ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 8, February 2016

Speech Recognition Using Signal Processing Techniques

Megha Agrawal[#1], Tina Raikwar[#2]
Department of Electronics and Communication, NIIST-Bhopal, India

Abstract— In this paper, the fundamentals of speech recognition are discussed and its recent progress is investigated. The various approaches available for developing an ASR system are clearly explained with its merits and demerits. This paper presents a speech recognition system based on signal processing techniques. The performance of the adopted ASR system based on the adopted feature extraction technique and the speech recognition approach for the particular language is compared in this paper.

Keywords— **Speech, ASR, Feature Extraction, Signal Processing**

I. INTRODUCTION

The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition.

Speech processing is one of the exciting areas of signal processing. The goal of speech recognition area is to developed technique and system to developed for speech input to machine based on major advanced in statically modelling of speech ,automatic speech recognition today find widespread application in task that require human machine interface such as automatic call processing.[1]. Since the 1960s computer scientists have been researching ways and means to make computers able to record interpret and understand human speech.

Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer. computer which can speak and recognize speech in native language [2].

To identify the voices of the unknown speaker we need to:

- ✓ Extract characteristic features of the speech of the known speakers.
- ✓ Create models of the features of the known speakers.
- ✓ Compare the features from the unknown speaker's utterances with the statistical models of the voices of the speakers known to the system.
- ✓ Make decision when we have identified that test utterance belongs to a certain speaker.

II. CLASSIFICATION OF SPEECH RECOGNITION SYSTEM

Speech recognition systems can be separated in several different classes by describing the type of speech utterance, type of speaker model, type of channel and the type of vocabulary that they have the ability to recognize. Speech recognition is becoming more complex and a challenging task because of this variability in the signal. These challenges are briefly explained below:

A. Types of Speech Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences [3]. The types of speech utterance are:

1) Isolated Words

Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results.

2) Connected Words

Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

3) Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation. It includes a great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words. Continuous speech recognition systems are most difficult to create because they must utilize special methods to determine utterance boundaries. As vocabulary

DOI:10.17605/OSF.IO/ZAB7G

grows larger, confusability between different word sequences grows.

4) *Spontaneous Speech*

This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and non- words.

B. *Types of Speaker Model*

All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into two main categories based on speaker models namely speaker dependent and speaker independent.

1) *Speaker dependent models*

Speaker dependent systems are designed for a specific speaker. They are generally more accurate for the particular speaker, but much less accurate for other speakers. These systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems.

2) *Speaker independent models*

Speaker independent systems are designed for variety of speakers. It recognizes 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. Some applications only require a few words (e.g. Numbers only), others require very large dictionaries (e.g. Dictation machines). In ASR systems the types of vocabularies can be classified as follows.

- ✓ Small vocabulary - tens of words
- ✓ Medium vocabulary - hundreds of words
- ✓ Large vocabulary - thousands of words
- ✓ Very-large vocabulary - tens of thousands of words
- ✓ Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word

III. ASR SYSTEM

The task of ASR is to take an acoustic waveform as an input and produce output as a string of words [4]. Basically, the problem of speech recognition can be stated as follows. When given with acoustic observation $X = X_1, X_2 \ldots X_n$, the goal is to find out the corresponding word sequence $W = W_1, W_2 \ldots W_m$ that has the maximum posterior probability $P(W|X)$ expressed using Bayes theorem as shown in equation (1). The figure 1 shows the overview of ASR system.

$$W{=}\mathrm{argmax}\ \frac{P(W)P(X/W)}{P(X)} \qquad \text{..... (1)}$$

Where P(W) is the probability of word W uttered and P(X|W) is the probability of acoustic observation of X when the word W is uttered.

An essential task of developing any ASR system is to choose the suitable feature extraction technique and the recognition approach. The suitable feature extraction and recognition technique can produce good accuracy for the given application. Hence, these two major components are reviewed and compared based on its merits and demerits to find out the best technique for speech recognition system.

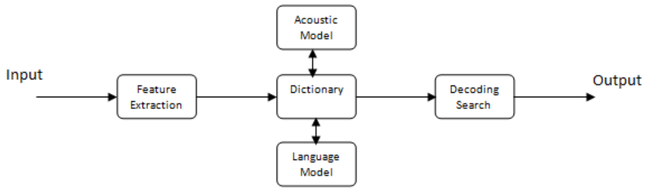

Fig 1. ASR System

IV. USED METHODOLOGY

Feature Extraction Techniques

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task [3]. But extracted feature should meet some criteria while dealing with the speech signal such as:

- ✓ Easy to measure extracted speech features
- ✓ It should not be susceptible to mimicry
- ✓ It should show little fluctuation from one speaking environment to another
- ✓ It should be stable over time
- ✓ It should occur frequently and naturally in speech

The most widely used feature extraction techniques are explained below.

A. *Linear Predictive Coding (LPC)*

One of the most powerful signal analysis techniques is the method of linear prediction. LPC [7][8] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through 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 [9]. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients.

B. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [7] [8] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [10], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the centre frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation [7] [8]. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula (2).

$$\mathrm{Mel}(f)=2595*\log10\left(1+\frac{f}{700}\right) \quad (2)$$

In speech recognition, the main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs some kind of spectra temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage (which is not always present) transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer. Although there is no real consensus as to what the optimal feature sets should look like, one usually would like them to have the following properties: they should allow an automatic system to discriminate between different through similar sounding speech sounds, they should allow for the automatic creation of acoustic models for these sounds without the need for an excessive amount of training data, and they should exhibit statistics which are largely invariant across speakers and speaking environment.

V. RESULT

This method incorporates the use of windowing technique at the pre processing stage of speech signal to smoothen it. Here, in this paper Hamming window. The result shown in next figure windows are .

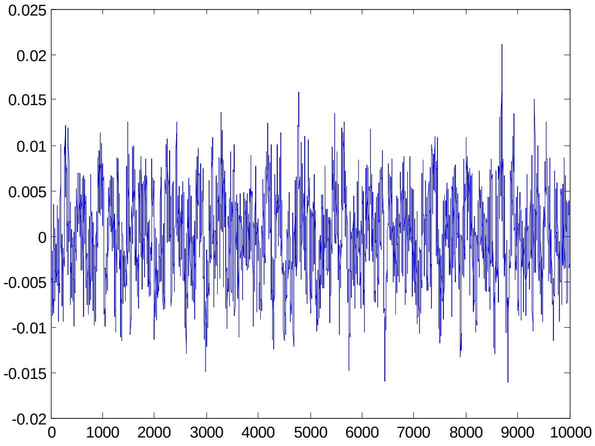

Fig 2 Speech Signal Pass through Hamming window

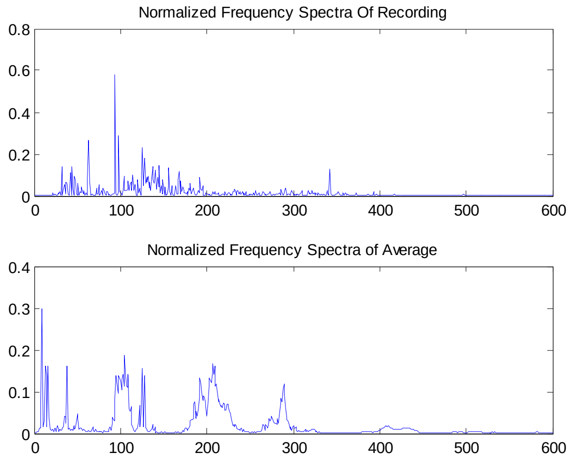

Fig 3 MFCC component of Speech signal pass through Hamming window

VI. CONCLUSION AND FUTURE WORK

Speech recognition has been in development for more than 50 years, and has been entertained as an alternative access method for individuals with disabilities for almost as long. In this paper, the fundamentals of speech recognition are discussed and its recent progress is investigated. The various approaches available for developing an ASR system are clearly explained with its merits and demerits. The performance of the ASR system based on the adopted feature extraction technique and the speech recognition approach for the particular language is compared in this paper. In recent years, the need for speech recognition research based on large vocabulary speaker independent continuous speech has highly increased. Based on the review, the potent advantage of HMM approach along with MFCC features is more suitable for these requirements and offers good recognition result. These techniques will enable us to create increasingly powerful systems, deployable on a worldwide basis in future.

REFERENCES

[1] Soon Suck Jarng," HMM Voice Recognition Algorithm Coding", International Conference on Information Science and Applications (ICISA), pp. 1-7, ISBN-978-1-4244-9222-0, 2011 IEEE.

[2] Peerapol Khunarsal et.al, "Singing Voice Recognition based on Matching of Spectrogram Pattern", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, pp.1595-1599, ISBN: 978-1-4244-3548-7, 2009 IEEE.

[3] Vimala.C, and Radha.V, "A Review on Speech Recognition Challenges and Approaches", ISSN: 2221-0741, Vol. 2 No. 1, pp.1-7, WCSIT, July 2012.

[4] Santosh K.Gaikwad, et.al," A Review on Speech Recognition Technique", IJCA, Volume 10– No.3, November 2010

[5] M.A. Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review ", IJCSIS, Vol. 6, No. 3, 2009.

[6] R.K.Moore,"Twenty things we still don t know about speech", Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology, UK, July 1994.

[7] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, ISBN: 953-7044-03-3, pp.115-118, 07-09 June 2006, Zadar, Croatia.

[8] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, ISSN: 0018-9219, pp. 1272-1305, IEEE.

[9] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 ISBN: 1793-8201.

[10] Bassam A. Q. Al-Qatab , Raja N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)", ISBN: 978-1-4244-6716-711, Vol. 2, pp. 557-562, 2010 IEEE.

[11] Wang. T," Two-Dimensional Speech-Signal Modeling", pp-1843-1856, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 6, AUGUST 2012.

[12] Chen.S, et.al, "A New Prosody-Assisted Mandarin ASR System", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, pp-1669-1684, 2012.

[13] M.Chen, et.al, "The Design of Voice Recognition Controller via Grey Relational Analysis", International Conference on System Science and Engineering, pp-477-481, IEEE-2011.

[14] Zhang. J, "A novel voice recognition model based on HMM and fuzzy PPM", ICSP, pp- 637-640, IEEE-2010.