A Review on Speech Recognition Technique

Shweta Tiwari[1], Manish Saxena [2]

Electronics and communication,RGPV(Bhopal)India.

Bansal institute of science and technology

1. Shweta Tiwari Student,Mtech(Digital Communication), Bansal Institute Of Science and Technology Bhopal. email-shweta.tiwari86@gmail.com Mobile 7828982256
2. Manish Saxena, M.Tech Coordinator, Bansal Institute Of Science and Technology Bhopal, Email:-manish.saxena2008@gmail.com, Mobile: +919826526247.

ABSTRACT: This paper we study of design a system to recognition voice commands. Most of voice recognition systems contain two main modules as follow "feature extraction" and "feature matching". In this project, MFCC algorithm is used to simulate feature extraction module. Using this algorithm, the cepstral coefficients are calculated on mel frequency scale. VQ (vector quantization) method will be used for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance is applied as similarity criterion. Because of high accuracy of used algorithms, the accuracy of this voice command system is high. Using these algorithms, by at least 5 times repetition for each command, in a single training session, and then twice in each testing session zero error rate in recognition of commands is achieved. Digital processing of speech signal and voice recognition algorithm is very important for fast and accurate automatic voice or speaker recognition technology. Speaker Recognition is a process of automatically recognizing who is speaking on the basis of the individual information included in speech waves. Speaker Recognition is one of the most useful biometric recognition techniques in this world where insecurity is a major threat. Many organizations like banks, institutions, industries etc are currently using this technology for providing greater security to their vast databases.

Keywords : Feature Extraction, Feature Matching, Mel Frequency Cepstral Coefficient (MFCC), dynamic Time Warping (DTW)

I. INTRODUCTION

Speaker recognition is the process of identifying a person on the basis of speech alone. It is a known fact that speech is a speaker dependent feature that enables us to recognize friends over the phone. During the years ahead, it is hoped that speaker recognition will make it possible to verify the identity of persons accessing systems; allow automated control of services by voice, such as banking transactions; and also control the flow of private and confidential data. While fingerprints and retinal scans are more reliable means of identification, speech can be seen as a non-evasive biometric that can be collected with or without the person's knowledge or even transmitted over long distances via telephone. Unlike other forms of identification, such as passwords or keys, a person's voice cannot be stolen, forgotten or lost. Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals.

In speaker recognition, all these differences can be used to discriminate between speakers. Speaker recognition allows for a secure method of authenticating speakers. During the enrollment phase, the speaker recognition system generates a speaker model based on the speaker's characteristics. The testing phase of the system involves making a claim on the identity of an unknown speaker using both the trained models and the characteristics of the given speech. Many speaker recognition systems exist and the following chapter will attempt to classify different types of speaker recognition systems.

II. SPEAKER RECOGNITION

Speaker identification [6] is one of the two categories of speaker recognition, with speaker verification being the other one. The main difference between the two categories will now be explained. Speaker verification performs a binary decision consisting of determining whether the person speaking is in fact the person he/she claims to be or in other words verifying their identity. Speaker identification performs multiple decisions and consists comparing the voice of the person speaking to a database of reference templates in an attempt to identify the speaker. Speaker identification will be the focus of the research in this case.

Speaker identification further divides into two subcategories, which are text dependent and text-independent speaker identification [10]. Text-dependent speaker identification differs from text-independent because in the aforementioned the identification is performed on a voiced instance of a specific word, whereas in the latter the speaker can say anything. The thesis will consider only the text-dependent speaker identification category.

The field of speaker recognition has been growing in popularity for various applications. Embedding recognition in a product allows a unique level of hands-free and intuitive user interaction. Popular applications include automated dictation and command interfaces. The various phases of the project lead to an in-depth understanding of the theory and implementation issues of speaker recognition, while becoming more involved with the speaker recognition community. Speaker recognition uses the technology of biometrics.

Biometric techniques based on intrinsic characteristics (such as voice, finger prints, retinal patterns) [17] have an advantage over artifacts for identification (keys, cards, passwords) because biometric attributes cannot be lost or forgotten as these are based on his/her physiological or behavioral characteristics. Biometric techniques are generally believed to offer a reliable method of identification, since all people are physically different to some degree. This does not include any passwords or PIN numbers which are likely to be forgotten or forged. Various types of biometric systems are in vogue. A biometric system is essentially a pattern recognition system, which makes a personal identification by determining the authenticity of a specific physiological or behavioral characteristics possessed by the user. An important issue in designing a practical system is to determine how an individual is identified. A biometric system can be either an identification system or a verification system. Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Automatic speaker identification and verification are often considered to be the most natural and economical methods for avoiding unauthorized access to physical locations or computer systems. Thanks to the low cost of microphones and the universal telephone network, the only cost for a speaker recognition system may be the software. The problem of speaker recognition is one that is rooted in the study of the speech signal. A very interesting problem is the analysis of the speech signal, and therein what characteristics make it unique among other signals and what makes one speech signal different from another.

When an individual recognizes the voice of someone familiar, he/she is able to match the speaker's name to his/her voice. This process is called speaker identification, and we do it all the time. Speaker identification exists in the realm of speaker recognition, which encompasses both identification and verification of speakers.

Automatic speaker recognition systems can be divided into two classes depending on their desired function; Automatic Speaker Identification (ASI) classification of

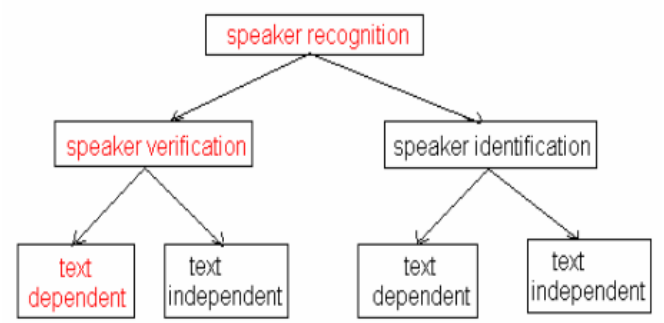

Fig 1: The Scope of Speaker Recognition

Speaker recognition methods can be divided into text independent and text dependent
methods. In a text independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. This system should be intelligent enough to capture the characteristics of all the words that the speaker can use. On the other hand in a text dependent system, the recognition of the speaker's identity is based on his/her speaking one or more specific phrases like passwords, card numbers, PIN codes etc.

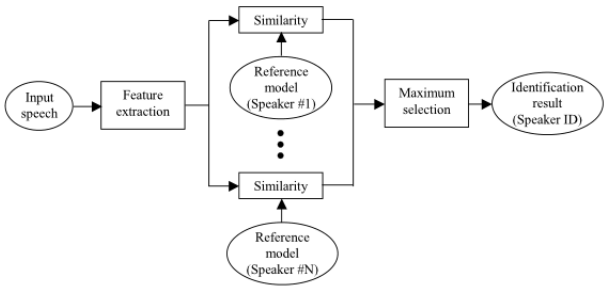

Fig. 2 Basic structure of speaker recognition systems: Speaker identification

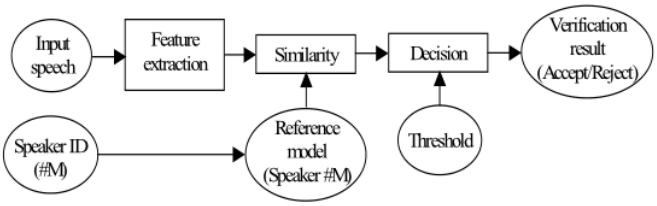

Fig .3 Basic structure of speaker recognition systems: Speaker verification

The vocabulary of digit is used very often in testing speaker recognition because of its applicability to many security applications. For example, users have to speak a PIN (Personal Identification Number) in order to gain access to the laboratory door, or users have to speak their credit card number over the telephone line. By checking the voice characteristics of the input utterance using an automatic speaker recognition system

similar to the one I will develop, the system is able to add an extra level of security.
In feature matching the input utterance of an unknown speaker is converted into MFCCs and then the total VQ distortion between these MFCCs and the codebooks stored in our database is measured. VQ distortion is the distance from a vector to the closest code word of a codebook. Based on this VQ distortion we decide whether the speaker is a valid person or an impostor. i.e. if the VQ distortion is less than the threshold value then the speaker is a valid person and if it exceeds the threshold value then he is considered as an impostor. This system is at its best roughly 80% accurate in identifying the correct speaker.

III. SPEECH FEATURE EXTRACTION

The purpose of this module is to convert the speech waveform to some type of parametric representation for further analysis and processing. This is often referred to as the signal-processing front end. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and this is used in this paper.

3.1 MFCC Computation

A block diagram of the structure of an MFCC processor is as shown (Fig. 7). The main purpose of the MFCC processor is to mimic the behavior of the human ears.
In first step, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). Typical values for N and M are N = 256 and M = 100.
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. Typically the Hamming window is used.

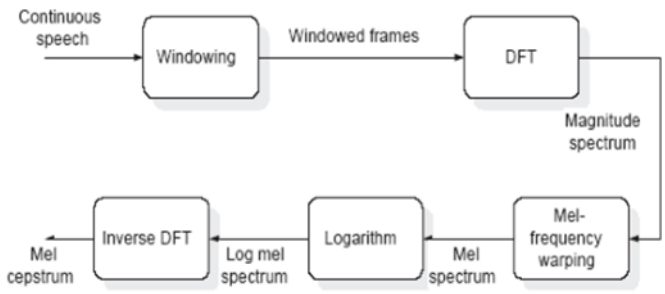

Fig.4 MFCC calculation

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. After that the scale of frequency is converted from linear to mel scale. Then logarithm is taken from the results. In final step, the log mel spectrum is converted back to time domain. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech cepstrum provides a good representation of the local spectral properties of the signal.
Using triangular filter bank, we obtain significant decrease in amount of data. But for more simplicity in next computations, more decreasing in amount of data is needed. For this purpose vector quantization algorithm is used [5].

3.2 Vector Quantization

Vector quantization (VQ) is used for command identification in our system. VQ is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a cluster and is represented by its center (called a centroid). A collection of all the centroids make up a codebook. The amount of data is significantly less, since the number of centroids is at least ten times smaller than the number of vectors in the original sample. This will reduce the amount of computations needed when comparing in later stages [2],[4]. Even though the codebook is smaller than the original sample, it still accurately represents command characteristics. The only difference is that there will be some spectral distortion.

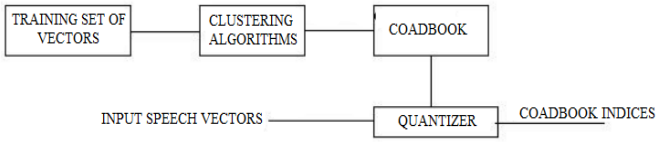

Fig 5. Block Diagram of the basic VQ Training and classification structure

Figure 5 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 triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors.

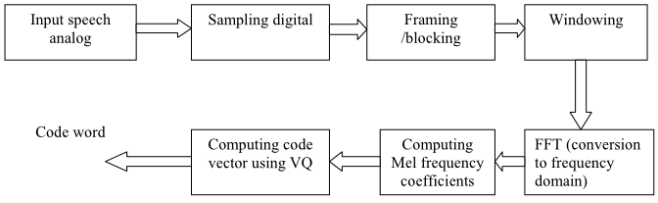

Fig 6. Steps for speaker recognition implementation.

Speaker recognition systems contain two main modules: feature extraction and classification.

1. Feature extraction is a process that extracts a small amount of data from the voice signal that can be used to represent each speaker. This module converts a speech waveform to some type of parametric representation for further analysis and processing. Short-time spectral analysis is the most common way to characterize a speech signal. The Mel-frequency Cepstrum coefficients (MFCC) are used to parametrically represent the speech signal for the speaker recognition task. The steps in this process are shown in Figure 6

(a) Block the speech signal into frames, each consisting of a fixed number of samples.

(b) Window each frame to minimize the signal discontinuities at the beginning and end of the frame.

(c) Use FFT to convert each frame from time to frequency domain.

(d) Convert the resulting spectrum into a Mel-frequency scale.

(e) Convert the Mel spectrum back to the time domain.

Classification consists of models for each speaker and decision logic necessary to render a decision. This module classifies extracted features according to the individual speakers whose voices have been stored. The recorded voice patterns of the speakers are used to derive a classification algorithm. Vector quantization (VQ) is used. This is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center, called a codeword. The collection of all clusters is a codebook. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. 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 the smallest total distortion is identified.

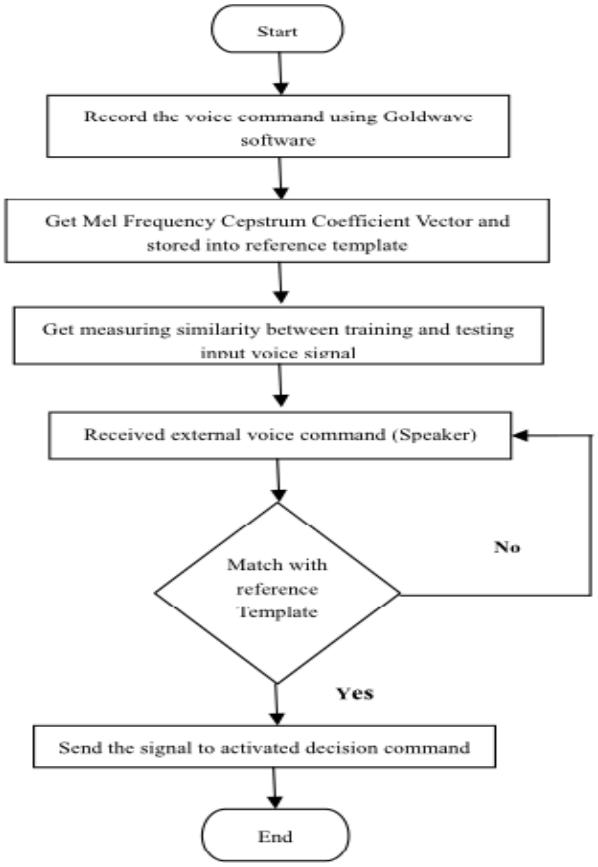

Fig 7. Voice Algorithm flow Chart

As mentioned in , voice recognition work on the premise that a person voice exhibits characteristic are unique to different speaker. The signal during training and testing session can be greatly different due to many factor such as people voice change with time, health condition (the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone. Figure 6 show the flowchart for overall voice recognition process.

IV. EXPECTS EXPERIMENTAL RESULT

Each speaker utters the same single digit, zero, once in a training session (then also in a testing session). A digit is often used for testing in speaker recognition systems because of its applicability to many security applications. This project was implemented on the C6711 DSK and can be transported to the C6713 DSK. Of the eight speakers, the system identified six correctly (a 75% identification rate). The identification rate can be improved by adding more vectors to the training code words. The performance of the system may be improved by

using two-dimensional or four dimensional VQ by changing the quantization method to dynamic time wrapping.
The results obtained in this paper using MFCC and VQ are applauding able. I will compute MFCCs corresponding to each speaker and these are vector quantized. The VQ distortion between the resultant codebook and MFCCs of an unknown speaker is taken as the basis for determining the speaker's authenticity. Here I used MFCCs because they follow the human ear's response to the sound signals.

V. CONCLUSION

The performance of this model is limited by a single coefficient having a very large VQ distortion with the corresponding codebook. The performance factor can be optimized by using high quality audio devices in a noise free environment. There is a possibility that the speech can be recorded and can be used in place of the original speaker .This would not be a problem in our case because the MFCCs of the original speech signal and the recorded signal are different. Psychophysical studies have shown that there is a probability that human speech may vary over a period of 2-3 years. So the 0training sessions have to be repeated so as to update the speaker specific codebooks in the database.

REFERENCES

1. Campbell, J.P., Jr.; "Speaker recognition: a tutorial" Proceedings of the IEEE Volume 85, Issue 9, Sept. 1997 Page(s):1437 – 1462.
2. Seddik, H.; Rahmouni, A.; Sayadi, M.; "Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier" First International Symposium on Control, Communications and Signal Processing, Proceedings of IEEE 2004 Page(s):631 – 634.
3. Childers, D.G.; Skinner, D.P.; Kemerait, R.C.; "The cepstrum: A guide to processing" Proceedings of the IEEE Volume 65, Issue 10, Oct. 1977 Page(s):1428 – 1443.
4. Roucos, S. Berouti, M. Bolt, Beranek and Newman, Inc., Cambridge, MA; "The application of probability density estimation to text-independent speaker identification" IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '82. Volume: 7, On page(s): 1649- 1652. Publication Date: May 1982.
5. Castellano, P.J.; Slomka, S.; Sridharan, S.; "Telephone based speaker recognition using multiple binary classifier and Gaussian mixture models" IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 Volume 2, Page(s) :1075 – 1078 April 1997.
6. Zilovic, M.S.; Ramachandran, R.P.; Mammone, R.J "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions".; IEEE Transactions on Speech and Audio Processing, Volume 6, May 1998 Page(s):260 - 267
7. Davis, S.; Mermelstein, P, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences" , IEEE Transactions on Acoustics, Speech, and Signal Processing Volume 28, Issue 4, Aug 1980 Page(s):357 – 366
8. Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, issue 1, Jan 1980 pp.84-95.
9. S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol.34, issue 1, Feb 1986, pp. 52-59.
10. Fu Zhonghua; Zhao Rongchun; "An overview of modeling technology of speaker recognition", IEEE Proceedings of the International Conference on Neural Networks and Signal Processing Volume 2, Page(s):887 – 891, Dec. 2003.
11. Moureaux, J.M., Gauthier P, Barlaud, M and Bellemain P."Vector quantization of raw SAR data", IEEE International Conference on Acoustics, Speech, and Signal Processing Volume 5, Page(s):189 - 192, April 1994.
12. Nakai, M.; Shimodaira, H.; Kimura, M.; "A fast VQ codebook design algorithm for a large number of data", IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, Page(s):109 – 112, March 1992.
13. Xiaolin Wu; Lian Guan; "Acceleration of the LBG algorithm" IEEE Transactions on Communications, Volume 42, Issue 234, Part 3Page(s):1518 - 1523, February/March/April 1994.
14. B. P. Bogert, M. J. R. Healy, and J. W. Tukey: "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking".Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed) Chapter 15, 209-243. New York: Wiley, 1963.
15. Martin, A. and Przybocki, M., "The NIST Speaker Recognition Evaluations: 1996-2000", Proc. OdysseyWorkshop, Crete, June 2001
16. Martin, A. and Przybocki, M., "The NIST 1999 Speaker Recognition Evaluation—An Overview", Digital Signal Processing, Vol. 10, Num. 1-3. January/April/July 2000
17. Claudio Becchetti and Lucio Prina Ricotti, "Speech Recognition", Chichester: John Wiley & Sons, 2004.
18. John G. Proakis and Dimitris G. Manolakis, "Digital Signal Processing", New Delhi: Prentice Hall of India. 2002.
19. Rudra Pratap. Getting Started with MATLAB 7. New Delhi: Oxford University Press, 2006
20. R. Chassaing, DSP Applications Using C and the TMS320C6x DSK, Wiley, New York, 2002.