Speech Signal Analysis and Speaker Recognition by Signal Processing

Seema P
Department of ECE,
RYMEC, Ballari , Karnataka, India.

***Abstract*— Recent developments in digital signal processing (DSP) technology make it easier for scientist to develop powerful personal computer based data acquisition and analysis system. Speaker recognition by signal processing technique is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves .Speaker recognition methods can be divided into text independent and text independent methods .In a text independent system, speaker models capture characteristics of somebody speech ,which show up irrespective of what one is saying .On the other hand in a text independent system the recognition of speaker identity is based on the person speaking one or more specific phrases like passwords ,card numbers etc .This technique makes it possible to use the speaker voice to verify their identity and control access to services such as voice dialing ,banking by telephone ,telephone shopping ,data base access services, information services ,voice mails ,security control for confidential information areas ,and remote access to computers. Area of artificial intelligence where machine performance can exceed human performance-using short utterances and large number of speakers, machine accuracy often exceed that of human being .Thus ,with the help of MATLAB .**

Keywords—Speech recognition, Digital signal processing , Telephone shopping , Artificial intelligence , MAT lab.

I. INTRODUCTION

The human speech conveys different types of information. The primary type is the meaning or words, which speaker tries to pass to the listener. But the other types that are also included in the speech are information about language being spoken, speaker emotions, gender and identity of the speaker.

The goal of automatic speaker recognition is to extract, characteristic and recognize the information about speaker identity. speaker recognition is usually divided into two different branches ,speaker verification and speaker identification. Speaker verification task is to verify the claimed identity of person from his voice .This process involves only binary decision about claimed identity. In the speaker identification there is no identity claim and the system decides who the speaking person is. Speech recognition will revolutionize the way people conduct business over the web and will ultimately differentiate the world class E-business voice-XML ties speech recognition and telephony together and provides the technology with business benefits. Ex: voice enabled web solutions. These solutions can greatly expand the accessibility .

II. LITERATURE REVIEW

[1] This book provides a theoretical sound, technically accurate, and complete description of the basic knowledge and ideas that constitute a modern system for speech recognition by machine. Covers production, perception and acoustic-phonetic characteristic of the speech signals ;signal processing and analysis methods for speech recognition by machine[7]. pattern comparision technique ;speech recognition system design and implementation ;theory and implementation of hidden markov model; speech recognition based on connected word models; large vocabulary continuous speech recognition ;and task oriented application of automatic speech recognition .For practicing engineers, scientists, linguists, and programmers interested in speech recognition.

[2] This book revels the intence fact of the human speech communication ,mechanisms and models of human speech production ,mechanism and models of the human auditory system ;digital coding of speech ,message synthesis from stored human speech components, phonetic synthesis by rule; introduction to stochastic modeling ,practical technique for improving speech recognition template matching .It also emphasis the future research directions in speech synthesis and recognition ,application and performance of the current technology.

[3] In this book the term voice recognition or speaker identification refers to identifying the speaker, rather than what they are speaking .Recognizing the speaker can simplify the task of the translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of the speaker as a part of the security process. From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has beneficial from advances in deep learning and big data .The advances are evidenced not only by the surge of academic papers published in the field ,but more importantly by the worldwide industry adoption of the variety of deep learning methods in designing and developing speech recognition systems.

[4] This book is intending to provide readers with an overview of speech communication in its wide range aspects ,from a discussion of how humans produce and perceive speech to details of computer based speech processing for diverse communication ,at time satisfying some technical depth for breath, so that readers may see the relationship between parts of the communication process that are often dealt with separately .A cohesive ,even – handed discussion of speech production and perception (both human and machine).

[5] This book intends speech recognition has a long

history
Of being one of the difficult problem in artificial intelligent
and computer science .As one goes from problem solving task such as puzzles and sketch to perceptual task such as speech and vision, the problem characteristic change dramatically: knowledge poor to knowledge rich; low data rates to high data rates; slow response time (minutes to hour) to instantaneously response time. These characteristic taken together increase the computational complexity of the problem by several orders of magnitude.
[6] In this paper the author clearly revels speech based mini-device is designed to assist the interphone for operating on the shaft in coal mining .some keys with separate speech tip are formed based on the detailed task in the shaft .Order in work can be transmitted by the interphone. The order is send from one of the couple interphone to the other after a key is pressed .the communication between the workers on the different work plate is formed. The mini –devices achieve automation when its output signal is connected to the relays driving the equipment .The mini-device with intelligent function is of high practical value.

III. PRINCIPALS OF SPEAKER RECOGNITION

Speaker recognition can be classified into identification and verification .Speaker identification is the process of determining which registered speaker provides a given utterance .Speaker verification, on the other hand, it is the process of accepting or rejecting the identity speaker's identity is based on his or her speaking one or more specific phrases ,like passwords, cardnumbers , pin codes ,etc .The difference between text-dependent and text- independent is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text .All technologies of the speaker recognition ,identification and verification, text-independent and text-dependent ,each has its own advantages and disadvantages and may require different techniques and treatments. The choice of which technology to use is application specific .The system that we are going to develop comes under text independent speaker identification system since its task is to identify the person who speaks regardless of what is saying .At the highest level, all speaker recognition system contains two main modules :feature extraction and feature matching .Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker .Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from the set of known speakers . The task can also be divided into text-dependent and text-independent identification. The difference is that in the first case the system knows the text spoken by the person while in the second case the system must be able to recognize the speaker from any text. The identification taxonomy is shown in Fig

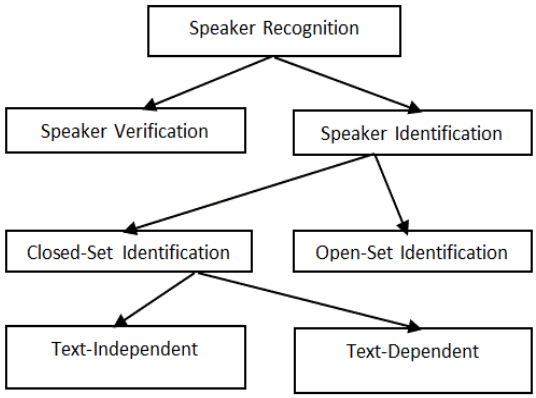

Fig 1: Identification taxonomy

The process of speaker identification is divided into two main phrases which is called speaker enrollment ,speech samples are collected from the speaker ,and they are used to train their models .The collection of enrolled models is also called a speaker database .In the second phase which is also known as speaker database .In the second phase which is also known as identification phases ,a test sample from an unknown speaker is compared against the speaker database. Both phases include the same first step, feature extraction, which is used to extract speaker dependent characteristic from speech .The main purpose of the step is to reduce the amount of test data while retaining speaker discriminative information .Then in the enrollment phase ,these features are modeled and stored in the speaker database.

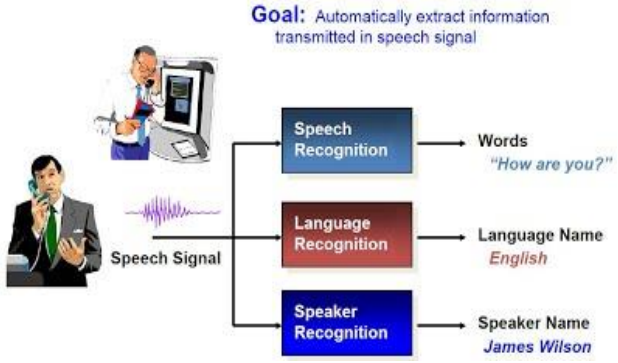

Fig 2: Extracting information from speech

All speaker recognition systems have to serve two distinguish phases. The first one referred to the enrollment session or training phase while the second one is referred to the operation session or testing phase. In the training phase ,each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker .In case of speaker verification systems, in addition, a speaker specific threshold is also computed from the training samples .During the testing phases ,the input speech is matched with stored reference model and recognition decision is made .Speaker recognition is a difficult task and it is an active research area. Our speaker recognition works based on the premise that a person's speech exhibits characteristics that are unique to the speaker. However this task has been challenged by the highly variant of input speech signals .The principle

source of variance is the speaker is the speaker himself .speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time ,health conditions (e, g. the speaker has a cold) , speaking rates, etc .There are also other factors , beyond speaker variability ,that present a challenge to speaker recognition technology .Example of these are acoustic noise and variations in recording environments (e.g. speaker uses different telephone handsets) [1][7].

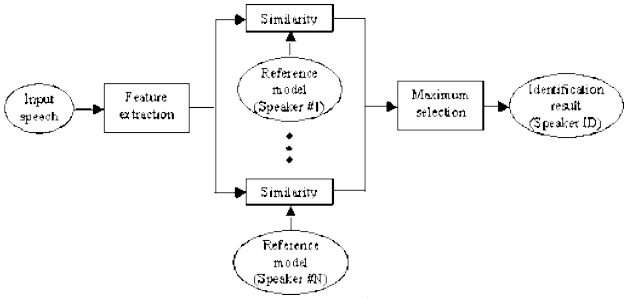

Fig 3: Speaker identification system

IV. FEATURE EXTRACTION

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for ,future analysis and processing .This is often referred as the signal -processing front end. The speech signal is a times varying signal (it is called quasi -stationary). An example of speech signal is shown in figure. When examined over a sufficiently short time (between 5 and common way to characterize the 100 msec), its characteristics are fairly stationary. However , overlong periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken .Therefore , short -time spectral analysis is the most speech signal MFCC TECHNIQUE is perhaps the best known and most popular, MFCC ' s are based on the known variation of the human ear's critical bandwidth with frequency, filters spaced linearly at low frequencies and logarithms at high frequencies have been used to capture the phonetically important characteristics of speech .This is expressed in the Mel- frequency scale, which is a linear frequency spacing below 1000 Hz and algorithms spacing above 1000 Hz . Automatic speech recognition (ASR) has made great strides with the development of digital signal processing hardware and software. But despite of all these advances, machines can not match the performance of their human counterparts in terms of accuracy and speed, specially in case of speaker independent speech recognition. So today significant portion of speech recognition research is focussed on speaker independent speech recognition problem. The reasons are its wide range of applications.

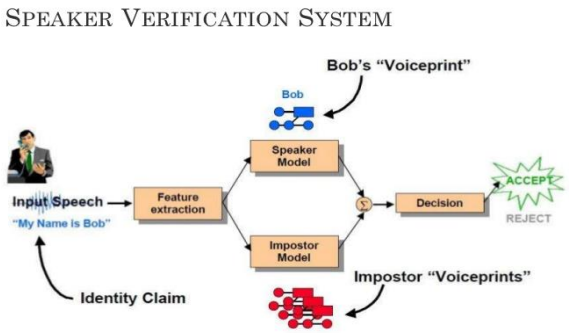

Fig 4: Speaker verification system

V. MEL-FREQUENCY CEPSTRUM COEFFICIENTS PROCESSOR

A block diagram of the structure of an MFCC processor is given in figure .The speech input is typically recorded at a sampling rate above 10000 Hz .This sampling frequency was chosen to minimize the effects of aliasing in the analog -to -digital conversation . These sampled signals can capture all frequencies up to 5 KHz which cover most energy of sounds that are generated by humans .As been discussed previously the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC 's are shown to be less susceptible to mentioned variations .

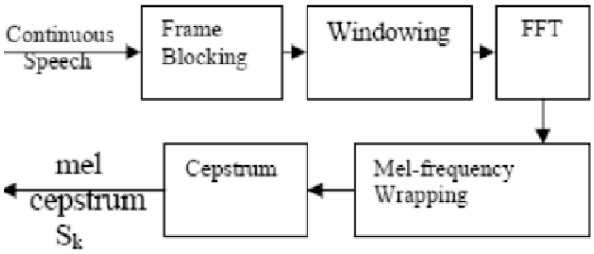

Fig 5: Block diagram of MFCC processor

a. FRAME BOCKING

In the step the continuous speech is blocked into frames of N samples with adjacent frames being separated by M(M<N). The first frame consists of the first N samples .The second frame begins M samples after the first frame ,and overlaps it by N-M samples. Similarly ,the third frame begins 2M samples after the first frame or M samples after the second frame and overlaps it by N-2M samples.This process continuous until all the speech is accounted for within one or more frames .Typical values for N and M are N=256 (which is equivalent to ~30 msec windowing and facilitate the fast radix-2 FFT)and M =100.

b WINDOWING

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities
at the beginning and end of each frame .The concept here is to minimize the spectral distortion by using the window to

taper the signal to zero at the beginning and end of each frame .If we define the window as W(n), 0<= n <= N-1 where N is the number of samples in each frame ,then the result of windowing is the signal Y(n)= X(n)W(n), 0<= n <= N-1.

c FAST FOURIER TRASFORM

The next processing step is the Fast Fourier Transform ,which converts each frame of N samples from the time domain into the frequency domain.The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples {xn} in general X n 'a are complex numbers .The resulting sequence {Xn} is interpreted as follows :the zero frequency corresponds to n=0,positive frequencies 0< f< Fs/2 corresponds to values 1<= n <=N/2 -1 while negative frequencies -Fs/2 < f < 0 corresponds to N/2 +1 <= N-1 .Here ,Fs denotes the sampling frequency. The result after this step is often referred to as spectrum or periodogram.

d MEL –FREQUENCY WRAPPING

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale .Thus for each tone with an actual frequency, f ,measured in Hz , a subjective pitch is measured on a scale called the 'mel' scale .The mel - frequency scale is a linear frequency spacing below 1000 Hz and a logarithms spacing above 1000 Hz .As a reference point ,the pitch of a 1 KHz tone ,40 db above the perceptual hearing threshold ,is defined as 1000 mels .Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz : mel(f) = 2595* log (1+ f/ 700**)**

e. CEPSTRUM

In this final step, we convert the log mel spectrum back to time .The result is called the mel frequency cepstrum coefficient (MFCC). The cepstrum representation of the speech spectrum provides a good representation of the local spectrum properties of the signal for the given frame analysis .Because the mel spectrum coefficient (and so their logarithm) are real numbers ,we can convert them to the time domain using the Discrete cosine Transform (DCT). Therefore if we denote those mel power spectrum co efficients that are the result of the step are Sk, k= 1,2,3,------ ,k .

VI. FEATURE MATCHING TECHNIQUE

Figure 6 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown .The circles refer to the acoustic vectors from the speaker 1 while the triangle are from the speaker 2 .In the training phase, a speaker - specific VQ code book is generated for each known speaker by clustering his/her training acoustic vectors .The result code words (centroids) are shown in Figure by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ- distortion. In the recognition phase, an input utterance of.

an unknown voice is vector -quantized using each trained codebook and the total VQ distortion is computed .The speaker corresponding to the VQ codebook with smallest total distortion is identified

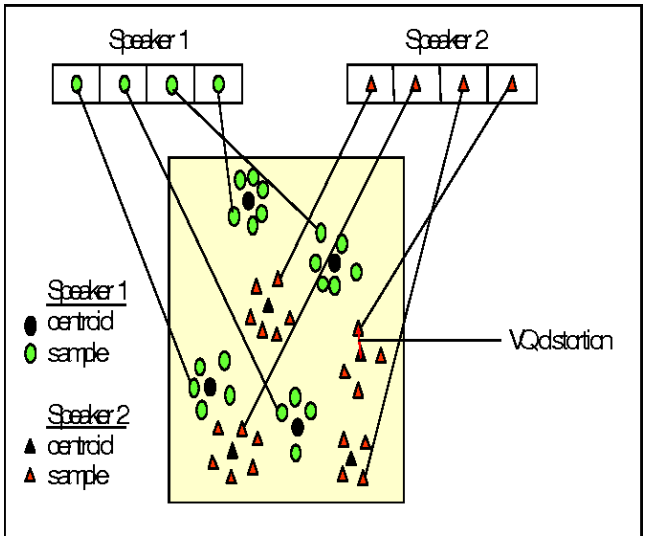

Fig 6. Conceptual diagram illustrating vector quantization codebook formation one speaker can be described from another based on the location of centroid

VII. CONCLUSION

Digital signal processing is one of the advancement in the field of electronic and communication engineering which has lead to several critical and intelligent application .An attempt has been made here to recognize a person based on his speech .A speech may be processed to recognize a particular person.

REFERENCES

[1] "Fundamentals of speech recognition", by Lawrence and Biing-Hwang Juang,1993,ISBN 0-13-015157-2

[2] “Speech synthesis and recognition” by J.N.Holmes Wokingham,1988

[3] "Electronic speech recognition; techniques, technology and application", by Geoff Bristow,1986

[4] “Speech communication: Human and machine”, by Douglas,1987

[5] “Automatic Speech Recognition: The development of the SPHINX system”,by Boston and Kal-Fu,1989

[6] “Speech signal processing”,by H.hu,Harbin university of industry press,2000.

[7] “Signal Conditioning of Audio Signal for Aircraft”, International Conference on Emerging Trends in Science and Engineering (ICETSE-2017) at Coorg Institute of Technology held 11th &12th May 2017.

[8] Prabu, S., V. Balamurugan, and K. Vengatesan. "Design of cognitive image filters for suppression of noise level in medical images." *Measurement* 141 (2019): 296-301.