A Comprehensive Review on Speech Recognition and Its Techniques

Er. Raman Kaur
Electronic and Communication Engineering
Chandigarh Group of college
Landran, India
E-mail: ramankaur2707@gmail.com

Abstract: This paper provides a review on speech recognition system and its techniques. And provide the advancement in the field of speech recognition system. As speech is a way for the communication between the sender and receiver. A speech recognition system takes speech signal as the input and gives the output in the form of text. This paper describes the basic Automatic Speech Recognition (ASR) System. Provide various Speech recognition techniques such as speech analysis, feature extraction techniques, and matching techniques. This paper gives brief description of feature extraction techniques such as Linear Prediction coding (LPC), Mel frequency Cepstral coefficient (MFCC) and Perceptual Linear Predictive (PLP) technique.

Keywords: Speech Recognition, Feature Extraction, MFCC, LPC, PLP.

I. INTRODUCTION

The basic purpose of speech is human communication i.e. the transmission of the message or any information between the speaker and the listener. In electronic communication system as well as in speech production, the message or any information to be transmitted by the speaker or transmitter is encoded in the form of analog waveform that can be transmitted, stored, manipulated and decoded by the listener or any receiver. The basic analog form of the message is an acoustic waveform that is called speech signal. The speech signal will further synthesized and recognized for many applications. The speech synthesis is the artificial production of human speech. A device used for this purpose is called a speech synthesizer [6].

The speech recognition is the process of converting speech into a sequence of words by a device, independent of the device used to store the speech, the speaker's accent or environment in which speaker is present. This is called Automatic Speech Recognition (ASR).

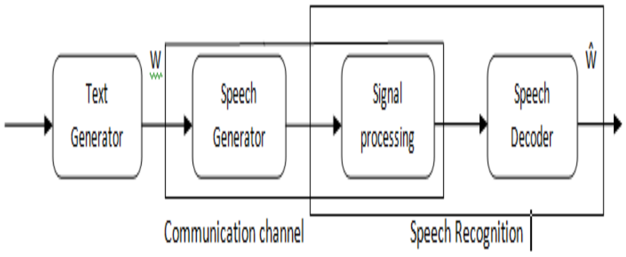

Figure 1.Basic Speech Recognition System

In speech recognition system, a text or any word (w) is processed by a speech generator which produces a speech signal having different pitch, utterance, vowel, and consonant. The combination of speech generator and signal processing is referred as communication channel. The speech is then converted back to the text by speech decoder (ŵ). Signal Processing with the combination of speech decoder is referred as speech recognition.

II. APPLICATION

The several application of speech recognition is as follows:-

A. Voice search: Voice search refers to the searching over the search engine by using the speech as command. The best example of voice search is goggle search by voice.

B. Applications controlled by voice: There are many applications and softwares which are controlled by voice. One can search anything by giving voice as input signal and get output in the form text and voice too.

C. Voice dialing: Voice dialing refers to making a call by giving voice or speech as input. There are many android and iOS devices which gives such feature. E.g. call Raman. It will automatically call that person whose name is given as command.

D. Robotics: The robots can assist a human in many ways or with their everyday activities in home like environment. By using a robust speech recognition system, the communication with robots is done in a natural manner.

E. Data entry: Data entry refers to entering the data using voice E.g. ATM machines.

III. LITERATURE SURVEY

Shreya Narang et.al [2] discussed about different techniques of speech recognition system. Speech recognition system is the system that takes speech as input for processing and provides text as output, they have discussed about various techniques of speech processing like analysis techniques, feature extraction techniques, modeling and matching techniques. They have further discussed and explained the various sub types of these techniques. They have analyzed that hidden Markov Model has been efficiently used for speech recognition and they have discussed the problem of speech recognition system with speech from multiple users and noise that distract the efficient recognition.

Moataz El Ayadi et.al [9] discussed about his survey on speech recognition on the basis of features, schemes of classification, and databases. They have efficiently described the quality of speech such as continuous, qualitative, spectral and TEO based. they have also surveyed the various techniques of speech processing like hidden Markov model, Gaussian mixture model, neural networks, support vector machine, multiple classifier system .They have efficiently analyzed the features used for various characterization of speech signal ,classification techniques that has been used for earlier researches and designs of various databases that has been used for speech recognition.

M.Kalamani et.al [8] discussed about their review on the segmentation of speech signal. They have discussed the two domains of signal processing time domain and frequency domain. Segmentation of speech into words and sub words has been done for efficient speech recognition.

Vimala.C et.al [1] discussed about various challenges and approaches of speech recognition system. They have discussed the different approaches of ASR system and merits and demerits of all the approaches. Various techniques for speech recognition are Template-Based Approaches, Knowledge-Based Approaches, Neural Network-Based Approaches, Dynamic Time Warping (DTW)-Based Approaches, and Statistical based Approaches.

Nidhi Desai et.al [7] discussed about various techniques of speech recognition system. Various types of speech in speech recognition system has been discussed as Isolated speech , Connected word, Continuous speech, Spontaneous speech .various speech recognition techniques that are being used for recognition purpose are Acoustic Phonetic Approach , Pattern Recognition Approach can be further classified into hidden Markov model and dynamic time wrapping ,Artificial Intelligence Approach.

Philippe Dreuw et.al [15] discussed their work on sign language recognition in the speech recognition techniques. Sign language recognition techniques have been efficiently along with the results. They have worked on transferring the many well known models of ASR into the new system for example pronunciation and language modeling.

IV. AUTOMATIC SPEECH RECOGNITION (ASR)

The aim of an ASR system is to accurately and expeditiously convert a speech signal into a text message of the spoken words which is independent of the speaker's accent, device used to store the speech (i.e. the device can be any transducer or microphone), or the environment in which speaker is present (i.e. the environment can be noisy room, quiet place, outdoors). For an accurate ASR system, the variations should be less. The modal of speech production and speech recognition processes or Basic ASR system is shown as below.

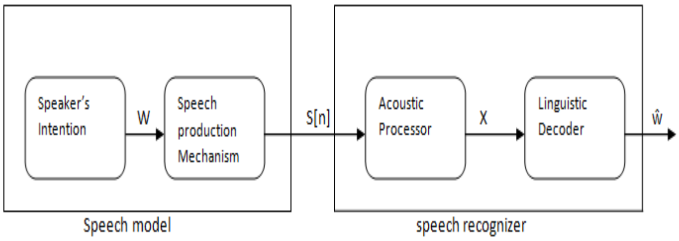

Figure 2.The modal of speech production and speech recognition processes or Basic ASR system.

The basic ASR system involves a speaker modal and a speech recognizer. The speaker modal consist of speaker's intention and speech production mechanisms. If a speaker wants to express some thoughts, the speaker must make a meaningful sentence, w, in the form of a sequence of words i.e. speaker must contain some pauses and other acoustic events such as uh's, um's etc. Once the word is chosen, the speaker sends proper control signal to the speech organs, which forms a speech waveform s[n]. Thus the process of creating the speech waveform from the speaker's intention is called speaker model. Further the speech recognizer consists of acoustic processor and linguistic decoder. The acoustic processor analyze the speech signal and provides the acoustic features to that speech signal and the linguistic decoding makes the words similar to the words of the spoken sentence, hence it results in the recognized sentence ŵ.

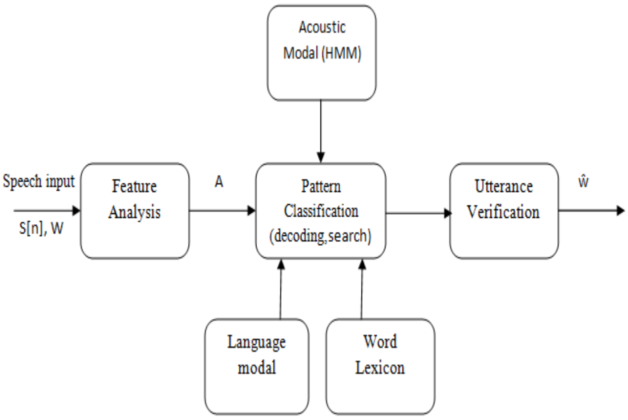

Figure 3.An Overall Automatic Speech Recognition (ASR) system.

The feature analysis provides acoustic features or acoustic vectors $A = \{A_1, A_2, A_N\}$ to the input speech s[n] or w. The pattern classification or the decoding of the speech signal provides the following equation.

$\hat{W} = \arg\max P(W|A)$

By using Baye's theorem;

$P(W|A) = P(A|W).P(W) \backslash P(A)$

Therefore $\hat{W}$ can be written as;

$\hat{W} = \arg\max \ P(A|W).P(W)$

Here P (A) is the probability of acoustic feature or vector. P(W) is the probability of sequence of words or the language model and P(A|W) is acoustic mode[6].

A. Acoustic model

The acoustic modeling plays a vital role in improving accuracy of automatic speech recognition (ASR) because accuracy is the important challenge in the research and development and it is determined by context variation, speaker variations and environment variations. Acoustic modeling deals with the pronunciation modeling which depicts that how a sequence and multi-sequence of speech signal is used to represent a large speech unit such as words or phrases. The most common type of acoustic model is Hidden Markov Model (HMM).

B. Language model

The role of the language modeling in speech recognition is to provide value P (W) i.e. the probability of sequence of words and it is based on possibility of occurrence of that word in the task performed by the speech recognition system. The probability of word W= “call car” is zero for a telephone number identification because that word makes no sense for the task. The language model can be trained for specific task by statistical learning and by rule-based learning.

V. SPEECH RECOGNITION TECHNIQUES

The goal of the ASR is to analyze, extract, characterize and recognize information about the speaker’s identity. The speech recognition can be done by following techniques.

Analysis Technique

Feature Extraction Technique

Matching Technique

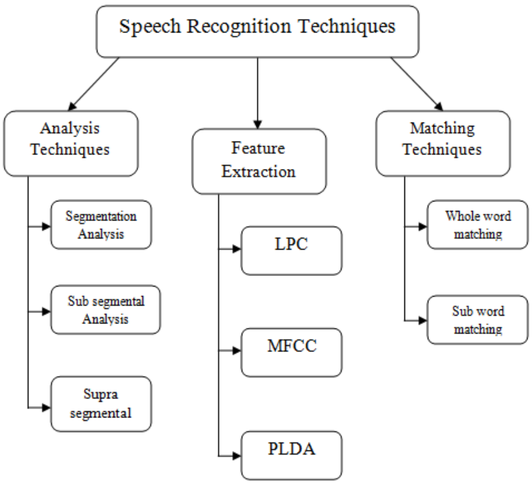

Figure 4.Speech recognition techniques [3].

A. Speech Analysis techniques

Speech signal contain different type of information regarding speaker’s identity. The information varies because of excitation source, vocal tract system, and behavior feature. Speech analysis stage deals with segmenting the speech signal for further analysis and extraction. The speech analysis can be done by following techniques.

1. Segmentation analysis

In this case the speech signal is analyzed using the frame size and shift in the range of 10-30 milliseconds to express the information regarding speaker.

2. *Sub segmental analysis*

In Sub segmental analysis, the extraction of the speaker’s information is done by using frame size and shift in the range of 3-5 milliseconds. The analysis and extraction of features of the excitation state is done by Sub segmental analysis.

3. Supra segmental analysis

The analysis to extract the behavior feature is done by Supra segmental analysis. The information of speaker is provided by using frame size as well as the shift in the range of 50-200 milliseconds.

B. Feature extraction techniques

In speech recognition system, feature extraction plays a vital role. As it helps in extracting the characteristics or features of a speech signal, required for the speaker’s identification. There are following techniques used for feature extraction [2].

1. Linear prediction coding (LPC)

LPC is an important and powerful feature extraction technique. It provides the basic features and estimate of the speech signal. The estimated signal is accurate and efficient. The input speech signal is compared to past signals. The difference between the actual speech sample and predicted speech values is squared and through minimize this value, a set of predicted coefficients and parameters can be observed. These coefficients are the basis for the LPC [1]. The more robust form of these coefficients is called Cepstral coefficients. The sound produced through this technique is very close to the vocal tract input signal. The Linear prediction coding (LPC) is a static technique in speech recognition system. This is a reliable and accurate technique. There is a good computational speech in this technique. The speech at low bit rate can be encoded by Linear prediction coding (LPC). But there are residual errors in this technique.

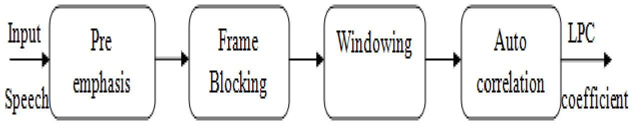

Figure 5.LPC feature extraction technique.

2. *Mel frequency Cepstral coefficient (MFCC)*

It is the important technique for the feature extraction and widely used. MFCC has good response as compared to the LPC. In this technique, the input is speech samples which are used to extract the coefficients. The discontinuities of the signal are minimized by using hamming window. The Mel filter bank is produced by using Discrete Fourier Transform

(DFT) [1].After Mel frequency wrapping the number of coefficients. MFCC is used for the speech processing task. The accuracy is high in MFCC technique as compared to LPC. There are number of advantage such as performance rate is high, low complexity. But the performance can be affected by number of filters which are used in this technique [5].

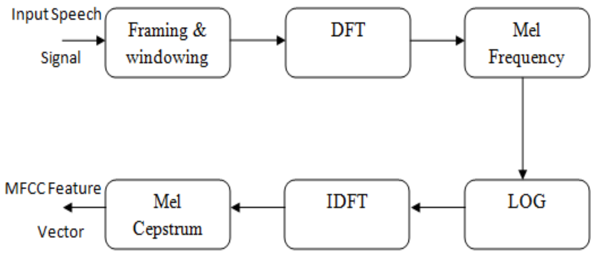

Figure 6.MFCC feature extraction technique.

3. Perceptual Linear Predictive (PLP)

This technique is based on the linear predictive analysis and the short term spectrum of speech. It is the combination of Discrete Fourier transforms (DFT) and linear Prediction (LP) [13].

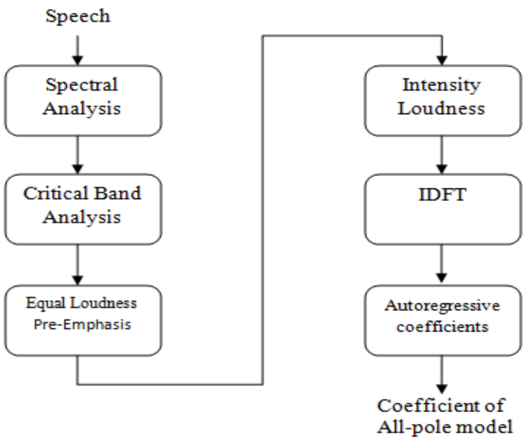
Speech
Spectral Analysis
Critical Band Analysis
Equal Loudness Pre-Emphasis
Intensity Loudness
IDFT
Autoregressive coefficients
Coefficient of All-pole model

Figure 7.Block diagram of PLP [13].

The spectral analysis generates a speech signal and converted to a power spectral density. The power spectrum is estimated and warped into the Bark frequency. To smooth the signal, the spectrum produced by the Bark scale and power spectrum produced by critical band filter is convolved. At different frequencies different loudness perception will compensate by pre emphasis. Perceived loudness is approximately the cube root of the intensity. The approximation is true for reasonable sound but not for very loud or very quite sound. Auto correlation function is generated by the Inverse Discrete Fourier Transform (IDFT). Then the auto regressive coefficients are translated into Cepstral coefficients of the all poll model.

The main difference between the MFCC and PLP is that MFCC uses Mel frequency for spectrum warping and PLP uses Bark scale for spectrum warping.

C. Matching techniques

The words spoken by the speaker is detected by the speech recognizer system to match that word with the set of words which are already in the database. The following techniques are used for this purpose [3].

1. Sub word matching

Sub words are the phonemes which are used in between the speech signal. These phonemes are looked up by search engine on which the system later performs pattern recognition. This technique requires storage of 5 to 20 bytes per word which is less and this technique also takes large amount of processing.

2. Whole word matching

In this technique, there are pre-recorded templates in the database. When a particular word is spoken then it will matched to pre-recorded templates and gives desired results. This technique requires less processing and the storage is also larger than the sub word matching technique. But in this technique every word is to be recorded before the processing. So that system will recognize the word appropriately. This may requires large amount of vocabulary which may create mismatch during processing. This technique needs storage of 50 to 512 bytes per word which is very large [2].

VI. SPEECH RECOGNITION APPROACHES

A. Template based approach

As speech signal is used as the input to the system and in order to find the best match, the unknown speech signal is compared with the words or template which are pre recorded in the system. These templates are used as the reference for the unknown speech signals. There are perfectly accurate words models used in this approach which become advantageous to this approach. But the pre recorded words or templates are fixed which is a drawback of this technique. In this approach a large size of vocabulary is needed. This method is inefficient for storage and processing. This approach is usually speaker dependent.

B. Knowledge based approach

The knowledge based or rule based approach is proposed by several researchers and then applied to speech recognition. The expert knowledge about linguistic, spectrogram and phonetics is used in this approach [4]. Knowledge about variations of speech is hand coded into the system and then stored into the database. The set of features from the speech are taken and then the system is to be trained to produce the set of rules automatically from the speech samples. These rules provide information about the classification. The modelling of variations in speech is explicitly done by this approach. But this expert knowledge is difficult to obtain and use successfully.So this approach is considered as impractical and not commonly used for the speech recognition system. Further approaches are used for the speech recognition system to improve the performance of the system.

C. Statistical based approach

In this approach, the speech variations are modelled statistically by using Hidden Markov Model (HMM)[10]. The general purpose speech recognition systems are based on statistical acoustic and language models. For parameter approximation, a large amount of acoustic and language data is required. This approach is inaccurate and has limited system performance.

D. Neural network based approach

This approach is also known as learning based approach because learning methods could be introduced such as neural networks and genetic algorithms. That's why this technique is referred as neural network based approach. This approach overcomes the problems related to the statistical based approach or Hidden Markov Model (HMM). This approach can handle noisy data, low quality, and speaker independency [11]. This approach provides better accuracy than the Hidden Markov Model. Phoneme recognition is also an approach which is using neural networks and is called as NN-HMM hybrid system.

E. Dynamic Time warping (DTW) based approach

This approach provides the similarity between the two sequences which may vary in time and speed. This approach is used in the ASR to deal with the different speed of the vocalization. With certain restrictions the optimal match between two given sequences can be produced with the use of a computer. The two sequences are warped non-linearly to match each other. The sequence alignment method is often used in the context of Hidden Markov Model. Continuity is less important in DTW than in other pattern matching algorithms. DTW is best suited to matching sequence with missing in formation [3].

VII. CONCLUSION

At the end, it is concluded that speech recognition is a good area for research in various fields. In this paper, the fundamentals of speech recognition and its recent progress is discussed. This paper provides a brief description of Automatic Speech Recognition (ASR) system. This is extensively used in many applications. The different techniques have been discussed among which feature extraction techniques are used for the better system performance and results. Feature extraction techniques such as MFCC, LPC, and PLP are explained. Among all these techniques PLP is better in order to provide accurate and robust performance and results. The various approaches for speech recognition are also discussed in this paper.

ACKNOWLEDGEMENT

Er. Raman Kaur pursing M.tech (ECE) from Chandigarh Group of Colleges (CGC), Landran would like to thank my friends and teachers for helping me in this paper.

REFERENCES

[1] Vimala .C et al.,"A Review on speech recognition challenges and approaches" , world of computer science and information technology journal ,vol.2 , pp-1-7 , 2012.

[2] Shreya Narang, Ms. Divya Gupta ," Speech feature Extraction Techniques : A Review" , International Journal of Computer Science and Mobile Computing, Vol.4 Issue.3, March- 2015.

[3] Santosh K.Gaikwad and Pravin Yannawar, A Review, International Journal of Computer Applications ***A*** Review on Speech Recognition Technique Volume 10– No.3, November 2010.

[4] Ranu Dixit, Navdeep Kaur, " Speech Recognition using Stochastic Approach : A Review" , International journal of innovative research in science ,engineering and technology , Vol.2 ,issue 2 ,Feb 2013.

[5] Douglas Shaughnessy, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September © 2003 IEEE.

[6] Bassam A. Q. Al-Qatab , Raja N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)", 978-1-4244-6716-711 0/ ©2010 IEEE.

[7] Nidhi Deasi et al , "Feature extraction and classification techniques for speech recognition : A review", International Journal of Emerging Technology and Advanced Engineering, Vol 3 ,isssue 12 ,2013.

[8] M.Kalamani1, Dr.S.Valarmathy et al , "Review of Speech Segmentation Algorithms for Speech Recognition" , International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 3, Issue 11, November 2014.

[9]Moataz El Ayadi, "Survey on speech emotion recognition: Features, classification schemes, and databases", Engineering Mathematics and Physics, Cairo University, Elsevier @2012.

[10] Suman K. Saksamundre et al, "A Review on different approaches for speech recognition system", International journal of computer application, Vol 115, April 2015.

[11] Vimal Krishnan V. R, Athulya Jayakumar and Babu Anto.P, "Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network", 4th IEEE International Symposium on Electronic Design, Test & Applications, 0-7695-3110-5/08 $25.00 © 2008 IEEE.

[12] Chia-Ping Chen Jeff Bilmes and Daniel P. W. Ellis, Department of Electrical Engineering University of Washington Seattle, WA on Speech Feature Smoothing for Robust ASR.

[13] Cedric Gaudard, "Speech Recognition based on template matching and phone posterior probabilities", IDIAP Communication Edition 2007.

[14] Philippe Dreuw, David Rybach, Thomas, "Speech Recognition Techniques for a Sign Language Recognition System", Human Language Technology and Pattern Recognition Computer Science Department , Germany.

AUTHOR's BIBLOGRAPHY

Er. Raman Kaur, Research scholar, pursing masters in Electronic and Communication Engineering from Chandigarh Group of Colleges, Landran and received bachelor's degree in Electronic and Communication Engineering from College of Engineering and Management, Kapurthala.