Mel Frequency Cepstral Coefficients Based Text Independent Automatic Speaker Recognition Using Matlab

Amit Kumar Singh, Rohit Singh,Ashutosh Dwivedi
Department of Electrical Engineering, School of Engineering
Shiv Nadar University
Gautam Budh Nagar, India
amitkumarsingh89@gmail.com, rohitsingh@asia.com, ashutosh.dwivedi@snu.edu.in

Abstract—- Speech feature extraction is the most significant step in any Automatic speaker recognition system. In the last 60 years a lot of research has gone into parametric representation of these speech features. Several techniques are currently being used for Automatic Speaker Recognition. Yet Automatic Speaker Recognition still remains a confront mainly due to variations in speaker's vocal tract with time and health, varying environmental conditions, disparities in the behavior and quality of speech recorders etc. MFCC is a extensively used technique in Automatic speaker recognition. In this paper the performance of MFCC technique was evaluated in a quiet environment. A speaker database containing 30 male and 30 female speakers was created. Two separate experiments were conducted for the performance evaluation of MFCC technique when applied to K means clustering. In the first case the speech features were directly matched. In the second case a VQ codebook was created by clustering the training features of these 60 speakers. A distortion easure based on the minimum Euclidean distance was used for speaker recognition. The failure rate of speaker recognition in first ase was found to be was found to be 10% while in the second case as found to be 14%. Matlab-7.10.0 was used for this study

Keywords—Mel Frequency Cepstral Coefficients (MFCC), Mel Window Design, K means, Vector Quantization, feature vector.

I. INTRODUCTION

Speech is the most fundamental signal that is used by human beings to convey information. Speech contains information not only about the message that is to be delivered to the other listener but it also contains information about the gender of the speaker, the language of the speaker, the age of the speaker, the ethnicity, the state of mind of the speaker (happiness, sadness and other emotions), the nature of the speaker (polite, kind, harsh etc). A human brain is trained in a manner so as to guess a lot of features about a person just by listening to the voice of a person. Designing a system that can behave just as a human being is a difficult and complex task and has not been achieved yet.

The most important application of an Automatic Speaker Recognition system is in security systems. A person may get access to a secured place based on his voice. This voice can either be text independent or text dependent (for example- a numeric code). These days it can also be used to buy products on order via telephone credit cards. Automatic Speaker recognition can also be used for monitoring of people based on their voices. This application can be used for tracking criminals in specific areas based on their previously recorded voices. Speaker recognition Systems can be broadly classified into two categories.

(1) Speaker Verification: In case of speaker Verification, a speaker's claimed identity is verified. Thus it is a yes or no problem.
(2) Speaker Identification: In case of Speaker Identification, the speaker is directly identified. Generally the Id or the name of the person is displayed if he/she is identified.

There are further two classifications of Speaker Identification namely,

(a) Open set Identification: In this case, a decision is made as to who is the speaker among all the already present speakers in the database (speaker data base is created by recording the sample speeches of various speakers). If the speaker database does not contain the recorded speech of the speaker, it is said that the speaker is not present in the database.
(b) Closed set Identification: In this case, a decision is made as to who is the most likely the speaker among all the already present speakers in the database. Thus the results are that of the nearest matching features.

1.3 Classification based on algorithm used for Identification are:

(i) Text dependent
(ii) Text independent

In this paper a text independent automatic speaker recognition system has been presented. The goal of a text-independent automatic speaker recognition is to identify a speaker independent of what text he/she speaks. This means that the system must be able to identify a person when he/she utters any speech. In order to make such a system an algorithm must be implemented that can extract unique features from a speaker that are independent of what is spoken. So application specific speech feature extraction is the first step in any speech based system. Here MFCC technique was used to extract speaker specific unique features.

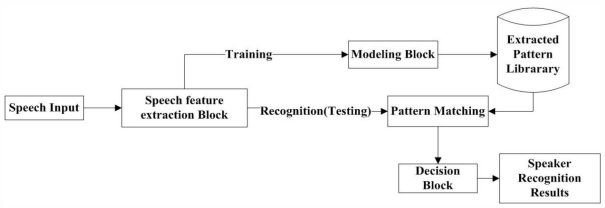

Fig 1: Framework of a basic Automatic Speaker Recognition System

978-1-4799-2995-5/14/$31.00©2014 IEEE

II. FEATURE EXTRACTION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

Various techniques have been used for speaker recognition; Amit Kumar Singh et al., 2013 have presented an overview of some of these prominent techniques [1]. In 1966 L.C.W Pols's research had given a way to results that suggested that the phonetically important characteristics of speech could be represented in a compact manner by a set of Mel-frequency Cepstrum Coefficients (MFFCs). In 1980 Davis and Mermelstein showed that a very high performance of the Mel Frequency Cepstrum coefficients is due to the reason that the perceptually useful and pertinent features of the short term speech spectrum can be represented in a better manner as compared to a linear frequency Cepstrum or a linear prediction spectrum [3]. Davis and Mermelstein also gave a conclusion that the Cepstrum parameters (MFCC, LFCC and LPCC) which apply frequency smoothed representations of the log magnitude spectrum perform better than the LPC and reflection Coefficients (RC) in representing the significant acoustic data. Since MFCCs were found to be very efficient in speaker recognition and thus various other modifications of MFCCs based on the window band design, log compressed window bank output and the approximate models of the nonlinear pitch sensitivity of human were proposed by various researchers. Young et al., 1995 described the HTK MFCC FB-24(they are computed through a window bank of 24 windows) [4]. MFCC features are extracted by passing the speech data through a set of triangular windows that are placed in a linear perceptual Mel scale. MFFCs are based on the well known variations of the human ears critical bandwidth with frequency and perhaps are the most frequently used parametric representation used in speaker recognition.

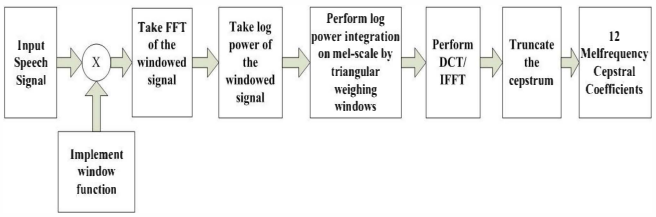

Fig 2: Block Diagrammatic representation of steps involved in the calculation of Mel frequency cepstral coefficients

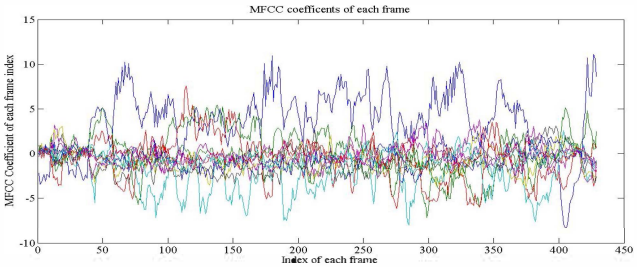

Fig 3: A waveform showing MFCC coefficients for a sample statement "it provided the basic information which is necessary for understanding the expo" spoken by speaker- m00026.

III MEL WINDOW DESIGN

The Mel frequency cepstrum is a short-time power spectrum of an input speech signal. Calculation of this short time power spectrum involves various steps. First the input time domain speech signal is windowed using any of the windowed functions (hamming, hanning, rectangular etc). Most generally hamming window is used because of its advantages over other windowing techniques. Once the signal is windowed, Fast Fourier Transform (DFT) of the windowed signal is calculated. Thus the input speech signal is now converted to frequency domain input signal in which there is a magnitude plot and a phase plot. The N-point DFT of the windowed speech signal can be calculated as [1]

$$X(k) = \sum_{n=0}^{N-1} x(n).\exp(\frac{-j2\pi nk}{N}), k = 0,1,...., N-1, \quad (1)$$

Since FFT gives only the amplitude and phase with respect to sample number. We are now in the need of frequencies against which these amplitudes and phases must be plotted in order to get a clearer picture of amplitude vs. the frequency plot. One of the techniques in this regard is the FFT bin technique.

$$f = \frac{sampling\ frequency}{Total\ number\ of\ samples}(sample\ number) \quad (2)$$

Thus corresponding to each sample number (increasing sample number) we get a frequency that is increasing. Also since the FFT results are symmetric in nature only half of the values are considered. The next step in Mel frequency cepstrum calculation is the calculation of squared power spectrum. A squared power spectrum in the context of the MFC calculation is the square of each FFT point corresponding to each sampled bit. The squared power spectrum is calculated as

$$p_{sq} = (b_1)(b_1) \quad (3)$$

where p_{aq} represents squared power corresponding to first FFT point. The log power spectrum is then calculated based on the formula.

$$p_{\log} = \log_{10}(p_{sq}) \quad (4)$$

Since melscale helps to space windows equally, the design and implementation becomes easier. In order to convert the frequency in Hz to frequency in mel, the following formula must be used. Mel frequency can be calculated using the formula given by (5).

$$f_{mel} = 2595\log_{10}(1+\frac{f}{100}) \quad (5)$$

Here f represents each FFT bin frequency corresponding to each sample number. The next step in MFC computation is the design of triangular windows. Each of these windows will be placed at equal spacing. These triangular windows are nearly linear up to 1000Hz (1 KHz) and are logarithmic (non-linear) post 1000Hz. A collection of these windows is called Mel-windowbank.

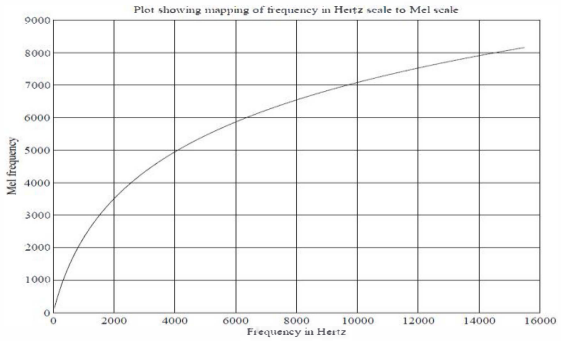

Fig 4:Plot Showing mapping of frequency in Hz to Mel scale.

The $$\psi_i(k) = \begin{cases} 0 & for \quad k \le k_{b_{i-1}} \\ \dfrac{(kk - _{b_{i-1}})}{(kk_{b_i} - _{b_{i-1}})} & for \quad k_{b_{i-1}} \le k \le k_{b_i} \\ \dfrac{(k_{b_{i+1}} - k)}{(k_{b_{i+1}} - k_{b_i})} & for \quad k_{b_i} \le k \le k_{b_{i+1}} \\ 0 & for \quad k > k_{b_{i+1}} \end{cases} \quad (6)$$

Number of these triangular windows may vary based on the need, design and accuracy requirements. In MFCC Extraction most commonly triangular windows are used. But Triangular windows may also be replaced by windows of trapezoidal shapes and other complex shapes that come from various auditory models. Sirko Molau et al. suggested that better results were sometimes seen with cosine shaped windows [5]. The generalized window equation is given by (6)[8,9].

With i =1,2,.....M.

Where M is the number of windows in the bank and various values of k_{b_i} are the boundary points of the windows.

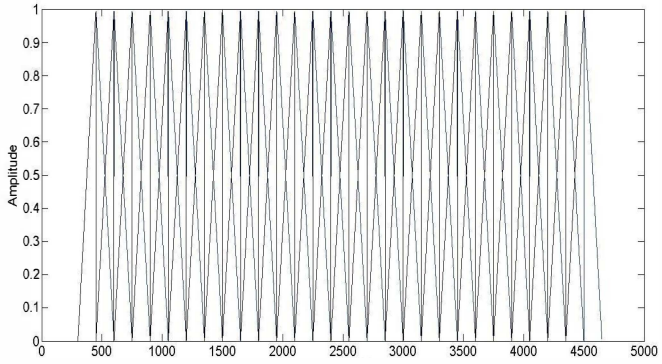

Fig 5: Mel-Windowbank structure

The next step is integrating all the values that lie in each window. Thus for M number of windows, we will get M integration values. We take the IFFT of integrated values. We are now in the cepstral domain. We may take initial 12, 14 or 16 coefficients from the M coefficients. These coefficients are basically the MFC coefficients. Here the melcepstrum was calculated using 12 coefficients.

IV. DATASET DESCRIPTION

A high quality large database was designed that consisted of 30 male and 30 female speakers. The motive behind creating this database was to take into consideration all possible variations of the speakers such as age, gender, state (region) etc. Each speech data consisted of a unique English sentence spoken by the speaker. Ten tests were conducted for each of the five users in a quiet environment (a common office room) that were chosen randomly. Every time the test signal was taken from test speaker and a distortion measure between the features based on the Euclidean distance between the feature sets of other users in the database was calculated. The results of distortion measures for two separate experiments that were conducted have been shown in Table 1. All recordings and tests were performed using Zoom H4n Handy recorder. All the speakers were asked to keep the microphone at a constant distance from their mouth. A handy recorder was used for ease in operation because of its high degree of mobility.

Table 1. Exp-1 : Distortion measure for direct feature matching,

Exp-2 Distortion measure for K-means clustering(C=64).

SpeakerID /Test no	T-1 (ED)	T-2 (ED)	T-3 (ED)	T-4 (ED)	T-5 (ED)	T-6 (ED)	T-7 (ED)	T-8 (ED)	T-9 (ED)	T-10 (ED)	Mean Distortion	Success Rate(%)	Failed Tests
Exp-1													
m00026	5.214	5.164	5.299	4.659	3.417	4.579	4.540	4.975	4.468	4.547	4.686	90	3
m00027	3.826	4.785	4.716	4.428	4.204	5.050	5.318	3.812	4.599	3.708	4.445	90	6
m00006	4.255	5.384	4.982	4.315	4.652	4.766	5.095	4.826	4.318	4.741	4.733	100	0
m00030	4.454	3.489	3.759	4.363	5.003	4.566	4.345	5.233	5.432	4.156	4.48	80	5.7
m00029	4.054	4.910	4.614	4.663	6.512	4.453	5.667	8.513	5.621	4.434	5.344	90	8
Exp-2													
m00026	5.240	5.909	5.323	6.401	6.076	5.709	5.289	6.121	6.966	7.932	6.097	80	3
m00027	5.212	4.032	5.372	4.355	4.965	5.153	4.458	4.654	4.600	5.061	4.786	90	6
m00006	4.114	5.064	5.130	6.349	5.739	5.322	5.365	5.820	4.914	5.047	5.287	90	4
m00030	5.721	6.412	4.585	4.781	4.698	4.531	4.964	5.378	6.720	5.181	5.297	80	2.9
m00029	4.054	4.910	4.614	4.663	6.512	4.453	5.667	8.513	5.621	4.434	5.344	90	8

A. *Recording scenario*

1 Microphone: Zoom H4n Handy Recorder [7]
2 Sampling Frequency: 48kHz
3 Language: English
4 Noise Environment : Quiet
5 Face contact: None
6 Recording Time: 2sec
7 Dialogue: Each speaker was given an English newspaper from where he/sheread a sentence of their choice. The recording was done in a quiet environment microphone connected to a Zoom H4n handy recorder.

B. Recorder Specifications

1 Recording Format: WAV (Quantization: 16/24bit, Sampling Frequency: 44.1/48/96kHz), MP3 (Bit Rate:48/56/64/80/96/112/128/160/192/224/256/320kbps/VBR, Sampling Frequency: 44.1kHz).
2 Playback Format: WAV (Quantization: 16/24bit, Sampling Frequency: 44.1/48/96kHz), MP3 (Bit Rate:32/40/48/56/64/80/96/112/128/160/192/224/256/320kbps/VBR, Sampling Frequency: 44.1/48kHz).
3 Input Impedance: balanced input = 1kΩ balanced / pin 2 hot, unbalanced input = 480kΩ unbalanced.
4 Input Level: balanced input = -10 to -42dBm, unbalanced input = +2 to -32dBm.
5 Built-in Stereo Mic: Unidirectional condenser microphone (Gain: +7 to +47dB).

V. THE SPEAKER RECOGNITION EXPERIMENT

Two separate experiments were performed to study the effects of clustering on the feature sets extracted from the experiment. In the first experiment the test features that were extracted were directly matched with the training features of all the speakers. In the second experiment, during the training phase, and also during the testing phase a VQ codebook was generated for every test speaker that was based on k-means clustering technique. If we have S Training feature vectors ($t_1t_2t_3....t_s$). These training features are partitioned into say J vector spaces. Each of these convex spaces are represented by a centroid which represent the feature vectors also called codeword. A collection of these codewords is a codebook[6].

During the recognition phase distortion measure based on Euclidian distance were computed for C = 64 as well as direct distortion measures were calculated. Five experiments were conducted in which a randomly chosen speaker underwent ten tests in each of the two experiments. The percentage change in the mean distortion gives us the margin for clustering and the range of error in the result that can be expected.

The test cases have been presented in Table I above. m00026, m00027, m00006, m00030 m00029, are speakerIDs corresponding to the test speakers. ED stands for Euclidean Distance. Failed cases- voice detected in this case was that of other than the test user. The performance of a text-independent speaker recognition system depends on the size of the database and the cluster size taken. A success rate of 75% was shown in [10] while in [11] a success rate 80% was shown with a database of 15 users. In [12] 87.5% success rate was shown with a database of 8 users. Here in the present case a success rate of speaker recognition in first case was found to be 90% while in the second case was found to be 86%.

VI. CONCLUSIONS

Two separate experiments were conducted for the performance evaluation of MFCC technique when applied to K means clustering. In the first case the speech features were directly matched. In the second case a VQ codebook was created by clustering the training features of these 60 speakers. The choice of number of clusters plays a vital role in the recognition rate. The failure rate of speaker recognition in first case was found to be was found to be 10% while in the second case was found to be 14%. The percentage rise in mean distortion for all the five test cases for the clustered case was found to be 13.18%. This gives intuitive idea regarding the choice of ideal number of clusters for a better recognition.

REFERENCES

[1] Amit Kumar Singh, Rohit Singh, Ashutosh Dwivedi. "Evolvement and recent research in parametric representations of speech features for automatic speaker recognition." In Proceedings of ICEECMPE International Conference, 2nd November 2013, New Delhi, India.

[2] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. Discrete-time signal processing. Vol. 5. Upper Saddle River: Prentice Hall, 1999.

[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP- 28, pp. 357-366, Aug. 1980.

[4] Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland(1995). The HTK Book, version 2.1, Department of Engineering, Cambridge University, UK.

[5] Molau, Sirko, Michael Pitz, Ralf Schluter, and Hermann Ney. "Computing mel-frequency cepstral coefficients on the power spectrum." In Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'01), IEEE International Conference on, vol. 1, pp. 73-76. IEEE, 2001.

[6] Soong, F., A. Rosenberg, L. Rabiner, and B. Juang. "A vector quantization approach to speaker recognition." In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85., vol. 10, pp. 387-390. IEEE, 1985.

[7] ZoomH4n, manual, Available: http://www.zoom.co.jp/english/products/h4n/ (as of 10-Dec-2013).

[8] Chakroborty, Sandipan, and Goutam Saha. "Improved text independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter." International Journal of Signal Processing 5, no. 1, pp. 11-19. (2009).

[9] Ganchev, Todor, Nikos Fakotakis, and George Kokkinakis. "Comparative evaluation of various MFCC implementations on the speaker verification task." In Proceedings of the SPECOM, vol. 1, pp. 191-194. 2005.

[10] Hemlata Eknath Kamale, Dr.R. S. Kawitkar , "Vector Quantization Approach for Speaker Recognition.", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 3, Special Issue, In proceedings of E-NSPIRE, Loni, India, pp. 110-114, March-April 2013.

[11] Gupta, Arnav, and Harshit Gupta. "Applications of MFCC and vector quantization in speaker recognition." International Conference on Intelligent Systems and Signal Processing (ISSP), pp. 170-173, 2013.

[12] Samudre, N. A. "Text-Independent Speaker Identification using Vector Quantization." International Journal of Engineering, Vol. 2, Issue 8, pp. 1787- 1792, August – 2013.