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

A Survey on Isolated Assamese words spoken by Male and Female speakers

Mr. Adarsh Pradhan[1], Antara Chowdhury[2]

Assistant Professor, Dept of Computer Science and Engineering, Girijananda College of Engineering and Technology, Guwahati, India[1]

M.Tech 4th Semester, Dept of Computer Science and Engineering, Girijananda College of Engineering and Technology, Guwahati, India [2]

Abstract: Speech is the most usual form of human communication and speech processing has been one of the most stimulating expanses of signal processing. Speech recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Speech technology and human computer interaction systems have witnessed a steady and significant improvement over the last two decades. Now – a - days, speech technologies are commercially available for an unlimited but exciting range of tasks. Using these technologies machines are able to respond correctly and dependably to human voices, and provide useful and valuable services. Recent research concentrates on developing systems that would be much more robust against unpredictability in environment, speaker and language. Hence today researchers mainly focus on Speech Recognition systems with a large vocabulary that support speaker independent operation with continuous speech in different languages.

Keywords: Speech Recognition; Feature Extraction; MFCC; LPC; ANN; VQ; HMM.

I. INTRODUCTION

The ability of machine or program to identify words and phrase from spoken language and convert them in to machine readable format is known as Speech Recognition. It is also known as Automatic Speech Recognition [3, 4]. Speech recognition intends to evolve techniques and system for speech input to machine.
We know that speech is the primary means of communication between humans, and hence motivates research efforts to allow speech to become a feasible human computer communication.

Signal is a measurable physical quantity that is produced by a system which is an existing physical entity. We can categorize signal into several classes based on their nature like - continuous or discrete, energy or power, periodic or aperiodic, deterministic or random, stationary or non-stationary etc. Speech is an example of non-stationary signal [7].

On the other hand synthetic signals like sine wave, triangular wave, square wave etc are stationary in nature. Hence we require different approaches and tools for processing the speech signals. Communication, i.e., message transmission is the main purpose of speech.

According to Shannon's information theory, "A message which is represented as a sequence of distinct symbols can be measured by its information content in bits, and the rate of transmission of information is measured in bits per second (bps)" [1]. In speech production, as well as in many electronic communication systems, the information to be transmitted is programmed in the form of a continuously varying waveform that can be transmitted, recorded, manipulated, and ultimately get decoded by a human listener. For speech, the fundamental analog form of the message is an auditory waveform called the speech signal. Speech signals can be converted to an electrical waveform using a microphone, which can further be manipulated both by analog and digital signal processing, and then converted back to acoustic form by a loudspeaker, a telephone handset or headphone, as required. Speech can be represented phonetically by a finite set of symbols called the phonemes of the language, the number of which depends upon the language and the enhancement of the analysis [1]. For most languages the number of phonemes is between 32 and 64.

A. Types Of Speech Uttered:
Different classes of speech recognition systems can be made based on the types of utterances of speech which are as follows [5]:

1. Isolated speech: A single utterance at a time is required requiring each utterance has little or no noise on both sides of the sample window. They often have "Listen/Not-Listen states", where it is required to have pause between utterances.
2. Connected word: They require minimum pause between utterances to make speech flow smoothly and are almost similar to isolated words.
3. Continuous speech: This is normal human speech, without silent pauses between words and makes machine understanding much more difficult.
4. Spontaneous speech: They can be thought of as speech that sound natural and have not being tried out before.

DOI 10.17148/IJARCCE.2016.5564

B. Types of Speaker Model:
All speakers have singular voices because of their unique physical body and personality [5]. Speech recognition system is broadly classified into two categories based on speaker models:

1) Speaker dependent models: They are designed for a specific speaker and are generally more accurate for the particular speaker, but much less accurate for other speakers. They are easier to develop, cheaper and more accurate, but not as flexible as speaker independent systems.
2) Speaker independent models: They are designed for variety of speakers. They recognize 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.

C. Types Of Vocabulary:
The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. The types of vocabularies can be classified as follows:
1. Small vocabulary – Consists of tens of words.
2. Medium vocabulary – Consists of hundreds of words.
3. Large vocabulary – Consists of thousands of words.
4. Very - large vocabulary – Consists of tens of thousands of words.
5. Out-of-Vocabulary - Maps a word from the vocabulary to the unknown word.

Apart from the above characteristics, the environment variability, channel variability, style of speaking, gender, age, speed of speech also makes Speech Recognition system more complex.

D. Non - Stationary Nature Of Speech:
A signal is said to be stationary if its frequency or spectral contents do not change with respect to time. This happens because while generating a sine wave using either a function generator or software, the frequency value will be selected and kept constant forever. Thus the frequency content of the sine wave will not change with time. If the frequency is changed, then it becomes a new sine wave. Further, we should not confuse the stationary nature with the time varying amplitude in the time domain as in the case of sine wave. It is linked to the behaviour of the frequency contents of the signal with respect to time and nothing else. The speech signal is a very complicated non-stationary signal [7]. First, there may many components in a given interval of time. Second, the interval itself will be very short, as short as about 10-30 msec. This means that the frequency contents of the speech signal will have many frequency contents which will change continuously with time.

II. LITERATURE REVIEW

In their paper “Assamese Speaker Recognition Using Artificial Neural Network”, Bhargab Medhi and Prof. P.H. Talukdar examined an effective Speaker recognition technique which gives a moderately high accuracy in recognition system. They proposed a new method where both LPC and MFCC are used parallely. They found good result in their method, but there are still many problems that need to further investigated because all the signals of their database were recorded in very good condition. The future scope of their work will be to perform speaker recognition in noisy environment.
Gurpreet Kaur and Harjeet Kaur in their paper “Multi Lingual Speaker Identification on Foreign Languages Using Artificial Neural Network with Clustering” carried out their experiment on a series of cluster based multi lingual speaker identification using neural network. The result shows that k-mean clustering can be used for multi lingual system. This research focuses on text dependent speaker identification. The minimum performance of the system is 92.08% while the best performance is being reached up to 100%. Overall performance of the system is reached on 96%. This system is unable to identify very small words. Future work includes small words also which can increase the performance of this system.
M.K. Deka, C.K. Nath, S.K. Sarma and P.H. Talukdar in their paper “An Approach to Noise Robust Speech Recognition using LPC - Cepstral Coefficient and MLP based Artificial Neural Network with respect to Assamese and Bodo Language” concluded that neural networks are inherently fault tolerant. This fault tolerant capability of the neural network can be enhanced by proper configuration of the network which can then be used as a robust speech recognizer.
It is found that in multi-layer perceptron, by increasing the number of hidden layers and nodes, better performance can be achieved in noisy environment. Since the increase in the number of layers also increases the time requirement for training as well as complexity of the training algorithm, so the proposed scheme sets a limit of maximum number of hidden layers (=3), after which both the training and recognition process become slow.”

III. SPEECH FEATURE EXTRACTION TECHNIQUES

Feature Extraction is the most important part of speech recognition as it plays an important role to separate one speech from the other. Because every speech has different individual characteristics embedded in utterances these characteristics can be extracted from a wide range of feature extraction techniques already proposed and successfully exploited for speech recognition task [6]. But extracted feature should be with met with some criteria while dealing with the speech signal such as:
a. It should be easy to measure extracted speech feature.
b. It should not be receptive to mimicry.
c. It should be balanced over time.
d. It should occur normally and naturally in speech.
e. It should show little fluctuation from one speaking environment to another [6].

Different techniques for feature extraction are LPC, MFCC, AMFCC, PLP, PCA, etc.

DOI 10.17148/IJARCCE.2016.5564

A. Linear Predictive Coding (LPC):

Linear predictive analysis is one of the most powerful and widely used speech analysis techniques for encoding good quality speech at a low bit – rate and provides extremely accurate estimates of speech parameters [3]. It is a tool most widely used in audio signal processing and speech processing for representing the spectral envelope of digital signal of speech in compressed form, using the information of a linear predictive model. This method is important as accurate estimates of the speech parameters can be provided and also for its computational speed. Linear prediction is a mathematical operation where future values of a discrete time signal are estimated as a linear function of previous samples. Linguists often use LPC as a formant extraction tool. LPC analysis is usually most appropriate for modelling vowels, periodic vowels, except nasalized vowels [1].

It is one of the most powerful signal analysis techniques. It has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and is also an efficient computational model of speech. The basic idea is that a speech sample can be approximated as a linear combination of past speech samples. Minimizing the sum of squared differences over a finite interval between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are transformed to a more robust set of parameters known as cepstral coefficients.

The basic idea behind the LPC model is that a given speech sample at a time n, s(n) can be approximated as a linear combination of the past p speech samples, such that

$$\mathbf{s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots + a_p s(n-p)}$$

where the co-efficients a_1, a_2, ..., a_p are assumed constant over the speech analysis frame [2]. The above equation is converted to an equality by including an excitation term, G u(n), giving:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G\,u(n)$$

where u(n) is normalized excitation and G is gain of excitation. By expressing the above equation in the z – domain we get the relation:

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + G\,U(z)$$

leading to the transfer function:

$$H(z) = \frac{S(z)}{G\,U(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)}$$

The interpretation of the above equation is given in Fig. 1 which shows the normalized excitation source, u(n), being scaled by the gain, G, and acting as input to the all- pole system to produce the speech signal, s(n).

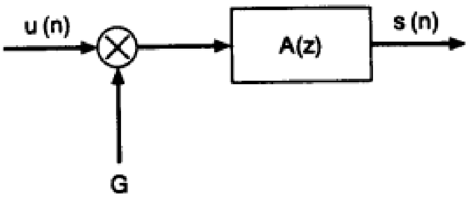

Fig. 1: Linear Prediction Model of speech [2]

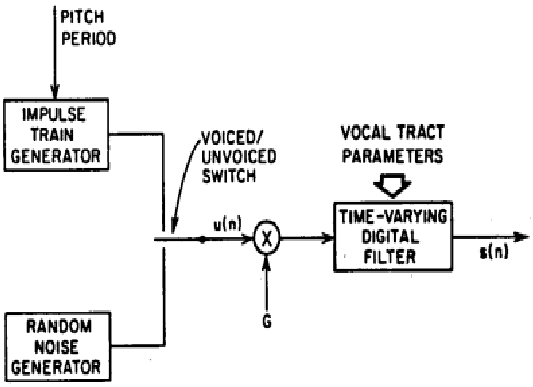

Fig. 2: Speech Synthesis Model based on LPC Model[2]

$$H(z) = \frac{1}{A(z)}$$

The normalized excitation source is chosen by the switch whose position is controlled by the voiced/ unvoiced character of speech. The appropriate gain of the source, G, is estimated from

the speech signal and the scaled source is used as input to digital filter (H(z)) which is controlled by the vocal parameters characterised by the produced speech. Thus the parameters of this model are: voiced – unvoiced classification, pitch period for voiced sound, gain parameter, and coefficients of digital filter $\{a_k\}$ which vary slowly with time.

In linear predictive analysis, the excitation is defined implicitly by the vocal tract system model. The major advantage of this model is that the gain parameter, G, and the filter coefficients {ak} can be estimated in a very straightforward and computationally efficient manner by the method of linear predictive analysis [1].

The basic problem of linear prediction analysis is to determine the set of predictor coefficients $\{\alpha_k\}$ directly from the speech signal in order to obtain a useful estimate of the time-varying vocal tract system. The basic approach is to find a set of predictor coefficients that will minimize the mean-squared prediction error over a short segment of the speech waveform.

The resulting parameters are then assumed to be the parameters of the system function H(z) in the model for production of the given segment of the speech waveform. This process is repeated periodically at a rate appropriate to track the phonetic variation of speech.

B. Mel Frequency Cepstral Coefficients (MFCC):

MFCC is the most widespread and leading method used to extract spectral features. They are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is

DOI 10.17148/IJARCCE.2016.5564

based on the human ear scale. MFCCs being considered as frequency domain features are much more accurate than time domain features.

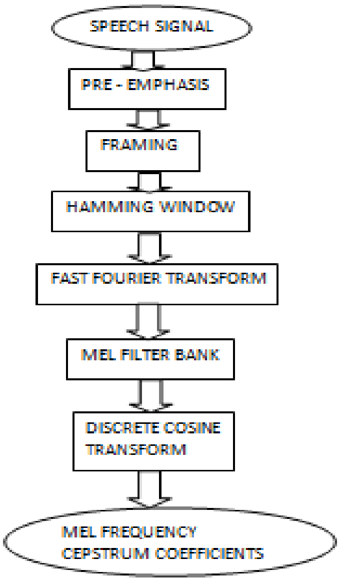

Fig. 3: Flow of MFCC

It is a representation of the real cepstral of a windowed short time signal derived from the Fast Fourier Transform (FFT) of that signal [7]. The variation from the real cepstral is that a nonlinear frequency scale is used, which approximates the behaviour of the auditory system In addition, these coefficients are robust and reliable to variations according to speakers and recording conditions. MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones used by humans for hearing speech, while at the same time, deemphasizes all other information. The speech signal is first divided into time frames consisting of an arbitrary number of samples. In most systems overlapping of the frames is used to smooth transition from frame to frame. Each time frame is then windowed with Hamming window to eliminate discontinuities at the edges.

IV. CLASSIFICATION TECHNIQUES

A. Artificial Neural Network (ANN):
An artificial neural network (ANN) is often just called a neural network (NN. It is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing. In most cases an ANN is an adaptive system that changes its structure depending on external or internal information flowing through the network. The MLP (Multi Layer Perceptron) is a type of neural network that has become very popular over the past several years. They are usually trained with an iterative gradient algorithm known as back propagation.

B. Vector Quantization (VQ):
Vector Quantization (VQ) is often applied to Automatic Speech Recognition. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the VQ's utility lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods.

C. Hidden Markov Model (HMM):
HMM is one of the key technologies developed in the 1980s. It is a doubly stochastic process that is not observable and hence the term hidden, but can be observed through another stochastic process that produces a sequence of observations. The HMM pattern-matching strategy was in due course adopted by each of the major companies pursuing the commercialization of speech recognition technology (SRT). The U.S. Department of Defence sponsored many practical research projects during the 70s that involved several contractors, including IBM, Dragon, AT&T, Philips and others.

V. CONCLUSION

In this paper we have studied various feature extraction techniques for speech recognition like LPC, MFCC and various classifier techniques such as ANN, VQ, HMM.

ACKNOWLEDGMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

Our thanks and appreciations go to all those people in recording the voices and all the people who have willingly helped us out with their abilities.

REFERENCES

[1] Lawrence R. Rabiner, Ronald W. Schafer, "Introduction to Digital Speech Processing", Foundations and Trends in Signal Processing, Vol. 1, Nos. 1–2 (2007), DOI: 10.1561/2000000001

[2] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993

[3] Bhargab Medhi, Prof. P.H.Talukdar, "Assamese Speaker Recognition Using Artificial Neural Network", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015

[4] Mayur R Gamit, Kinnal Dhameliya, "ISOLATED WORDS RECOGNITION USING MFCC, LPC AND NEURAL NETWORK", IJRET, eISSN: 2319-1163 | pISSN: 2321-7308, Volume: 04 Issue: 06 | June-2015

[5] Shanthi Therese S., Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition", International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.2, Issue No.6,

[6] Nidhi Desai, Prof.Kinnal Dhameliya, Prof.Vijayendra Desai, "Feature Extraction and Classification Techniques for Speech Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 12, December 2013)

[7] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal For Advance Research In Engineering And Technology, Volume 1, Issue VI, July 2013

[8] Shreya Narang, Ms. Divya Gupta, "Speech Feature Extraction Techniques: A Review", International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IJCSMC, Vol. 4, Issue. 3, March 2015, pg.107 – 114.

DOI 10.17148/IJARCCE.2016.5564