International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-5, June 2013

Speaker Recognition System Based On MFCC and DCT

Garima Vyas, Barkha Kumari

Abstract- This paper examines and presents an approach to the recognition of speech signal using frequency spectral information with Mel frequency. It is a dominant feature for speech recognition. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non linear mel scale of frequency. The performance of MFCC is affected by the number of filters, the shape of filters, the way that filters are spaced, and the way that the power spectrum is warped. In this paper the optimum values of above parameters are chosen to get an efficiency of 99.5 % over a very small length of audio file.

Keywords: Speech recognition, Feature extraction, Feature Matching, DCT, MFCCs

I. INTRODUCTION

All of the man made sounds, which affect our lives; speech and music are most significant. Speech is the natural and one of the secure interface between human and machine. Speech recognition is an important and emerging technology with great potential .It is a process used to recognize speech uttered by a speaker and has been in the field of research for more than six decades since 1950s [1]. It can be used in many applications like, security devices, household appliances, cellular phones, ATM machines and computers. MFCCs are short- term spectral- based features. They are derived from a type of cepstral representation of the audio clip . The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound. The organisation of this paper is as follows. Part I describes MFCCs for modelling speech and voice recognition. Part II shows some of the experimental results with some definite set of parameters. Part III is about the conclusion and future work.

II. MFCCs FOR SPEECH MODELLING

The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The name mel comes from the word melody to indicate that the scale is based on pitch comparisons. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale.

Manuscript published on 30 June 2013.
***** Correspondence Author (s)
Garima Vyas, Deperment of ECE, Amity University , Noida, India.
Barkha Kumari Deperment of ECE, Amity University , Noida, India.

$$\text{Fmel} = 2595 \log_{10}(1+ f/700)$$

The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic 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. Equation 1 shows frequency to mel scale conversion:
Figure 1 shows the flow chart of MFCC feature extraction.

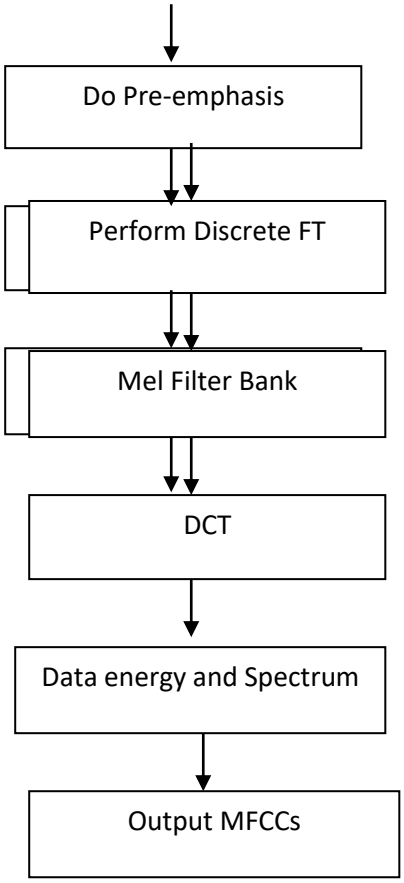

Figure 1. Process model for extracting MFCCs from an audio speech

a) Pre-emphasis: The digitized speech is put through low order system, to spectrally flatten the signal and to make it less susceptible to finite precision effects. Normally, a one coefficient FIR filter is known as pre emphasis filter.

b) Framing: Frames are typically 20-30ms with an overlap of 10-15ms. The justification for such segmentation is that the speech signals are non stationary and exhibit quasi-stationary behaviour at shorter durations. The conventional speech recognition system use features that are extracted with single frame size and frame rate.

c) Windowing: The hamming or hanning window functions can be used. The window function is convoluted with the input signal. Equation 2 and 3 shows hann and hamming window functions respectively.

$W(n)= 0.5(1\pm \cos (2\Pi n/ N-1))$

The generalized hamming window is in form of

$W(n) = \alpha-\beta \cos (2\Pi n/ N-1)$

d) **DFT:** To convert each frame of N samples from time domain into frequency domain FFT is being used. The Fourier Transform is used to convert the convolution of the glottal pulse X[n] and the vocal tract impulse response H[n] in the time domain. This statement supports as shown in Eq. (4) below:

$$Y(w)=FFT[h(t)*X(t)]=H(w)*X(w)$$

e) Mel filtering: Each filter's magnitude frequency response is triangular in shape and equal to unity at the Centre frequency and decrease linearly to zero at centre frequency of two adjacent filters.

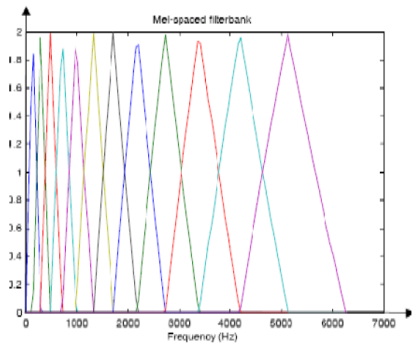

Figure 2: Examples of Mel filter bank

f) DCT: A discrete cosine transform (DCT) expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. This is the process to convert the log Mel spectrum into time domain using DCT. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

III. EXPERIMENTAL RESULTS

The original image is read and displayed. The peaks of audio speech lies between -1 to 1 as it are clearly shown in figure 1.

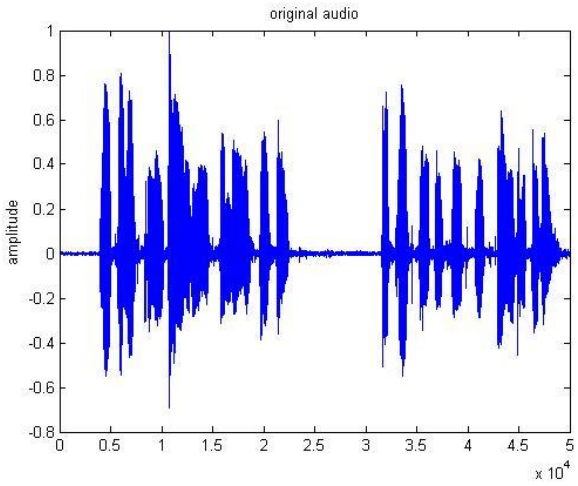

Figure 3: Plot of original Audio

Now the original audio signal is passed through the hanning window (shown in figure 4) and to convert this time domain signal to frequency domain signal its 512 point FFT is shown in figure 5.

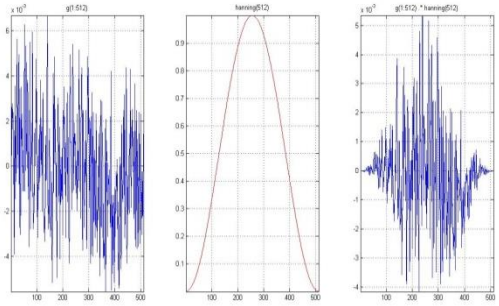

Figure 4: Output of hanning window

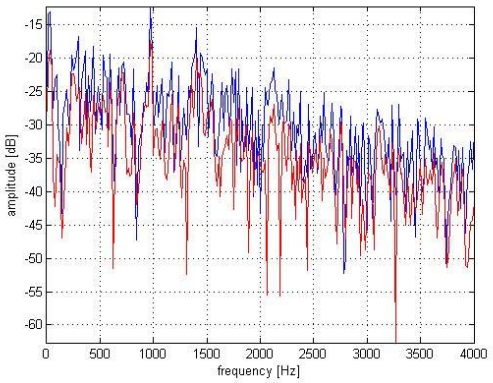

Figure 5: Windowed signal in frequency domain

The windowed signal is passed through 24 level triangular filter bank (see figure 6) and the output is in the form of coefficients. This output is the cepstral coefficients and are plotted. (Shown in figure 7).

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

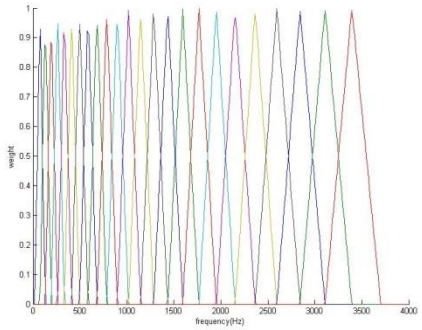

Figure 6: Triangular Filter Bank

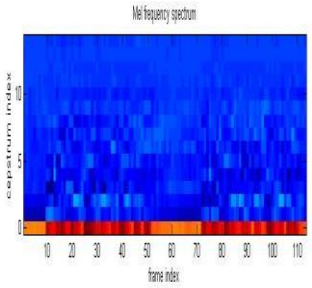

Figure 7: Mel frequency spectrum

Now we have extracted the MFCC feature from an audio clip. Next phase is the matching phase. A database is created which has some pre recorded audio clips. Now the query is passed and matched with all the audio in database. The recognition is done by matching the centroid of query with the centroid of stored audio. The result of speaker recognition is shown in figure 8.

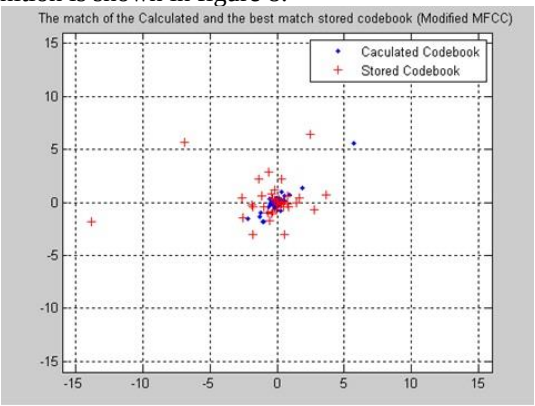

Figure 8: Centroids of query and stored audio.

IV. CONCLUSION AND FUTURE WORK

This paper presents the novel and efficient scheme for speaker recognition. It works quite well for its simplicity and lack of any more complicated techniques, though it is not super-simple. In the end, the features that we chose to collect ended up being good enough to identify between people. Indeed, the cepstrum is a very useful regime in which to analyze and characterize speech.

We are aware that the proposed scheme has not been tested in real noisy environment in which the audio signal will be subject to a variety of noises and degradations. In future we'll focus on testing using degraded audios samples and improve the proposed algorithm accordingly.

REFREENCES

1. Md Sahidullah, and Goutam Saha, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition" IEEE signal processing letters, vol. 20, no. 2, 2013, pp 149-153.
2. A.S.Bhalerao and V.B.Malode, "Implementation of Automatic Speaker Recognition on TMS320C6713 Using MFCC", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Coimbatore, India, 2013, pp 1-4.
3. Genevieve I. Sapijaszko and Wasfy B. Mikhael, "An Overview of Recent Window Based Feature Extraction Algorithms for Speaker Recognition" IEEE signal processing letters, vol. 12, 2012, pp 880-884.
4. R. Rajalakshmi and A. Revathy, "Comparison of MFCC and PLP in Speaker Identification using GMM" International Conference on Computing and Control Engineering (ICCCE 2012), Coimbatore , 2012, pp 110-114.
5. Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, Shihab Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition" IEEE signal processing letters, Maryland ,vol.12, 2011, pp 80-84.
6. Ibrahim Patel, Dr. Y. Srinivas Rao, "Speech Recognition Using HMM with MFCC- an Analysis Using Frequency Spectral Decomposition Technique" Signal & Image Processing : An International Journal, Vol.1, No.2, 2010, pp 101-110.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)