_i = [R(i) - \sum_{j-1}^{i-1} a_j^{i-1} R(i-j)] / E_{i-1}, 1 \le i \le p$$

$$a_i^i = k_i$$

$$a_i^i = a_i^{i-1} - k_i a_{i-j,}^{i-1} 1 \le j \le i-1$$

$$E_i = (1-k_i^2)E_{i-1}$$

The above set of equations are solved recursively for i = 1,2, p, where p is the order of the LPC analysis. The k_i are the reflection coefficients. The a_j are the LPC coefficients. The final solution for the LPC coefficients is computed as follow

$$a_j = a_j^{(p)}, 1 \le j \le p$$

In this experiment the parameters for LPC are Sampling Frequency=16000Hz, Frame Size=256samples, Frame Overlap=128 samples, Window Type= Hamming (size 256), LPC size=12.

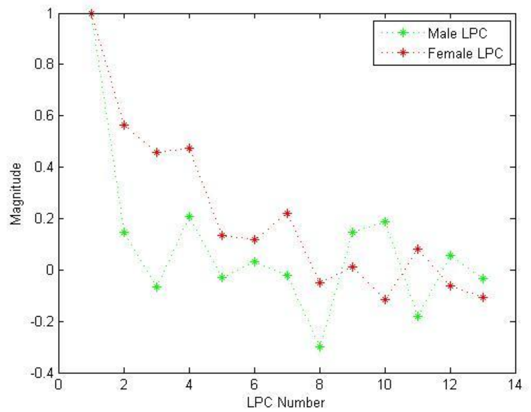

Fig. 2 the 5^{nt} frame 12 LPC comparison of both male and female speaker of Assamese word অসমীয়া(IPA:/ɔxɔmija/)

B. Mel Frequency Cepstral Coefficient(MFCC)

The Mel-frequency Cepstral Coefficients (MFCCs), introduced by Davis and Mermelstein, is possibly the most popular and common feature for ASR systems [10]. This may be certified because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better. The following figure fig.3 shows the block diagram of MFCC computation.

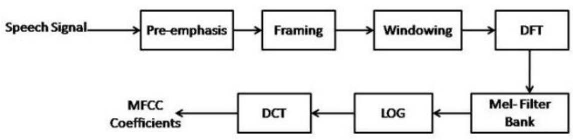

Fig. 3 Block diagram of MFCC

To calculate the MFCCs of a speech signal sample, the signal is first passed to pre-emphasis filter. Then the speech is processed on a frame-by-frame basis in what is called framing. Normally, a frame size of 20ms to 30ms is used and Windowing of these frames are done to compensate discontinuities within the speech signal as a result of segmentation and overlapped frames [8,9]. Windowing means multiplying the window function w(n) with the framed speech signal s(n) to obtain the windowed speech signal $s_{0w}(n)$.

$s_{0w}(n)=s(n)w(n)$

The discrete Fourier transform (DFT) of the windowed speech signal is then calculated by the following equations:

$$\hat{S}_{0w}(k)=\sum_{n=0}^{N-1} s_{0w}(n)e^{\frac{-j2\pi kn}{N}}$$

The mel-filterbank is a triangular bandpass filter which is equally spaced around the Mel-Scale. A Mel is a unit of perceived pitch or frequency of a tone [5,7,9]. The mapping between real frequency (hz) and Mel-frequency is given by the following equations as[5,6,7]:

$$f_{mel}=2595\log(1+\frac{f}{700})$$

The power spectrum from the DFT step is then binned by correlating it with each triangular filter in order to reflect the frequency resolution of the human ear. Binning means multiplying the power spectrum coefficients with the triangular filter gain or coefficients and summing the resultant values to obtain the Mel-Cepstral coefficients as in equation:

$$G(k)=\sum_{n=0}^{\frac{N}{2}} \eta_{kn}.[\hat{s}_{0w}(k)]^2$$

Where η_{kn} is the triangular filter coefficients, k=0,1,2,…..k-1, n=0,1,2,…..N/2 and G(k) is the Mel-Cepstral coefficients. After that, the log of the Mel-Cepstral coefficients G(k), is taken. This step is to plane unwanted ripples in the spectrum and done the following equation.

$$m_k=\log G(k)$$

Finally, DCT is applied to the log mel-cepstrum m_k as in equation to obtain the Mel-frequency Cepstral Coefficients (MFCC) c_i of the i^{th} frame:

$$c_i=\sqrt{\frac{2}{N}}\sum_{k=1}^{N} m_k \cos(\frac{\pi i}{N}(k-0.5))$$

In this computation, the parameters for MFCC are Sampling Frequency=16000Hz, Frame Size=256samples, Frame Overlap=128 samples, Window Type= Hamming (size 256), cepstral coefficient=12, no. of filter bank=24.

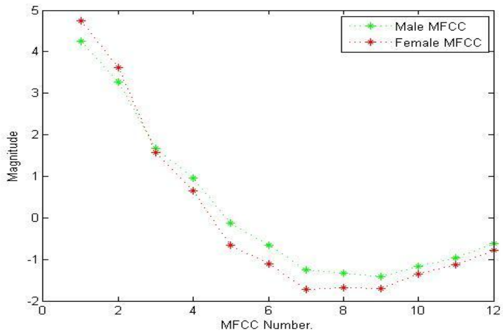

Fig. 4 the 2nd frame 12 MFCC comparison of both male and female speaker of Assamese word মা (/Maa/)

III.NEURAL NETWORK DESIGN

In this paper we have created an Artificial Neural Network (ANN) for speaker recognition. A Neural Network is composed of simple elements which are operated in parallel. These elements are inspired by biological nervous system called neurons. Each neuron computes a nonlinear weighted sum of its inputs, and sends the result over its outgoing connections to other neurons [2,3,4].

We train a neural network to carry out a particular function by adjusting the values of the connections (weights) between elements. Learning is a process of training the network. During the training phase of the network the weights are adjusted. After using all the training data one time , it is called learn cycle or epoch.

The most useful neural network used in function approximation is Multi Layer Perceptron (MLP). A MLP consists of an input layer, one or more hidden layers, and an output layer. 12 LPC and 12 MFCC are calculated from each speech sample frame wise to construct the feature vector. Because of the strong randomness of a speech signal i.e. different sizes feature vectors, it is necessary to merge some of the elements of the feature vector. **K-means** is one of the simplest and popular methods to cluster the vectors to get compressed feature vectors.

Each feature vector of LPC or MFCC is clustered into different sizes 5, 10, 15, 20, and 25 to overcome the problem of variable length feature vector. Both LPC and MFCC feature combined parallely to construct the feature vector for the neural network. 80% speech sample is used for training phase and the remaining 20% is used for testing.

The proposed method is used for both text dependent speaker recognition and text independent speaker recognition. In case of text dependent speaker recognition the testing sample is considered from known speakers. But in the text independent speaker recognition the testing is performed with the samples which are not there in training.

An MLP is shown in the figure fig.5.

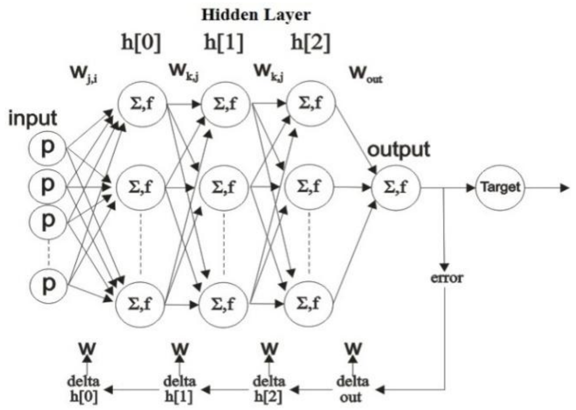

Fig. 5 A multilayer Perceptron network with three hidden layers

The matlab command *newff* generates a MLP neural network which is called net

net=newff(PR,[S1 S2...SN*l*],{TF1 TF2...TFN*l*},BTF)
where,
PR=Min Max values
Si=Number of neurons in the i^{th} layer, i=1,2..,*l*.
TFi=Transfer Function of the i^{th} layer.
BTF=Network training function.

Here only one hidden layer is considered and number of neurons in the hidden layer is tuned depending on the correct classification.

The recognition rate for both text dependent speaker recognition and text independent speaker recognition are mentioned in table 1 and table 2 respectively. In text dependent speaker recognition, an Assamese word 'ব'হাগ'(bohag)is used as speech sample to recognize who is speaking among the speakers. Each of ten speakers uttered the word 20 times. In text independent speaker recognition, we use a neural network of five outputs to indicate five different speakers no matter which registered speeches are given.

TABLE 1
TEXT DEPENDENT SPEAKER RECOGNITION

Assamese Speaker	No. of Samples for Testing					No. of Properly Recognized Speaker					Recognition Rate (%)					
	Cluster 5	cluster 10	Cluster 15	Cluster 20	Cluster 25	Cluster 5	cluster 10	cluster 15	cluster 20	cluster 25	Cluster 5	cluster 10	cluster 15	cluster 20	cluster 25	Average Recognition Rate
speaker1	20	20	20	20	20	19	19	20	20	20	95%	95%	100%	100%	100%	98%
speaker2	20	20	20	20	20	19	19	20	20	20	95%	95%	100%	100%	100%	98%
speaker3	20	20	20	20	20	18	19	19	19	20	90%	95%	95%	95%	100%	95%
speaker4	20	20	20	20	20	19	19	19	20	20	95%	95%	95%	100%	100%	97%
speaker5	20	20	20	20	20	19	19	19	20	20	95%	95%	95%	100%	100%	97%
Total	100	100	100	100	100	94	95	97	99	100	94%	95%	97%	99%	100%	97%

TABLE 2
TEXT INDEPENDENT SPEAKER RECOGNITION

Assamese Speaker	No. of Samples for Testing					No. of Properly Recognized Speaker					Recognition Rate (%)					
	Cluster 5	cluster 10	Cluster 15	Cluster 20	Cluster 25	Cluster 5	cluster 10	cluster 15	cluster 20	cluster 25	Cluster 5	cluster 10	cluster 15	cluster 20	cluster 25	Average Recognition Rate
speaker1	20	20	20	20	20	20	19	20	20	20	100%	95%	100%	100%	100%	99%
speaker2	20	20	20	20	20	19	19	19	20	20	95%	95%	95%	100%	100%	97%
speaker3	20	20	20	20	20	18	20	19	19	20	90%	100%	95%	95%	100%	96%
speaker4	20	20	20	20	20	19	19	19	20	19	95%	95%	95%	100%	95%	96%
speaker5	20	20	20	20	20	19	19	18	20	20	95%	95%	90%	100%	100%	96%
Total	100	100	100	100	100	95	96	95	99	99	95%	96%	95%	99%	99%	96.8%

IV. CONCLUSION

In this paper, an effective Speaker recognition technique is used which gives a moderately high accuracy in recognition system. Though MFCC is widely used alone in speaker recognition technique, but MFCC failed to give a satisfactory result in case of text independent speaker recognition. So we proposed a new method where both LPC and MFCC are used parallely. Even though the good result in our method, there are still many problems that need to further investigated because all the signals of my database are recorded in very good condition. The future scope of my work will be to perform speaker recognition in noisy environment. We hope that this paper brings out understand and inspiration amongst the research group of ASR.

ACKNOWLEDGMENT

We are very much grateful to the different speakers for whom the record is possible to create the whole database. It is only because of the data sets that make it possible to accomplish the entire task.

REFERENCES

[1] Banikanta Kakati, "Assamese, its Formation and Development",5th ed.,Guwahati, India ,LBS publication, 2007.
[2] Gold and N. Morgan, "speech and Audio Processing: Processing and Perception of speech and music", New York,2000.
[3] G.K. Vallabha , B. Tuller, "Systematic errors in the formant analysis of steady-state vowels, Speech Communication, Vol. 38,2002,pp. 141.
[4] L.Flanagan, "Speech Analysis, Synthesis and Perception",2nd edition, New York, Springer-Verlag, 1972.
[5] L.R. Rabiner, R.Schafer, "Digital Processing of Speech Signals", Englewood Cliffs, NJ, Prentice Hall, 1979.
[6] F. Jelinek, Statistical Methods for Speech recognition", Cambridge, The MIT Press, 1998.
[7] T. B. Adam , Md. Salam, "Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Network", IJCA, vol 42,March 2012.
[8] T.K. Das , P.H. Talukdar, "Cepstral Analysis of Assamese vowel Phonemes",IJACST, Aug. vol 2 2013.
[9] B. Medhi, P.H. Talukdar, "LPC and MFCC analysis of Assamese vowel phonemes",IJARCSSE, Vol 5, 2015.
[10] Dr. Joseph Picone, Fundamental of speech recognition: A short coures", ISIP.
.

DOI 10.17148/IJARCCE.2015.4377