Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1,Issue 12, December -2014

# SPEECH RECOGNITION: A REVIEW

Bhoomika Dave[1], Prof. D. S. Pipalia [2]

[1]Electronics and communication Department, R.K.University Email ID: bhoomikadave30@gmail.com
[2]Electronics and communication Department, R.K.University Email ID: dhaval.pipalia@rku.ac.in

***Abstract:*** *Speech interface to computer is the next big step that the technology needs to take for general users. Automatic speech recognition (ASR) will play an important role in taking technology to the people. There are numerous applications of speech recognition such as direct voice input in aircraft, data entry, speech-to-text processing, voice user interfaces such as voice dialing. ASR system can be divided into two different parts, namely feature extraction and feature recognition. This paper provides an overview for Speech recognition where Acoustic modeling techniques, Feature extraction techniques for Speech recognition are briefly discussed.*

***Keywords:*** *Automatic Speech Recognition, Hidden Markov Model, Artificial Neural Network, Feature Extraction*

## I. INTRODUCTION

The process of automatically recognizing spoken words of speaker based on information in speech signal is called Speech Recognition. In automatic speech recognition computer captures the words spoken by a human with a help of microphone. These words are then recognized by automatic speech recognizer, and in the end, system displays the recognized words on the screen[2]. Speech processing can be performed at different three levels. Signal level processing considers the anatomy of human auditory system and process signal in form of small chunks called frames. In phoneme level processing, speech phonemes are acquired and processed. Phoneme is the basic unit of speech. Third level processing is known as word level processing. This model concentrates on linguistic entity of speech. There are few problems faced in speech processing which can be listed as Robustness, Portability, Adaptability, Language Modeling, Out of Vocabulary, and Accent[3].

### 1.1 Speech Recognition Process:

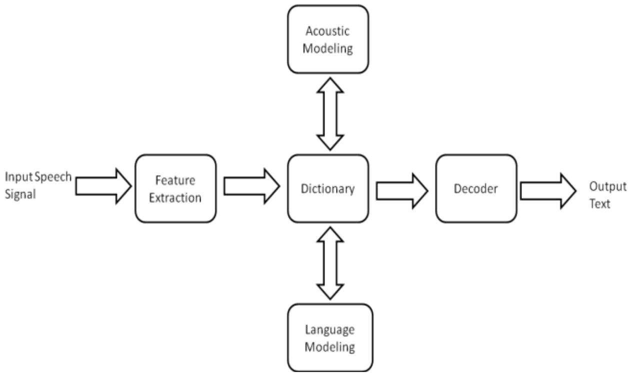

***Figure 1: Steps involved in Speech Recognition.***

Audio input: the spoken speech is taken as an input to the system by means of microphone or any such device.
Feature Extraction: this block does the recording of various speech samples. Digitization of signal takes place in this block where sampling and quantization are performed. Elimination of noise in speech signal is done by quantization.
Acoustic Modeling: the Acoustic modeling provides statistical model of speech signal. The acoustic model assigns probabilities to phonemes, words, or sentences.
Language Modeling: Plain acoustic information is not enough knowledge about the language is also necessary. Comparison of symbols from acoustic model is done with the set of words present in the Dictionary to produce required sequence of words. Decoder is then used for recognizing the exact word.

## II. CLASSIFICATION OF SPEECH PROCESSING

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in Fig.2.

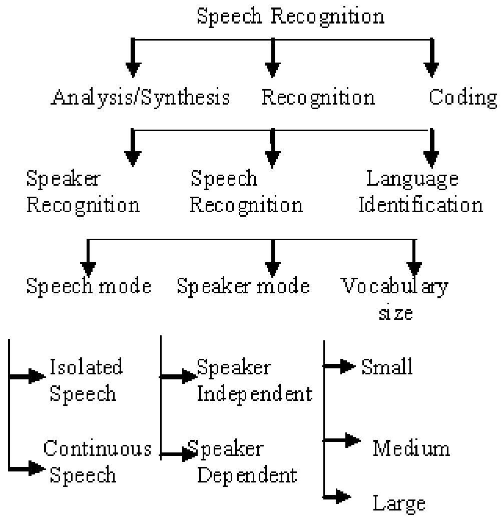

*Figure 2: Classification of Speech Recognition*

- Speech Synthesis: Text- to- Speech.
- Speech Recognition: Speech –to- Text.
- Speech Coding: Is an application of data compression of digital audio signals containing speech.
- Speaker Recognition: Is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves.
- Language identification: The problem of classifying a sample of characters based on its language this is a critical preprocessing stage in the application that apply language specific modeling.
- Speech Mode:

Isolated Speech: Isolated word recognizer usually set necessary condition that each utterance having little or no noise on both sides of sample window. It requires single utterance at a time. Often, these types of speech have “Listen/Not-Listen states”, where they require the speaker to have pause between utterances. Isolated word might be better name for this type.

Continuous Speech: It is normal human speech, without silent pauses between words. This kind of speech makes machine understanding much more difficult.

- Speaker Mode:

Speaker Independent: Trained to respond to word regardless of who speaks therefore system must respond to a large variety of speech patterns.

Speaker Dependent: Trained by individual who will use the system and responds accurately on to the speaker who trained the system.

- Vocabulary Size:

One crucial variable affecting the complexity of a recognition task is the size of the vocabulary. With larger vocabularies the probability of confusing one word to another increases, requiring more accurate modeling in order to avoid recognition errors. Larger vocabularies also mean increased computational load required for decoding utterances. The vocabulary sizes used in ASR have gradually increased from the small vocabulary tasks of number and control word recognition to modern large vocabulary systems with tens of thousands of words in the vocabulary.

## III. FEATURE EXTRACTION TECHNIQUES

As any pattern recognition task, speech recognition process begins by the extraction of relevant features from the input signal. The basic principle in ASR is to extract a sequence of features for each short-time *frame* of the input signal, usually about 100 times in a second. Typically each feature represents a segment of about 20 ms of speech. The assumption under this procedure is that such a small segment of speech is sufficiently stationary to allow meaningful modeling.

### 3.1 Mel Frequency Cepstrum Coefficients (MFCC):

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction method in speech recognition. The technique is called FFT based which means that feature vectors are extracted from the frequency spectra of the windowed speech frames. The Mel frequency filter bank is a series of triangular bandpass filters. The filter bank is based on a non-linear frequency scale called the mel-scale. A 1000 Hz tone is defined as having a pitch of 1000 mel. Below 1000 Hz, the Mel scale is approximately linear to the linear frequency scale. Above the 1000 Hz reference point, the relationship between Mel scale and the linear frequency scale is non-linear and approximately
Logarithmic. One advantage of MFCC is that it is able to mimic Human Auditory System well. But it is very sensitive to noise and has no prediction algorithm.

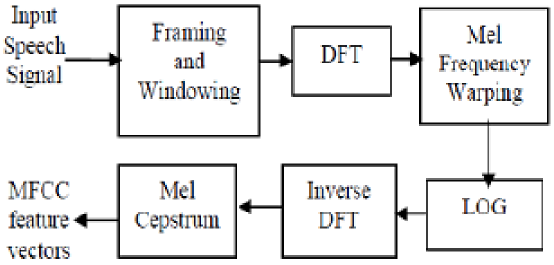

*Figure 3: Steps in MFCC*

### 3.2 Linear Prediction Coefficients (LPC):

LPC (Linear Predictive coding) analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, 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.The coefficients of the difference equation (the prediction coefficients) characterize the formants.LPC provides a good prediction algorithm but it is not well able to mimic the human auditory system.

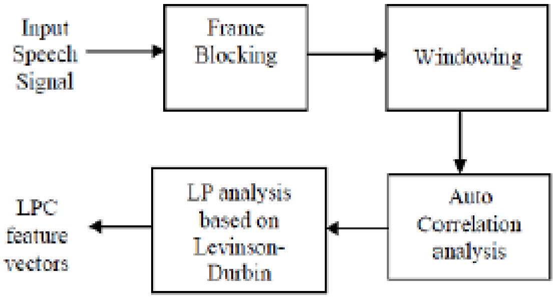
Input Speech Signal
Frame Blocking
Windowing
Auto Correlation analysis
LP analysis based on Levinson-Durbin
LPC feature vectors

*Figure 4: Steps in LPC*

### 3.3 Perceptually Based Linear Predictive Analysis (PLP):

H.Hermansky, B. A. Hanson, H. Wakita proposed a new PLP analysis which models perceptually motivated auditory spectrum by a low order all pole function, using the autocorrelation LP technique. This technique was mainly focused in cross-speaker isolated word recognition. PLP analysis results also demonstrated that speech representation is more consistent than the standard LP method. Basic concept of PLP method is shown in block diagram of Fig. 5. It involves two major steps: Obtaining auditory spectrum and approximating the auditory spectrum by an all pole model. Auditory spectrum is derived from the speech waveform by critical-band filtering, equal loudness curve pre-emphasis, and intensity loudness root compression. The PLP analysis provides similar results as with LPC analysis but the order of PLP model is half of LP model. This allows computational and storage saving for ASR. PLP due to above mentioned advantages is able to mimic Human Auditory system well as well as provides a good prediction algorithm thus covering advantages of both MFCC and LPC.

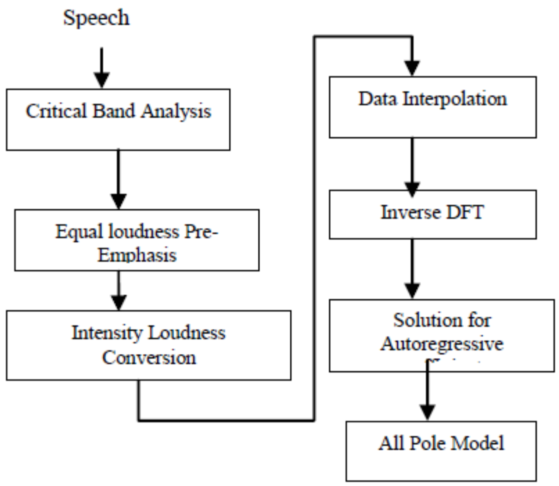
Speech
Critical Band Analysis
Equal loudness Pre-Emphasis
Intensity Loudness Conversion
Data Interpolation
Inverse DFT
Solution for Autoregressive
All Pole Model

*Figure 5: Steps in PLP*

## IV. MODELING TECHNIQUES :

The probabilistic framework used in ASR builds on a statistical model of the speech signal. The role of the acoustic model is to evaluate acoustic feature sequences and assign probabilities for different acoustic classes. Depending on the recognizer, the acoustic model assigns probabilities to phonemes, words, or full sentences.

### 4.1. Hidden Markov model (HMM):

A hidden markov model is signalizing by a finite state markov model and a set of output distributions. The alteration parameter in the Markov chain models are temporal variability's, while the output distribution model parameters are spectral variability. These two types of variability are essential for speech recognition. Hidden Markov modeling is more general and has a secure mathematical foundation compared to template based approach. Compared to knowledge base approach, HMM enables easy incorporation of knowledge sources into organized architecture.HMM do not provide much insight on the recognition process, is negative side effect of HMM. To improve performance of HMM system, analyses of errors of system is made, but it is quite difficult. However, judicious incorporation of knowledge has significantly improved HMM based system.

### 4.2. Dynamic time warping (DTW):

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed.DTW has been applied to video,audio,graphic, infect any data which can be develop into a linear representation can be analyzed with DTW. In general, DTW allows a computer to search an optimal match between two time series if one time series may be "warped" non-linearly by pulling or shriveling it along its time axis. This warping between two time series can then be used to find equivalently regions among the two time series or to diagnose the similarity between two times series.Continuity is not much important in DTW than in other pattern matching algorithms.

### 4.3. Artificial neural networks (ANN):

An artificial neural network contains potentially large number of simple processing element that is called units or neurons, which impact each other's performance via a network of excitatory or repressive weights. It is a feed-forward artificial neural network which has more than one layer of hidden units between its inputs and its outputs. Neural Network provides three types of learning methods namely supervised, unsupervised and reinforced.

**Table 1 below shows a list of techniques used for Automatic speech recognition along with the accuracy provide by the techniques**

| .TECHNIQUE | ACCURACY | REFERENCE |
|---|---|---|
| DTW<br>(MFCC) | 90% (approx) | Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi<br>Ms. Rupali S Chavan*1, Dr. Ganesh S. Sable2 |
| DTW<br>(LPC) | 69% | Ms. Rupali S Chavan*1, Dr. Ganesh S. Sable2 |
| HMM<br>(MFCC) | 90-96% | Ms. Rupali S Chavan*1, Dr. Ganesh S. Sable2<br>**Xinguang Li1, Jiahua Chen1, Zhenjiang Li2** |
| HMM<br>(LPC) | 77% | Ms. Rupali S Chavan*1, Dr. Ganesh S. Sable2 |
| ANN<br>(LPC) | 70% | Viresh Moonsar K.Venayagamoorthy |
| ANN<br>(MFCC) | 80% | Siddhant C. Joshi1, Dr. A.N.Cheeran2 |
| MLP<br>(PLP) | 86% (approx) | Mondher Frikha*, Ahmed Ben Hamida |
| MLP<br>(MFCC) | 84% (approx) | Mondher Frikha*, Ahmed Ben Hamida |

**Table. 1 List of techniques**

## V. PERFORMANCE OF THE SYSTEM

The performance of speech recognition systems is normally referred in terms of accuracy and speed. Accuracy is labeled with word error rate (WER), Whereas speed is labelled with the real time factor. Alternately accuracy can be labelled Single Word Error Rate (SWER) and Command Success Rate

(CSR)[21].Word Error Rate (WER)[20], is a common measurement of the performance of a speech recognition. Normally problems occurred in the performance of system due to mismatch of recognized word sequence with reference word sequence. The WER is derived from the Levenshtein distance [20],operating at the word level. Word error rate calculated as

$$\mathbf{WER} = \frac{\mathbf{S} + \mathbf{D} + \mathbf{I}}{\mathbf{N}}$$

Where
• S is the number of substitutions,
• D is the number of the deletions,
• I is the number of the insertions,
N is the number of words in the reference
Sometimes word recognition rate (WRR) is used

$$\mathbf{WRR} = \mathbf{1} - \mathbf{WER} = \frac{\mathbf{N} - \mathbf{S} - \mathbf{D} - \mathbf{I}}{\mathbf{N}} = \frac{\mathbf{H} - \mathbf{I}}{\mathbf{N}}$$

Where
• H is N-(S+D), the number of correctly recognized words.

## VI. CONCLUSION

In this paper basics of speech recognition are discussed. The performance of the ASR system based on the feature extraction technique and their accuracy is compared in this paper. Based on this review, the advantage of PLP features is more suitable which reduces the complexity of the calculation and offers good recognition result, Along with reduction in time consumption. Other side recognition part is performed by using the HMM or ANN which are used now a days. These methods have their own advantages and disadvantages. HMM is the most popular but, these methods had been implemented with speaker-dependent, with low percentage of accuracy. A Hybrid Approach that can contain Advantages of both the approaches can provide a great improvement in accuracy.

## REFERENCES

[1] Ishan Bhardwaj and Narendra D Londhe. Article: Speaker Dependent and Independent Isolated Hindi Word Recognizer using Hidden Markov Model (HMM). International Journal of Computer Applications 52(7):34-40, August 2012.

[2] Rupali Chavan and Dr. Ganesh Sable. Article: An Implementation of Text Dependent Speaker Independent Isolated Word Speech Recognition Using HMM International Journal of Engineering Sciences& Research Technology 2(9): 2311-2318 September 2013.

[3] Namrata Dave Article: Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. International Journal For Advance Research in Engineering And Technology (ISSN 2320-6802) 07/2013; Volume 1(Issue VI).

[4] Kavita S.Yadav , And M.M.Mukhedkar Article: Review on Speech Recognition. International Journal of Science and Engineering Volume 1, Number 2 – 2013 PP-61-70.

[5] Sanjivani Bhabad and Gajanan K. Kharate Article: An Overview of Technical Progress in Speech Recognition. International Journal of Advanced Research in Computer Science and Software Engineering 3(3): ISSN: 2277 128X March 2013.

[6] Janne Pylkkönen: Decoding ASR 2013, Aalto University publication series Doctoral dissertations.

[7] Nidhi Desai And Prof.Vijayendra Desai Article: Feature Extraction and Classification Techniques for Speech Recognition: A Review. International Journal of Emerging Technology and Advanced Engineering .Volume 3 Issue 12 December 2013 ISSN 2250-2459.

[8] Leena R Mehta , S.P.Mahajan , Amol S Dabhade: COMPARATIVE STUDY OF MFCC AND LPC FOR MARATHI ISOLATED WORD RECOGNITION SYSTEM. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 6, ISSN (Print) : 2320 – 3765 ISSN (Online): 2278 – 8875 June 2013.

[9] Shanthi Therese S. And 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, pp : 479-484 June 2013.

[10] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi: Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617.

[11] Xinguang Li, Jiahua Chen, Zhenjiang Li: English Sentence Recognition Based on HMM and Clustering. American Journal of Computational Mathematics, March 2013, 3, 37-42.

[12] Viresh Moonsar And Ganesh K.Venayagamoorthy : "ARTIFICIAL NEURAL NETWORK BASED AUTOMATIC SPEAKER RECOGNITION USING HYBRID TECHNIQUE FOR FEATURE EXTRACTION".

[13] Siddhant C. Joshi, Dr. A.N.Cheeran: MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition. International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering. Vol. 3, Issue 7, ISSN (Print) : 2320 – 3765 ISSN (Online): 2278 – 8875 , July 2014.

[14] Mondher Frikha, Ahmed Ben Hamida: A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition. American Journal of Intelligent Systems 20112 2(1): 1-8.