Analysis of Feature Extraction Techniques for Speech Recognition System

Rajeev Ranjan, Abhishek Thakur

Abstract—The audio signal is filtered using a method known as feature extraction technique. In this article, the feature extraction technique for speech recognition and voice classification is analyzed and also centered to comparative analysis of different type of mel-frequency cepstral coefficients (MFCC) feature extraction method. The MFCC technique is used for deduction of noise in voice signal and also used for voice classification and speaker identification. The statistical results of the different MFCC techniques are discussed and finally concluded that the delta-delta MFCC feature extraction technique is better than the other feature extraction techniques.

Keywords— Feature Extraction; Voice Data; MFCC; Delta-Delta MFCC; Cepstral Coefficient.

I. INTRODUCTION

For the audio signal analysis used a feature extraction technique known as MFCC. The objective of this paper is to transform the audio waveform to frequency domain representation, for advanced signal processing and analysis. Here also discuss the comparative analysis different MFCC methods. The important parameter of speech signal in feature extraction method isCepstral coefficients andpitch frequency. It is used for the speech recognition, speech synthesis and speaker verification, etc [1-3]. Here extract the highlights of the discourse section, for example, essential frequency, groups, Cepstral coefficient line spectral pairs, MFCC and spectrogram [4, 5]. Here we are mainly discussing Cepstral coefficient for the speech recognition and different types of the MFCC. The process of the feature extraction technique is first the speech is analyzed over short frame window and then each short frame window, obtained a spectrum by Fast Fourier transform (FFT) [5]. The mel spectrum is obtained, when the output of the FFT is passed through a Mel-filter. For the MFCC the mel spectrum is performed Cepstral coefficients [6-8]. Therefore, the audio signal is signified as a sequence of the Cepstral vectors.

The rest of this paper, is structured as follow: In section II demonstrate the feature extraction techniques. The architecture of proposed model in section III. In the section IV described the types of the MFCC. The simulated results and discussion are in section V. Finally, concluded the results in section VI.

Revised Manuscript Received on May 29, 2019.

Rajeev Ranjan,Electronics & Communication Engineering, Thapar Institute of Engineering and Technology, Patiala, India. (Email: rajeevranjan1134@gmail.com)

Abhishek Thakur,Electronics & Communication Engineering, Thapar Institute of Engineering and Technology, Patiala, India. (Email: abhithakur25@gmail.com)

II. FEATURE EXTRACTION TECHNIQUES

The feature extraction block diagram is shown below in Figure 1. It is divided in to three parts as pre-emphasis, frame blocking & windowing and feature extraction.

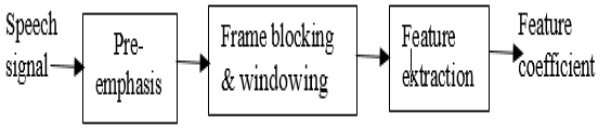

Figure 1: Block diagram of feature extraction

Mel frequency cepstral coefficients

Feature extraction of the speech signal is the primary step for any speech recognition system [2]. It is accountable for extracting significant information from each frames, as a feature parameters and vectors in [7,8], here used MFCC technique for feature extraction. It is a mathematical representation of speech data. It is also used in other area of speech, such as speech processing, speech synthesis, and speaker recognition.

III. METHODOLOGY

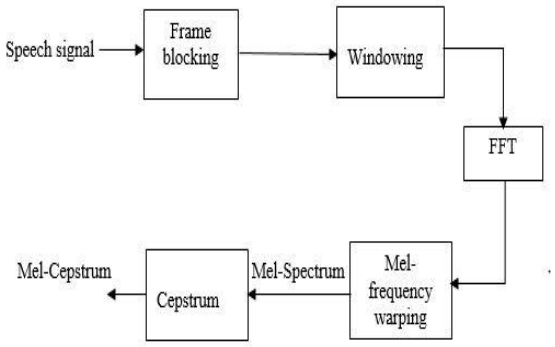

Figure 2: MFCC extraction process

Per-emphasis

The aim of this step is to minimize the high frequency section of the audio which was stifled in the human's audio production mechanism. It can also intensify the prominence of high frequency formats. The speech after the pre-emphasis sounds sharper with a small volume. The speech signal $s[n]$ is sent to a high pass lower order filter [5].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$s_y[n] = s[n] - \alpha s[n-1]$ (1) where $S_y[n]$ output signal and α lies between, $0.9 < \alpha < 1$

The Z transform of the filter is defined as -

$$H(Z) = 1 - \alpha Z^{-1} \quad \textbf{(2)}$$

Frame blocking

In this step, an audio signal can be divided into a numbers of frames and some parts of each frame is overlap to each other. So the each frame can be analyzed and synthesized individually without loss of information. Therefore, the each frame can be denoted by a single vector. In this method a continuous audio signal is divided into N numbers of samples in a frame.Where each neighboring frames are disjointed by M (M < N). Here the first frame contains N numbers of samples and the next frame starts M numbers of the samples so the first frame and second frame overlap by (N – M) numbers of samples. Correspondingly, the next frame contains 2M number of samples. Therefore, the first frame and the third frame are overlap by (N – 2M) numbers of samples. In the similar manner the speech signal is framing so no discontinuity occurs in audio samples.

The frame blocking of the audio data is important for the angle of computational complexity and analyzing sufficiently over a short duration of time. so its characteristics are stationary during that interval. Whereas for a long time interval the characteristics is change to mirror the individual speech sounds being talked. For computation the value of N is equal to 149 as a resolution between frequency and time resolution. When the frequency and time coefficients are observed then the corresponding power spectrum of the audio data will be appeared in the outcome segment.

Hamming window

In this step,the hamming window is multiplied with each frames so the continuity will be occurs in between the first point of the first frame and last point of the last frame.

If the signal is denoted as $s[n]$, and hamming windowing is denoted as $w[n]$, then it is mathematically defined as-

$$= (s[n] \times w[n]) \quad (4) \text{ where,}$$

$$w[n,\alpha] = (1-\alpha) - \alpha \cos(2\pi n / N-1); \quad 0 \le N \le N-1 \quad \textbf{(5)}$$

Triangular band pass filter

In this step, gives the log energy of the each triangular passband filter.When we will multiply by a set of twenty triangular passband filters. The location of these filters are equal and it gives a relation in between linear frequency and mel frequency and it is defined as-

$$mel(f) = 2595 \times \log_{10}(1 + \frac{f}{100}) \quad \textbf{(6)}$$

The purpose of the triangular bandpass filter is to moderate the features and magnitude spectrum of the speech signal. Therefore, the harmonics are compressed to achieve envelope of the spectrum.

Mel frequency warping

The human ear hears frequency being nonlinear. Where the scaling of frequency straight upto 1 kHz. The audio sound-related framework is scaled by frequency called mel scale frequency. The talked signs for every edge is permitted through the mel scale frequency band pass channel to mirror the audio sound-related discernment. Consequently, for each tone with a real frequency f, estimated in Hz, an abstract pitch is estimated on a scale called the mel scale. The mel-frequency scale is straight frequency dividing beneath 1000Hz and a log is logarithmic separating above 1000Hz. The mel-frequency is relative to the logarithm of the direct frequency, duplicating comparative impacts in the human's sound-related observation level. The beneath precedent demonstrates the connection in between mel- frequency and direct frequencies.

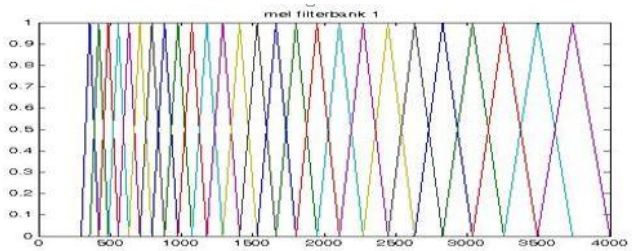

Figure 3: Mel frequency triangular filter.

The reason for the triangular bandpass filter is to smooth the magnitude spectrum of a speech signal and compressed the cepstral coefficients of that signal. Therefore, the compressed harmonics are obtain.

Mel frequency Cepstral coefficients with discrete cosine transform

In this step, we apply DCT on the $20\log_{10} E_k$ are obtained from the triangular bandpass filter to have L mel-scale Cepstral coefficient.

$$C_m = \sum_{K=1}^{N} \cos[m \times (K-0.5) \times \pi / N] \times E_k \quad \text{where } m = 1,2,..,L \quad \textbf{(7)}$$

where MFCC is denoted by C_m, number of triangular filters is denoted by N and M represents order of cepstral coefficients. After all the analytical and mathematical operation of these steps, we get the MFCC coefficients and its spectrum

IV. TYPES OF THE MFCC

In this section, illustrate the different types of MFCC techniques for feature extraction and classification.

MFCC:

Figure 4 shows that the all 12 coefficients of each frame of the MFCC. Each frame of MFCC is represented by a different color.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

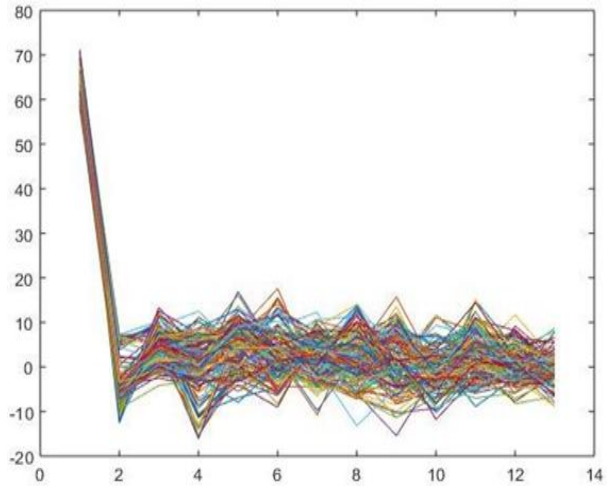

Figure 4:Coefficients of MFCC

Delta and Delta-Delta MFCC

The delta and delta-delta (double delta) MFCC is another types of MFCC which is known as differential and acceleration coefficients. Using the MFCC technique described only power spectrum of a single frame, where as it appears like speech signal would also having some data in the dynamics that is how the MFCC coefficients are arranged over time. They find out the computation of the MFCC coefficients over the original feature vector performance. When we have been twelve MFCC coefficients, then we will also get twelve delta coefficients and twelve delta-delta coefficients. After the combining the feature vector length is thirty six

For compute the differential coefficients as-

$$D_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad \textbf{(8)}$$

where delta coefficient is denoted by D_t, from frame t. They can be calculated in terms of the static coefficients c_{t+n} to c_{t-n} , where the typical value for N is 2. The acceleration coefficients or delta-delta MFCC are computed in the same manner, but it has been computed from the differential coefficients or delta MFCC, not static coefficients. Figure 5 and figure 6 shows that the all 12 coefficients of each frame of the Delta MFCC and Delta-Delta MFCC respectively. Each frame of Delta MFCC and Delta-Delta MFCC is represented by a different color.

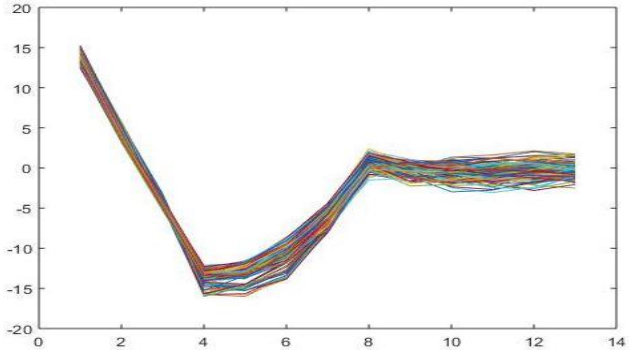

Figure 5: Coefficients of Delta MFCC

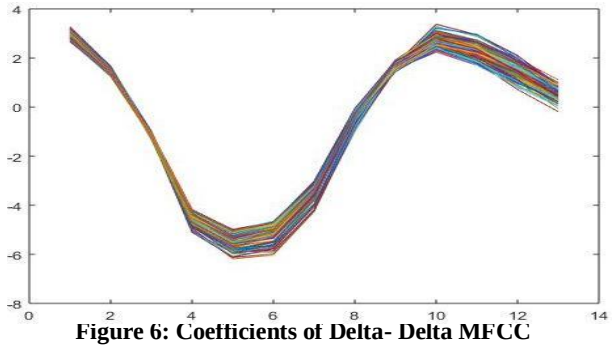

Figure 6: Coefficients of Delta- Delta MFCC

V. EXPERIMENTAL RESULTS

In this section, synthesized and analyzed the simulated results and performance of the system. In this paper, all the simulation work has been done using MATLAB. In figure 7 and figure 8 shows that the first and second coefficients of MFCC in time and frequency domain respectively. Similarly, we can also find the remaining feature coefficients of MFCC.

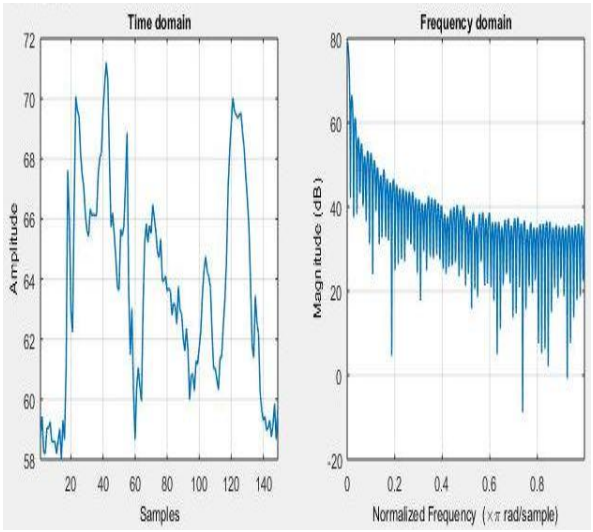

Figure 7: First coefficient of MFCC

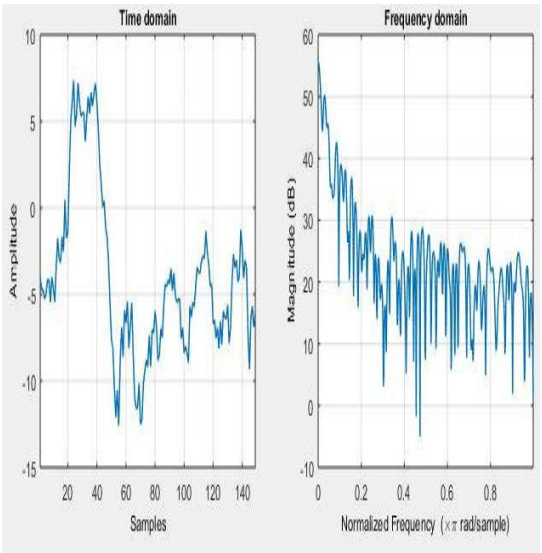

Figure 8: Second coefficient of MFCC

In figure 9 describe the mean value of MFCC, Delta MFC, Delta-Delta MFCC at each frames and figure 10 is describe the Standard deviation at each frames.

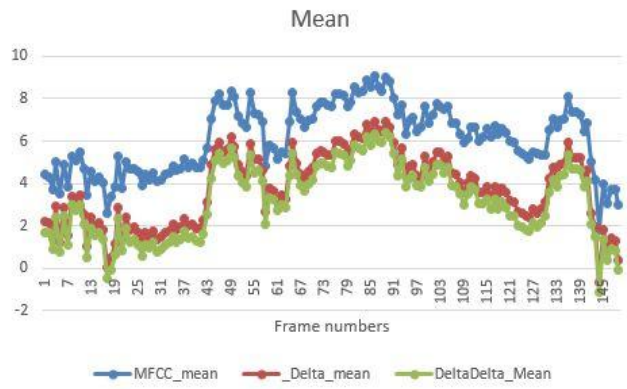

Figure 9: Mean values of each frames

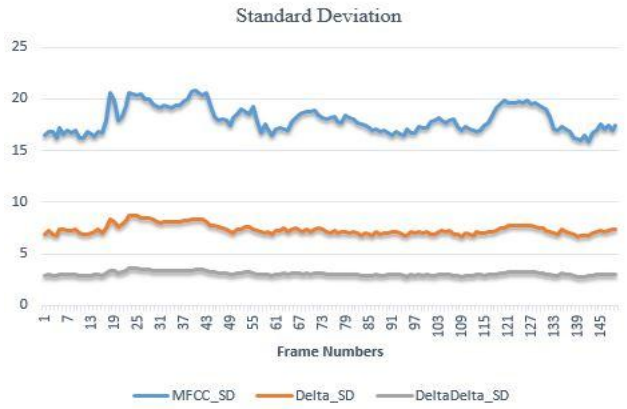

Figure 10: Standard deviation of each frames

VI. CEPSTRUM ANALYSIS

Figure 11 show the mel frequency cepstrum value of each frame. Here the dark region indicates the formats or peaks in the spectrum, in this region high amplitude is occurs. The formants is identified the voice and their transition and formants carry the identity of the voice. Forments and smooth curve connecting them this smooth curve is referred to as spectral envelope.

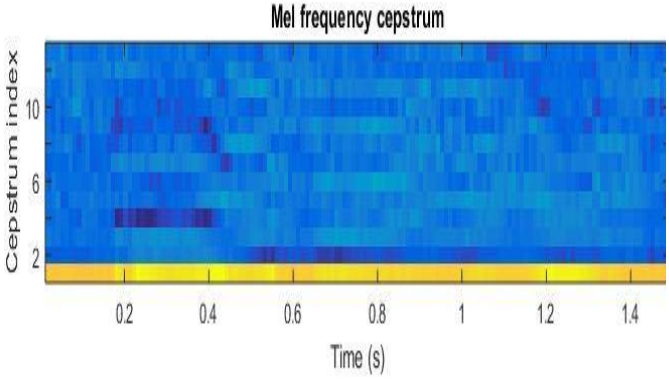

Figure 11: Cepstrum values of each frames

VII. CONCLUSION

The feature extraction technique is very important technique that are used for speech and speaker recognition and also used for feature classification of voices. There are several types of feature extraction techniques such as PLP, LPC and different types of MFCC etc. We concluded that the minimum standard deviation occurs in Delta-Delta MFCC and better Cepstral coefficients compare to other form of MFCC. So it is a best feature extraction technique in compare to other.

REFERENCE:

1. L. Deng, J. WU, J. Droppo and A. Acero, "Analysis and comparison of two speech feature extraction / Compensation algorithms," *IEEE signal processing letters*, vol. 12, no. 6, 2005.
2. R. Ranjan and R. K. Dubey, "Isolated word recognition using HMM for Maithili dialect," *IEEEInternational conference on signal processing and communication*, pp. 32-328, 2016.
3. S. Boruah and S. Basishtha, "A study on HMM based speech recognition system," *IEEE Int. Conf. on computational intelligence and computing research*, pp.4799-1597, 2013.
4. J. P. Campbell and Jr. "Speaker recognition: A Tutorial" *Proceeding of the IEEE*, vol. 85, pp. 1437- 1462, 1997.
5. S. Furui. "Fifty years of progress in speech and speaker recognition," *Proc. 148th ASA Meeting*, 2004.
6. A. Rosenberg, "Automatic speaker recognition: A review," *Proc. IEEE*, vol. 64, pp. 475-487, 1976.
7. M. A. Hossan, S. Menon, and M. A. Gregory, "A novel approach for MFCC feature extraction," *school electrical and computer engineering RMIT University*, vol.8, pp. 978-993, 2010.
8. H.S. Jayannaand S. R. M. Prasanna, " Analysis, feature extraction, modelling and testing techniques for speaker recognition," *IETE Technical Review*, vol. 26, 2009.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication