International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 10 (November 2014) www.ijirae.com

Voice Recognition Using MFCC Algorithm

Koustav Chakraborty
EXTC Department
FCRIT Vashi, India

Asmita Talele
EXTC Department
FCRIT Vashi, India

Prof. Savitha Upadhya
EXTC Department
FCRIT Vashi, India

Abstract---Human voice plays a very important role as a vital biometric parameter for authentication and identification. Voice recognition is a biometric technology used to identify one particular person's voice. It provides enhanced security, convenient authentication and considerable cost saving. It can be performed using many algorithms and speech models. Mel Frequency Cepstral Coefficients (MFCC) algorithm is generally preferred as a feature extraction technique to perform voice recognition as it involves generation of coefficients from the voice of the user that are unique to every user.

Keywords— Voice, Biometric technology, Feature extraction, Authentication, MFCC

I. INTRODUCTION

Voice can combine what people say and how they say it by two-factor authentication in a single action. Other identifications like fingerprints, handwriting, iris, retina, face scans can also help in biometrics but voice identification is needed as an authentication that is both secure and unique. Voice can combine two factors, namely, personal voice recognition and telephone recognition. Voice recognition systems are cheap and easily understood by users. In today's smart world, voice recognition plays a very critical role in many aspects. Voice based banking, home automation and voice recognition based gadgets are some of the many applications of voice recognition.[1]

II. MFCC AS A VOICE RECOGNITION ALGORITHM

Mel frequency Cepstral coefficients algorithm is a technique which takes voice sample as inputs. After processing, it calculates coefficients unique to a particular sample. In this project, a simulation software called MATLAB R2013a is used to perform MFCC. The simplicity of the procedure for implementation of MFCC makes it most preferred technique for voice recognition.

A.GENERATION OF COEFFICIENTS USING MFCC

MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. The step-by-step computation of MFCC is explained.[2]

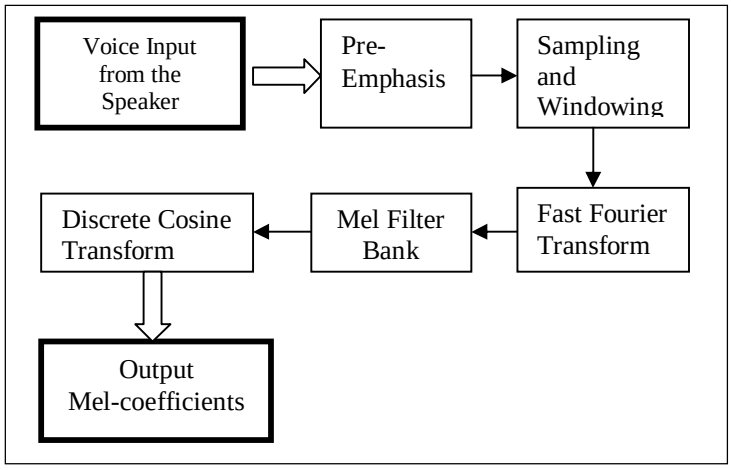

Fig 1 Block diagram for obtaining MFC coefficients

B. PRE EMPHASIS

The speech signal x(n) is sent to a high-pass filter :

$$y(n) = x(n) - a * x(n - 1) \quad (1)$$

where y(n) is the output signal and the value of a is usually between 0.9 and 1.0.
The Z transform of this equation is given by :

$$H(z) = 1 - a * z^{-1} \quad (2)$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.[3]

B. FRAME BLOCKING

The input speech signal is segmented into frames of 15~20 ms with overlap of 50% of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, zero padding is done to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 256 sample points, then the frame duration is 256/16000 = 0.016 sec = 16 ms. Additional, for 50% overlap meaning 128 points, then the frame rate is 16000/(256-128) = 125 frames per second. Overlapping is used to produce continuity within frames.

C.HAMMING WINDOW

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by
x(n), n = 0,…N-1, then the signal after Hamming windowing is,

$$x(n) * w(n) \quad (3)$$

where w(n) is the Hamming window defined by

$$w(n) = 0.54 - 0.46 * \cos(2\pi n/(N-1)) \quad (4)$$

where $0 \leq n \leq N-1$

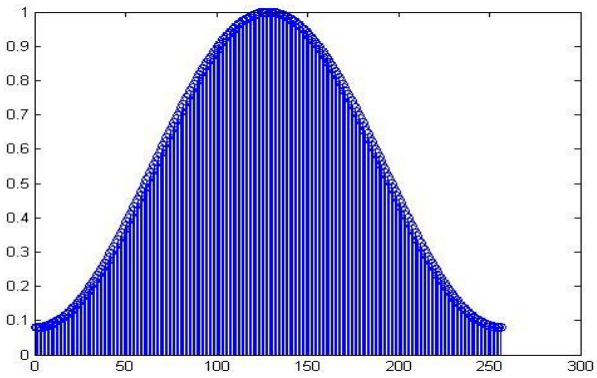

Fig 2 Plot of Hamming Window

D.FAST FOURIER TRANSFORM

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore FFT is performed to obtain the magnitude frequency response of each frame.
When FFT is performed on a frame, it is assumed that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, FFT can still be performed 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 multiply each frame by a hamming window to increase its continuity at the first and last points.[3]

E.TRIANGULAR BANDPASS FILTERS

The magnitude frequency response is multiplied by a set of 40 triangular band pass filters to get the log energy of each triangular band pass filter. The positions of these filters are equally spaced along the Mel frequency.
From centre frequencies from 133.33 Hz to 1 kHz, there are 13 overlapping (50%) linear filters, while for centre frequencies from 1 kHz to 8 kHz there are 27 overlapping filters spaced logarithmically.

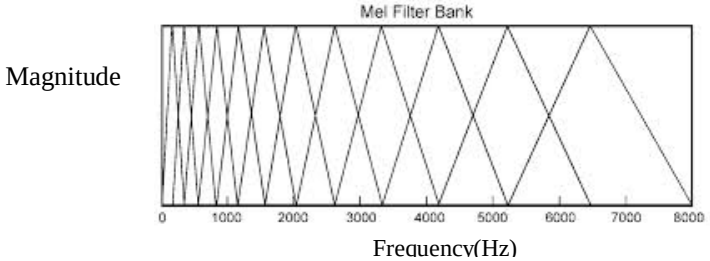

Fig. 3 Mel Filter bank

F.DISCRETE FOURIER TRANSFORM

In this step, DCT is applied to the output of the N triangular bandpass filters to obtain L mel-scale cepstral coefficients. The formula for DCT is,

$$C(n) = \sum Ek * \cos(n * (k - 0.5) * \pi/40)) \quad (5)$$

where n = 0,1,..to N

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. In this project, there are N = 40 and L = 13. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition.

III. MFCC IMPLEMENTATION

The following are the major steps involved in the implementation of the MFCC algorithm:

A. RECORDING AND SAMPLING

The recorded speech signals are sampled and stored using Audacity. The sampling is done at a rate of 16000 samples per second. Each speech signal is divided into windows of 16 ms each and hence, 256 samples each.

MFCC is implemented for each of these windows and a set of parameters is extracted per window. The first window consists of first 256 samples. The second window overlaps half of the first window and consists of 128 samples of the first window and 128 samples after it. Hence a 50% overlap is used.

It is observed that the same speaker saying the same word at two different instants have many variations. So it is important to calculate of the coefficients which almost remain same for a speaker at different instants becomes important.

B.MEL FILTER BANK

There are 40 Mel filters that form Mel filter Bank. Each filter passes a particular set of frequencies corresponding to samples from a frame. For a 256 sample frame, the filter bank spreads over 128 samples only because the FFT is symmetric.

C. MEL FREQUENCY CEPSTRAL COEFFICIENTS

Voice samples of two speakers saying the same word "HELLO" at two different instants were passed through the MFCC algorithm and their respective MFCC Coefficients were extracted, considering two voice samples per speaker, that is one that is stored as template in the database and the other is real time input.

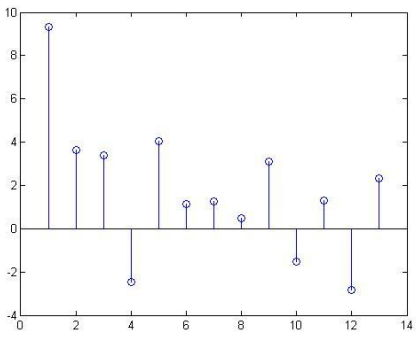

Fig.4(a) Speaker 1(Male) template

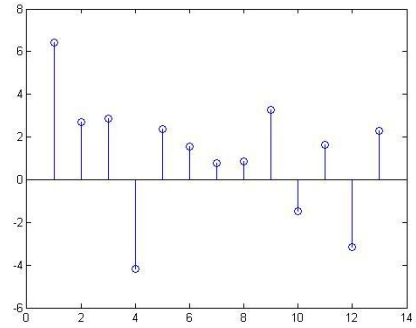

Fig.4(b) Speaker 1(Male) Real Time Input

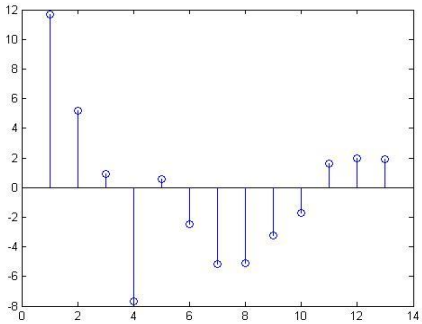

Fig.5(a) Speaker 2(Female) template

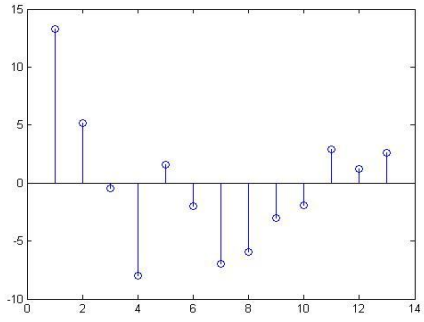

Fig.5(b) Speaker 2(Female) Real Time Input

III. CONCLUSION

It was observed that MFCCs for every individual user was unique. Certain variations were observed due to difference in the locality of the recording area. These MFCCs are then compared, that is, the MFCCs of the template and real time input are compared for every user. In programming, Euclidean distance is used to compare the template and real time input. In this manner, MFCC algorithm is used for voice recognition**.**

ACKNOWLEDGMENT

We take this opportunity to thank our project guide, ***Ms. Savitha Upadhya*** for her guidance and support throughout the course duration. Her efforts to clear our concepts and to help us code the entire algorithm was valuable for the development of this project. Her role as a the project coordinator helped us to meet all our deadlines.

We would like to express our gratitude towards ***our Head of the Department Dr. KTV Reddy,*** **ou**r ***Principal Dr. Rollin Fernandes*** and all the professors of EXTC Department for their support, encouragement and suggestions. We would also like to thank all the lab assistants because without their assistance in permitting the use of all the laboratory equipments, this project would not have been completed in the stipulated time duration.

Last but not least, we would like to thank our family members and our classmates for their valuable suggestions and constant motivation.

REFERENCES

[1]http://biometrics.pbworks.com/w/page/14811349/Advantages%20and%20disadvantages%20of%20technologies
[2] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
[3] L. R.Rabiner and R. W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

