Speaker Recognition System – A Review

Agrani Vishwakarma
Master of Technology (Digital Communication) Babulal Tarabai Institute of Research and Technology (BTIRT), Sagar (M.P.) India.

Megha Soni
Assistant Professor, Babulal Tarabai Institute of Research and Technology (BTIRT), Sagar (M.P.) India.

Abstract – Speech processing is emerged as one of the important application area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. Speaker recognition has been an interesting research field for the last many decades, which still have a number of unsolved problems. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. A direct analysis and synthesizing the complex voice signal is due to too much information contained in the signal. Therefore the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal. Several methods such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN) etc are evaluated with a view to identify a straight forward and effective method for speech signal. In this paper, the Mel Frequency Cepstrum Coefficient (MFCC) technique is been explained for designing a speaker recognition system.

Index Terms – Mel-Frequency Cepstral Coefficients, Speech Signal, Vector Quantization, Speaker Recognition System, Windowing.

1. INTRODUCTION

The speech signal has enormous capacity of carrying information. Speaker recognition, such as speaker identification and speaker verification is based on the fact that one's speech reflects his/her unique characteristics. Speech signal can be seen as a non-evasive biometric that can be collected with or without the persons 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. Speaker recognition allows for a secure method of authenticating speakers.

While many other biometric systems like fingerprint recognition, retinal scans, face recognition etc. are more reliable means of identification and are now a day's used in various security and access control system. In future, it is hoped that speaker recognition will make 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.

Research and development on speaker recognition methods and techniques has been undertaken for well over six decades and still it continues to be an active area. During this period of six decades various feature extraction and feature matching methods are introduced to represent the voice signal. The non-parametric method for modeling the human auditory perception system, Mel Frequency Cepstral Coefficients (MFCCs) is discussed in this paper as feature extraction technique and VQ as feature matching technique.

2. WORK ALREADY DONE IN THE FIELD

Although in last six decades many techniques have been developed, many challenges have yet to be overcome before we can achieve the ultimate goal of creating machines that can communicate naturally with people. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, research in automatic speech and speaker recognition by machines has attracted a great deal of attention for six decades.

In the 1950's, Bell Labs created a recognition system capable of identifying the spoken digits 0 through 9 was conducted by Biddulph and Balashek, while later systems were able to recognize vowels. Major attempts for automatic speaker recognition were made in the 1960s, one decade later than that for automatic speech recognition. Pruzansky at Bell Labs was among the first to initiate research by using filter banks and correlating two digital spectrograms for a similarity measure. Pruzansky and Mathews improved upon this technique and further developed it by using linear discriminators.

Growth of speaker recognition system during the last six decades is as follows:

1950's –1960's

1. Use of Spectral Resonance.
2. Use of filter bank and logic circuits.
3. Speaker recognition system were Limited in extent.
4. Replacement of filter banks by formant analysis.

1970's

1. Isolated word and discrete utterance recognition.
2. Pattern recognition idea.
3. LPC Spectral parameters.
4. Large vocabulary but simple queries.
5. First speech recognition commercial company called Threshold Technology, Inc. and developed the first real ASR product called the VIP-100 System

1980's

1. Recognition spoken strings.
2. Various matching algorithms were developed.
3. HMM, Polynomial coefficients.
4. Artificial neural networks (ANN) approach was developed.
5. Development of software tools that enabled many individual research programs all over the world.

1990's

1. Error minimization concet.
2. Description training and kernel-based methods.
3. Maximum Mutual Information (MMI) training.
4. Wide application: Telephone system.

2000's

1. Speech-to-Text (Automatic Transcription technology).
2. Spontaneous speech recognition.
3. Various other techniques were developed and improvements over some earlier techniques were proposed.

3. SPEAKER RECOGNITION SYSTEM

At the highest level, all speaker recognition systems contain 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 a set of known speakers.

All speaker recognition systems have to serve two distinguishes phases. The first one is referred to the enrollment sessions or training phase while the second one is referred to as the operation sessions 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 (operational) phase, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using some matching technique. The level of match is used to arrive at the decision.

A. Feature Extraction (MFCC)

The purpose of this module is to convert the speech waveform to some type of parametric representation for further analysis and processing. The speech signal is called quasi-stationary because it is a slowly timed varying signal. When examined over a sufficiently short period of time (Between 10 and 50 ms), its characteristics are fairly stationary. However, over long periods of time (On the order of 1/4 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. 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 will be discussed in this paper, because MFCC is perhaps the best known and most popular and also, it shows high accuracy results for clean speech and also experiments show that the parameterization of the Mel frequency cepstral coefficients is best for discriminating speakers and is different from the one usually used for speech recognition applications**.**

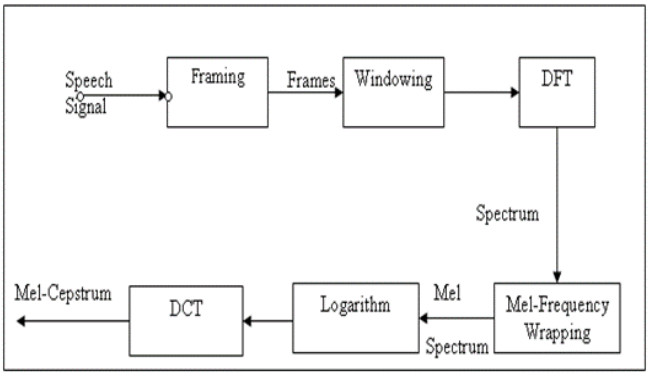

Fig. 1 MFCC Block Diagram.

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1KHz. In other words, MFCC is

based on known variation of the human ear's critical bandwidth with frequency. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. A block diagram of the structure of an MFCC processor is given in Figure 1.

1) Pre-Emphasis:

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing includes digital filtering and speech signal detection. Filtering includes pre-emphasis filter and filtering out any surrounding noise using several algorithms of digital filtering. Pre-emphasis refers to a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of the others(usually lower) frequencies in order to improve the overall SNR. Hence, this step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

2) Framing:

In this step the continuous speech signal 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 continues until all the speech is accounted for within one or more frames.. The reason for this overlapping is that on each individual frame we will also be applying a hamming window which will get rid of some of the information at the beginning and end of each frame. Overlapping will then reincorporate this information back into our extracted features Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing) and M = 100.

3) Windowing:

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. The multiplication of the speech wave by the window function has two effects:-

a. It gradually attenuates the amplitude at both ends of extraction interval to prevent an abrupt change at the endpoints.

b. It produces the convolution for the Fourier transform of the window function and the speech spectrum.

The choice of the window depends on several factors. In speaker recognition, the most commonly used window shape is the hamming window

The hamming window is defined as:-

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

The use for hamming windows is due to the fact that mfcc will be used which involves the frequency domain so windows will decrease the possibility of high frequency components in each frame due to such abrupt slicing of the signal.

4) Fourier Transform:

The next processing step is the Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. We usually perform FT to obtain the magnitude frequency response of each frame. When we perform DFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FT but the discontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies:

a. Multiply each frame by a Hamming window to increase its continuity at the first and last points.

b. Take a frame of a variable size such that it always contains an integer multiple number of the fundamental periods of the speech signal.

The second strategy encounters difficulty in practice since the identification of the fundamental period is not a trivial problem. Moreover, unvoiced sounds do not have a fundamental period at all. Consequently, we usually adopt the first strategy to multiply the frame by a Hamming window before performing FT.

5) Mel-Frequency Wrapping:

Human perception of the frequency contents of sounds for speech signals does not follow a linear scale. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.The formula to compute the mels for a given frequency f in Hz is:

Mel (f) = 2595*log10 (1+f/700)

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth

is determined by a constant Mel-frequency interval. The modified spectrum thus consists of the output power of these filters. The number of Mel spectrum coefficients, K, is typically chosen as 20.

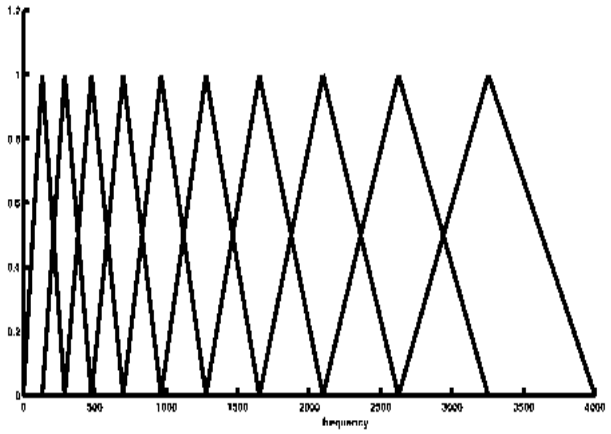

Fig. 2 Mel Scale Filter bank

6) Cepstrum:

In the final step, the log Mel spectrum has to be converted back to time. The result is called the Mel frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients are real numbers and so are their logarithms, they may be converted to the time domain using the Discrete Cosine Transform (DCT). It is known that the logarithm has the effect of changing multiplication into addition. Therefore we can simply convert the multiplication of the magnitude of the Fourier transform into addition. Then by taking the DCT of the logarithm of the magnitude spectrum, MFCC can be calculated.

B. FEATURE MATCHING

The feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this paper, the VQ approach is used, due to ease of implementation and high accuracy. Vector quantization, also called "block quantization" or "pattern matching quantization" is often used in lossy data compression. VQ works by dividing a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point.

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. In the figure, only two speakers and two dimensions of the acoustic space are shown.

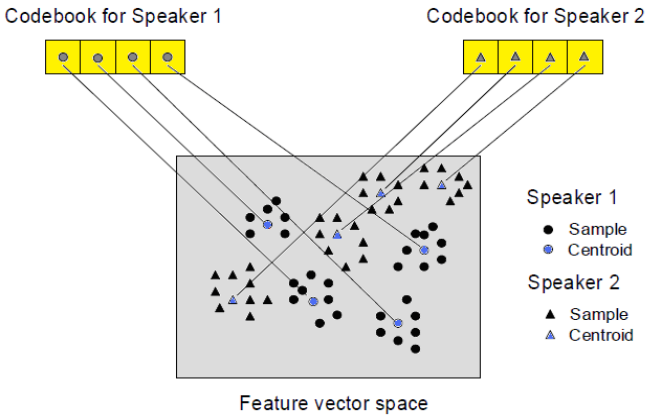

Fig. 3 Diagram illustrating vector quantization codebook formation

After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. The next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray], for clustering a set of L training vectors into a set of M codebook vectors. Linde, Buzo and Gray proposed an improvement of the Lloyd's technique. They extended Lloyd's results from mono- to k-dimensional cases. For this reason their algorithm is known as the Generalized Lloyd Algorithm (GLA) or LBG from the initials of its authors. In a few words, the LBG algorithm is a finite sequence of steps in which, at every step, a new quantizer, with a total distortion less or equal to the previous one, is produced.

The algorithm is formally implemented by the following recursive procedure: Find the centroid of the entire set of training vectors this is one vector codebook. Double the size of the codebook by splitting each current codebook y_n according to the rule

$$y_n^+ = y_n(1+\varepsilon)$$

$$y_n^- = y_n(1-\varepsilon)$$

Where n varies from 1 to the current size of the codebook, and $\in$ is a splitting parameter (we choose $\in$ =0.01). For each training vector, find the codeword in the current codebook that is closest and assign that vector to the corresponding cell (associated with the closest codeword). Update the codeword in each cell using the centroid of the training vectors assigned to that cell. Go on splitting and updating the centroids until a codebook size of M is designed.

Figure shows, in a flow diagram, the detailed steps of the LBG algorithm. "Cluster vectors" is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. "Find centroids" is the

centroid update procedure. "Compute D (distortion)" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged.

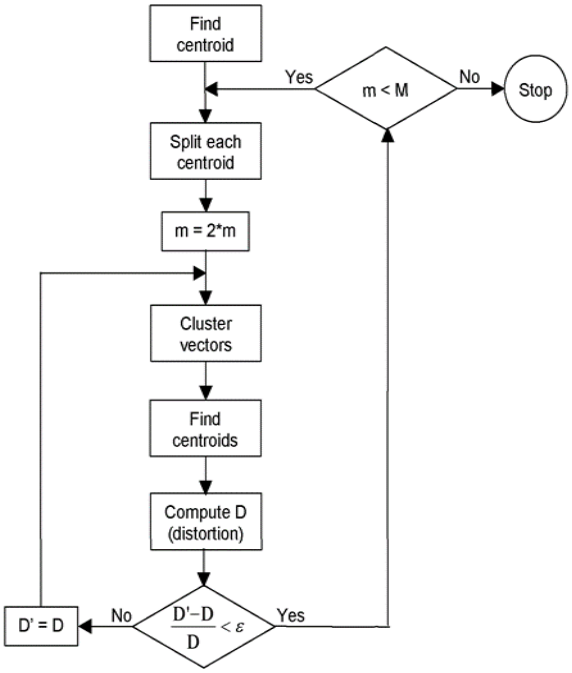

Fig. 4 LBG algorithm

4. CONCLUSION

From the study it can be concluded that by using Mel-Frequency Cepstral Coefficients technique for feature extraction and LBG-VQ matching technique for feature matching the speaker recognition system can be made more robust and efficient and hence the performance of Speaker recognition system can be improved.

REFERENCES

[1] Balsam Z. Khojah, Wadee S. AIhalabi – "*Text-Independent Speaker Identification System Using Different Pattern Matching Algorithms",* International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-1, Jan.-2017.

[2] N. Praveen and T. Thomas, "Text dependent speaker recognition using mfcc features and bpann," International Journal of Computer Applications, vol. 74, no. 5,pp. 31–39, 2013.

[3] Anjali Bala, Abhijeet Kumar, Nidhika Birla – "*Voice Command Recgnition using system based on MFCC and DTW"*, International journal of engineering science and technology vol.2 (12), 2010.

[4] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi- *"Voice Recognition using MFCC and DTW techniques"*, journal of computing vol.2, issue 3, march 2010.

[5] Shi-Huangchen and Yoy-Ren Luo - "*Speaker verification using MFCC and support vector machine"*, proceedings of the international multiconference of engineers and computer scientist, vol 1, 2009.

[6] Marshalleno Skosan and Daniel Mashao - "An Overview of speaker recognition technology", chi-sa, 2005.

[7] S.M. Ahadi, R.L. Brennan and G.H. Freeman - *"An efficient front-end for automatic speech recognition"*- Dec, 14-17, IEEE, 2003.

[8] Douglas A. Reynolds –"An Overview of automatic speaker recognition technology", IEEE, 2002.

[9] Sodaoki Furui – "50 years of progress in speech and speaker recognition".

[10] Jaemoon Lee and Mignon Park - "Design of the robotic system for human robot interaction using sound source localization, mapping data and voice recognition".

[11] Keiichi Tokuda - "Speech coding based on adaptive mel-cepstral analysis", IEEE, 1994.

[12] L R.Rabiner, R.W. Schafer, "*Digital Processing of Speech Signals"*, Prentice Hall, 1978.

[13] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani, and Md. Saifur Rahman, "*Speaker Identification Using Mel-Frequency Cepstral Coefficients*", ICECE 2004, 28-30 December 2004 Dhaka, Bangladesh.

[14] Wang Chen, Miao Zhenjiang and Meng Xiao, *"Comparison of different implementations of mfcc,"* J. Computer Science & Technology, 2001, pp. 16(16): 582-589.

[15] J.R. Deller, J.L.H Hansen & J.G. Proakis, (1993), "*Discrete-Time Processing of Speech Signal",* New York, Macmillan Publishing Company.

[16] J.G. Prokis and D.G. Monalakis, *Digital Signal Processing,* Third Edition, PHI,New Delhi, 2003.

[17] Li Tan and Montri Karnjanadecha "Modified Mel-Frequency Cepstral Coefficients" Thailand.

[18] Ashish Jain, Hohn Harris "*Speaker identification using MFCC and HMM based techniques"*, university Of Florida, April 25, 2004.

[19] Shi-Huang Chen and Yu-Ren Luo *"Speaker Verification Using MFCC and Support Vector Machine"* Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18 - 20, 2009, Hong Kong.

[20] Santosh K.Gaikwad, Bharti W.Gawali, Pravin Yannawar *"A Review on Speech Recognition Technique"* International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010.

[21] Ganesh K Venayagamoorthy, Viresh Moonasar and Kumbes Sandrasegaran *"Voice Recognition Using Neural Networks"* Electronics Engineering Department, M L Sultan Technikon, Durban, South Africa, IEEE, 1998.

[22] Douglas A. Reynolds "Automatic Speaker Recognition: Current Approaches and Future Trends" MIT Lincoln Laboratory, Lexington, MA USA.

[23] Joseph P. Campbell "*Speaker Recognition: A Tutorial"* Proceedings of IEEE, VOL. 85, NO. 9, September 1997

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

[25] Z. Bin, W. Xihong, C. Huisheng, *"On the Importance of Components of the MFCC in Speech and Speaker Recognition",* Center for Information Science, Peking University, China, 2001.

[26] P. Fränti, T. Kaukoranta, O. Nevalainen, *"On the Splitting Method for Vector Quantization Codebook Generation", Optical Engineering*, 36 (11), pp. 3043- 3051, November 1997.

[27] P. Fränti, J. Kivijärvi, *"Randomized Local Search Algorithm for the Clustering Problem"*, Pattern Analysis and Applications, 3 (4), 358-369, 2000.

[28] H. Gish and M. Schmidt, *"Text Independent Speaker Identification*", IEEE Signal Processing Magazine, Vol. 11, No. 4, 1994, pp. 18-32.

[29] T. Kinnunen, T. Kilpeläinen, P. Fränti, *"Comparison of Clustering Algorithms in Speaker Identification"*, Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000), pp. 222-227, Marbella, Spain, 2000.

[30] V. Mantha, R. Duncan, Y. Wu, J. Zhao, A. Ganapathiraju, J. Picone, *"Implementation and Analysis of Speech Recognition Front-Ends"*, Southeastcon '99. Proceedings. IEEE, 1999, pp. 32- 35.

[31] F.K. Song, A.E. Rosenberg and B.H. Juang, *"A vector quantisation approach to speaker recognition",* AT&T Technical Journal, Vol. 66-2, pp. 14-26, March 1987.