A STUDY ON SPEECH RECOGNITION SYSTEM: A LITERATURE REVIEW

Shikha Gupta[1],Mr.Amit Pathak[2],Mr.Achal Saraf[3]

1 (Research Student ,Dept of Electronics &Comm Engg,SRIST ,M.P ,Jabalpur)

2 (Head of Dept, Dept of Electronics &Comm Engg ,SRIST ,M.P ,Jabalpur)

3(Assistant Professor,Dept of Electronics &Comm Engg ,SRIST,MP,Jabalpur)

Abstract-Speech recognition are becoming more and more useful nowadays. Various interactive speech aware applications are available in the market. Speech recognition systems are the efficient alternatives for such devices where typing becomes difficult. But they are usually meant for and executed on the traditional general-purpose computers. With growth in the needs for embedded computing and the demand for emerging embedded platforms, it is required that the speech recognition systems (SRS) are available on them too. PDAs and other handheld devices are becoming more and more powerful and affordable as well. It has become possible to run multimedia on these devices.

Keywords-Feature extraction, Feature Matching, Modeling of speech

I.INTRODUCTION

The Speech is the most common & primary mode of communication among human beings. It is the most natural and efficient form of exchanging information among humans . Human voice conveys much more information such as gender, emotion and identity of the speaker. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means an Algorithm.

The objective of speech recognition is to determine which speaker is present based on the individual's characterization [1].Several techniques have been proposed for compensating the mismatch occurred between the testing and training sessions. The communication among human computer interaction is called human computer interface.

Since 1960s computer scientists have been researching ways and means to make computers able to record, interpret and understand human speech. In computer science, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Speaker recognition is the identification of the person who is speaking by characteristics of their voices (voice biometrics), also called voice recognition.

II.CLASSIFICATION OF SPEECH

A number of parameters define the capability of a speech recognition\ system[2].

i)Isolated word: The Isolated word have sample windows.it accepts single word or single utterances at a time.Isolated utterance might be a better name of this work[3].

ii)Connected word: The Connected word system are similar to isolated words but allow separate utterance to be "run together minimum pause between them.

iii)Continuous speech : It allows user to speak almost naturally, while the computer will examine the content.there are special methods used to determine utterance boundaries and various difficulties occurred in it.

iv)Spontaneous speech:A System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together.

III.SPEECH RECOGNITION TECHNIQUES

The goal of speech recognition is to analyze, extract, characterize and recognize information about the speaker identity. Variety of the techniques are used for determining the speech characteristics.

Speech analysis technique

The speech data contain different type of information that shows the speaker identity. This includes speaker specific information due to vocal tract, excitation

ISSN: 2278 – 7798

source and behavior feature. The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting [4].These are of three types.

i) **Segmentation analysis**
In this work, speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information. This method is used to extract vocal tract information of speaker recognition.

ii)**Sub segmental analysis**
Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis. This technique is used to mainly analyze and extract the characteristic of the excitation state. [5].

iii)**Supra segmental analysis**
In this work, speech is also analyzed using the frame size.This technique Is mainly used to analyze and characteristic the behaviour character of the speaker.

IV.MODELING TECHNIQUE

The aim of modeling technique is to use the specific feature of the speaker for creating speaker models. The speaker modeling technique is basically classified as speaker recognition and speaker identification. The speaker identification technique defines who is speaking on basis of individual information obtained from speech signal. The speaker recognition is further divided into two parts i.e speaker dependent and speaker independent. In the speaker independent mode of the speech recognition the computer ignore the speaker specific characteristics of the speech signal and extract the useful message. On the other hand in case of speaker recognition machine should extract speaker characteristics in the acoustic signal [7].Then comparsion of speech signal from an unknown speaker to a database of known speaker has been done.

Speaker recognition can also be divided into two methods, text- dependent and text independent methods. In text dependent method the speaker speakes key words or sentences having the same text for both training and testing trials whereas text independent does not rely on a specific texts being spoken [8]. Following are the methods used in speech recognition process are as follows:

i).Pattern Recognition approach
A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. A pattern recognition has been developed over two decades and received much attention and applied widely in many practical problem .It involves two essential steps namely pattern training and pattern comparison. The essential feature of this approach is that it uses a well defined mathematical framework and then creates speech pattern representations. The pattern-matching approach has become the predominant method for speech recognition in the last six decades ([9] pg.87.

ii). The acoustic-phonetic approach
This method has been studied and used for more than 40 years.This approach is based upon theory of acoustic phonetics and postulates [10]. The work done before to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units in spoken language and these units are broadly characterized by a set of acoustics properties that are changed in the speech signal over time.There are three methods that have been applied to the language identification i.e Problem phone recognition, Gaussian mixture modeling, and support vector machine classification.

iii).Learning based approaches
To overcome the disadvantage of the HMMs machine learning methods which was introduced in neural networks and genetic algorithm programming learning based approaches has been taken. In learning based approaches ,they can be learned automatically through emulations or evolutionary process.

iv) Knowledge based approaches
The guidance should be taken from an expert knowledge about variations in speech is hand coded into a system.This approach gives the advantage of explicit modeling but this situation is difficult to obtain and cannot used successfully. Knowledge based approach uses the information
regarding linguistic, phonetic and spectrogram.The test speech is considered by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure [11]. Vector Quantization (VQ)[12] is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction.

v)Artificial intelligence approach
The artificial intelligence approach coordinate the recognition procedure according to the person who applies it.The intelligence of a person such as visualizing, analyzing etc are used for making a decision on the measured acoustic features. The

ISSN: 2278 – 7798

Artificial Intelligence approach [13] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In its pure form,knowledge engineering design involves the direct and explicit incorporation of experts speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Knowledge enables the algorithms to work better. This form of knowledge based system increases the contribution and hence successful designs and strategies has been reported.

V.FEATURE EXTRACTION

The extraction of the features of the parameters which represent an acoustic signal is an important task to produce a better recognition performance. The efficiency of this method is important for the next method since it affects its behaviour. Various are the feature extraction methods available with their features.

i) In Principal Component analysis (PCA) It uses Non linear feature extraction method and gives Linear map and is fast and eigenvector-based.

ii)In Linear Discriminate Analysis(LDA),it depends on Non linear feature extraction method, it has Supervised linear map and are fast and eigenvector-based.This method is better than PCA for classification[6]

iii)The Linear Predictive coding uses Static feature extraction method which has 10 to 16 lower order coefficient. It is used for extracting features at the lower order.

iv)In Mel-frequency cepstrum (MFCCs).it has the property that the Power spectrum is computed by performing Fourier Analysis.

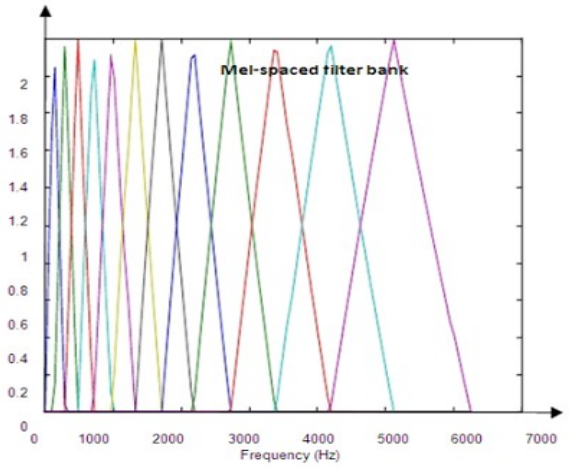

Fig.1 Mel Frequency Cepstral Coefficients.

v)The wavelet analysis gives better time resolution than Fourier Transform because It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform.

VI.FEATURE MATCHING

Various are the techniques used in feature extraction such as Dynamic Time Wraping (DTW),Vector Quantization (VQ) ,LBG etc .Each technique has its own feature matching function and specification

DTW: Dynamic time warping is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. DTW is a method that calculates an optimal match between two given sequences. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data which can be turned into a linear sequence can be analyzed with DTW. Applications include speaker recognition and online signature recognition.

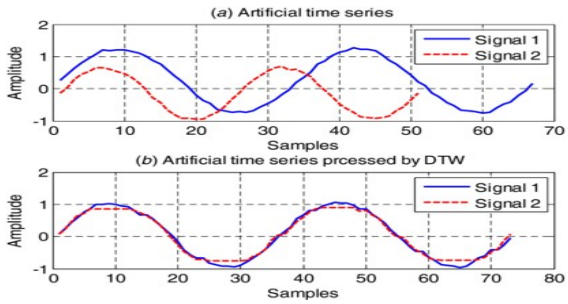

Fig.2.Dynamic Time Wrapping of two speech signal.

VQ: Vector quantization (VQ) is a classical quantization technique from signal processing. It was originally used for data compression[14]. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms. The density matching property of vector quantization is very powerful for large and high-dimensioned data. Hence VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation.

ISSN: 2278 – 7798

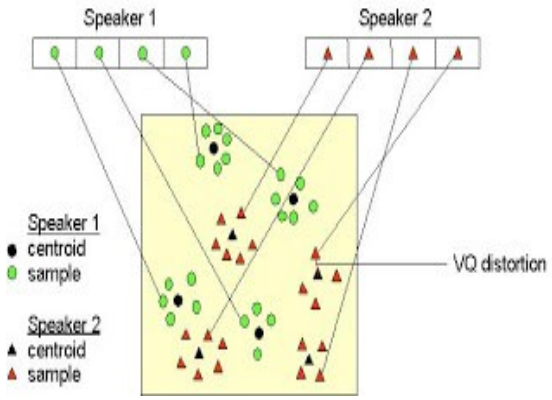

Fig.3 Vector Quantization of two speech signals.

LBG: Linde-Buzo-Gray (LBG) Algorithm: This is an algorithm developed in the community of vector quantization for the purpose of data compression[15].One speaker can be discriminated from another based on the location of centroids codebook for this speaker using those training vectors for clustering a set of L training vectors into a set of M codebook vectors.

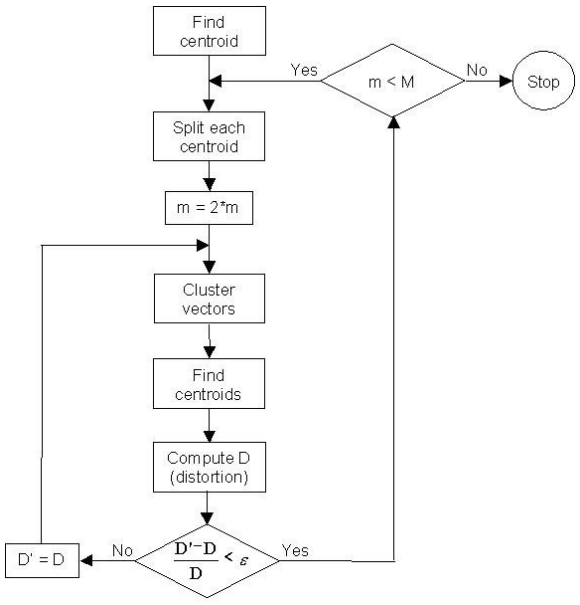

Fig.4 Linde-Buzo-Gray (LBG) Algorithm

VII.CONCLUSION AND FUTURE WORKS

In this paper, various techniques are discussed about speech recognition system. This paper also present the list of technique with their properties of Feature extraction and Feature matching .Through this review paper it is found that MFCC is widely used for feature Extraction and VQ is better over DTW. Comprehensive investigation, both experimental and theoretical of the problem have to be done for better results and for making the system more robust.

VIII.REFERENCES

[1]. Cheong Soo Yee and abdul Manan ahmad, Malay Language Text Independent Speaker Verification using NN-MLP classifier with MFCC, 2008 international Conference on Electronic Design.

[2].http://crdo.up.univ aix.fr/ExternalDisk0/preview/000836/node303.html

[3] Zahi N.Karam,William M.Campbell "A new Kernel for SVM MIIR based Speaker recognition "MIT Lincoln Laboratory, Lexington, MA, USA.

[4] GIN-DER WU AND YING LEI " A Register Array based Low power FFT Processor for speech recognition"Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan.

[5] Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 "MFCC Compensation for improved recognition filtered and band limited speech" Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA.

[6] M.A.Anusuya , S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[7] Samudravijay K "Speech and Speaker recognition report" source:http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html Viewed on 23 Feb. 2010

[8] Sannella, M Speaker recognition Project Report report" From http://cs.joensuu.fi/pages/tkinnu/research/index.html Viewed 23 Feb. 2010

[9] C.S.Myers and L.R.Rabiner, A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition , IEEE Trans. Acoustics, Speech Signal Proc.,ASSP-29:284- 297,April 1981.

[10] IBM (2010) online IBM Research Source:- http://www.research.ibm.com/Viewed 12 Jan 2010.

[11] L.R.Bahl et.al, A method of Construction of acoustic Markov Model for words, IEEE Transaction on Audio ,speech and Language Processing ,Vol.1,1993

ISSN: 2278 – 7798

[12] Keh-Yih Su et.al., Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.

[13] Tavel 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 1994.

[14]http://en.wikipedia.org/wiki/Vector_quantization

[15]Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, Vol. COM-28, pp. 84-95, Jan. 1980.

Authors Biblography

Miss.Shikha Gupta

Miss.Shikha Gupta passed B.E. and pursuing M.Tech. She has now studing at Electronics and Communication Engineering at Shriram Institute of Science and Technology, Jabalpur.

Assist Prof. Mr.Amit Pathak

Assist Prof.Mr.Amit Pathak passed B.E. and M.Tech. He has 7 years of teaching experience in different engineering colleges and now working as Assist Prof, Electronics and Communication Engineering at Shriram Institute of Science and Technology, Jabalpur. He has five research publications in national and international journals.

Assist Prof. Mr.Achal Saraf

Assist Prof.Mr.Achal Saraf passed B.E. and M.Tech. He has 5 years of teaching experience in different engineering colleges and now working as Assist Prof, Electronics and Communication Engineering at Shriram Institute of Science and Technology, Jabalpur. He has two research publications in national and international journals.

ISSN: 2278 – 7798