International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

MFCC AND DTW BASED SPEECH RECOGNITION

R. Harshavardhini[1], P. Jahnavi[2], SK. Zaiba Afrin[3], S.Harika[4], Y.Annapa[5], N. Naga Swathi[6]

[1,2,3,4,5]Student, Dept. Of Electronics and Communication Engineering, Bapatla Engineering College, Andhra Pradesh, India
[6] Asst.Prof., Dept. Of Electronics and Communication Engineering, Bapatla Engineering College, Andhra Pradesh, 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. Speech Recognition is the process in which certain words of a particular speaker will automatically recognized that are based on the information included in individual speech waves. Speech recognition has wide range of applications in security systems, health care, telephony, military and equipment designed for handicapped. Since speech is a continuous varying signal, proper digital processing algorithm has to be selected for speech recognition. To obtain required information from the speech sample features have to be extracted from it. MFCC and DTW are two algorithms adopted for feature extraction and pattern matching respectively. In this paper, for identifying the good feature of the speech signal MFCCs are derived by applying various window functions. The proposed algorithms are implemented in MATLAB.*

Key Words: **MFCC*(*Mel-Frequency Cepstral Coefficients), DTW(Dynamic Time Warping), Window Technique, Speech Recognition, Pattern matching.**

1. INTRODUCTION

Speech is one of the ways to express anything. Now-a-days these speech signals are also used in communicating with machine and biometric recognition technologies. Speech recognition is the process of enabling a computer to identify and respond to the sounds produced in human speech. Speech Recognition has a wide range of applications and is effectively deployed in contact centers, IVR systems, mobile and embedded devices, dictation solutions and assistive applications. In general speech signals are quasi-stationary. For sufficient short period of time(5-100ms), characteristics of the speech signals are fairly stationary. The signal characteristics changes if there is a change in period of time which shows the different speech sounds being spoken. The information contained in speech signal is represented by the short term amplitude spectrum of the speech waveform. This helps us to extract the features. Some of the speech recognition systems require training where an individual speaker reads text into the system. The system analyzes the person's particular voice and uses it to fine-tune the recognition of that person's speech with more accuracy. The speech recognition algorithm can classified in two types. One is speaker dependent and other one is speaker independent. The speaker dependent system focuses on developing a system to recognize unique voiceprint of individuals. Speaker independent system involves identifying the word uttered by the speaker. It is further divided as isolated word detection and continuous speech recognition. Input for isolated word detection is single words separated by pauses. This is simple compared to continuous speech recognition as the system doesn't need to learn sequence of dictionary words. The basic difficulty of speech recognition is that the speech signal is highly variable due to different speakers, speaking rates, contents and acoustic conditions.

Feature analysis component of speech recognition system plays a crucial role in the overall performance of the system. Many feature extraction techniques are available, these include

1. Linear Predictive Analysis (LPC)
2. Linear Predictive Cepstral Coefficients (LPCC)
3. Perceptual Linear Predictive Coefficients (PLP)
4. Mel-Frequency Cepstral Coefficients (MFCC)
5. Power Spectral Analysis (FFT)
6. Mel -scale Cepstral Analysis (MEL)
7. Relative spectra Filtering of log domain Coefficients (RASTA)
8. First Order Derivative (DELTA)

2. LITERATURE REVIEW

Various methodologies have been proposed for isolated word detection and continuous speech recognition over the years. Out of which Hidden Markov models (HMM) has been extensively used in lot of speech recognition applications because of its high reliability [1]. Artificial Neural Networks (ANN) is another classifier of speech recognition with acceptable accuracy. For simple isolated word detection MFCC and DTW approach is enough and efficient even for the implementation of speech recognition engine in embedded systems MFCC and DTW algorithms are proved to be simpler enough compared with neural networks and HMM.

3. METHODOLOGY

A. Recognition Module:

Isolated word detection involves two digital signal processes which are Feature Extraction and Feature Matching. Feature

extraction involves calculation of MFCCs for each frame. MFCCs are the coefficients that are collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear MEL scale of frequency. DTW is one of the algorithms used for measuring the similarity between two temporal sequences, which may vary in speeds. This method is used for Feature Matching.

B. Feature Extraction:

Speech recognition system starts with a pre-processing stage where, the speech waveform is taken as its input, and feature vectors are extracted from it which represents the information required to perform recognition. This stage is performed by software efficiently. Different operations are performed on the signal such as pre-emphasis, framing, windowing, and Mel-Cepstrum analysis. MFCC is used to extract features from the speech signal. It is based on the human peripheral auditorium system.

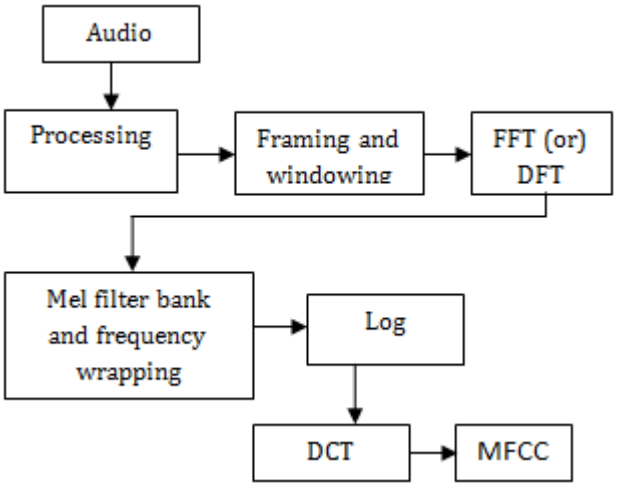


Fig.1 MFCC Block Diagram

Pre-emphasis:

Noise has a greater effect on the higher modulating frequencies than the lower ones. Hence, the higher frequencies are artificially boosted to increase the SNR. In MFCC extraction process firstly the input speech signal is pre-emphasized to enhance the high frequency part of the signal at the time of speech generation. Pre-emphasis process performs spectral flattening using a first order finite impulse response (FIR) filter. The speech signal s(n) is sent to a high-pass filter:

$$s2(n) = s(n) - a*s(n-1) \quad (1)$$

where s2 (n) is the output signal, and the value of a is usually between 0.9 and 1.0.

Framing:

The pre-emphasized speech wave is non-stationary and is getting divided into number of frames to analyze the speech where it supposed to be stationary speech signal. The frame length should not be too short and too long. If frame length is short then we don't get enough samples and for long length the signal changes.

Windowing:

Windowing of a signal is done to eliminate the discontinuities at the edges of the frames. Windowing is used to reduce the spectral effects, smooth's the signal for computation of the FFT[3]. Overlapping is used to produce continuity within the frames. Hamming, Hanning, Rectangular, Triangular etc., windows types are used for windowing of speech signal. If, the windowing function is defined as w(n), 0<n where, N is the number of samples in each frame, then the resulting signal will be; y(n)=x(n) w(n). Mathematically, framing is basically equivalent to multiplying the signal with a series of sliding rectangular windows. However, the use of rectangular windows may give rise to spectral leakage because the power contained in the side lobes is significantly higher. Hamming window is most used window shape in speech recognition technology because high resolution is not required. To avoid this, we have used a Hamming Window which has the form :

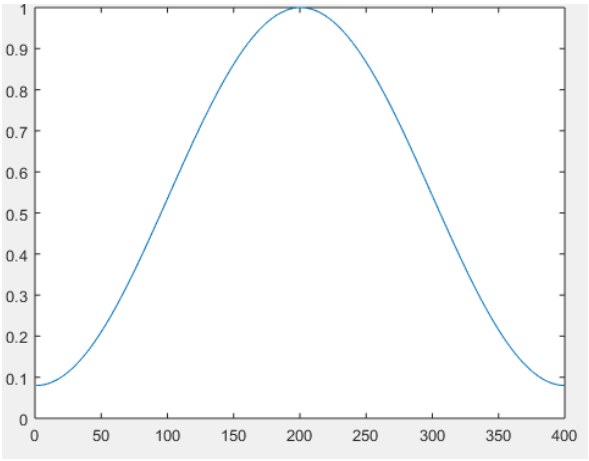


Fig.2 Hamming Window

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

Fast Fourier Transform (FFT):

A Fast Fourier transform (FFT) is an algorithm that samples a signal over a period of time (or space) and divides it into its frequency components. An FFT algorithm computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IFFT). Fourier analysis converts a signal from its original domain to a representation in the frequency domain and vice versa. An FFT rapidly computes such

transformations by factorizing the DFT matrix into a product of sparse (mostly zero) factors. As a result, a result it manages to reduce the complexity of computing the DFT [4].

$$X_n = \sum_{K=0}^{N-1} x_k \; e^{\frac{-2\pi jkn}{N}} \quad , n = 0, 1, 2 \ldots . N-1 \qquad (3)$$

Mel-Scale:

In this step, the magnitude spectrums calculated above are mapped on the Mel scale to know the approximation about the existing energy at each spot with the use of Triangular overlapping window which is also known as triangular filterbank . A Mel scale mapping has to be done between the given real frequencies (Hz) and the required frequency scale (Mels). The filter bank is used to transform the spectrum of a signal into a representation which more closely reflects the behavior of the human ear. Mel scale-Pitch is denoted in Mel scale[2]. This scale approximates to the human hearing system. As, human can easily distinguish the minor variations in pitch at low frequencies than at high frequencies.

Formula from Frequency to Mel scale:

$$M(f) = 1125 \ln(1 + f/700) \qquad (4)$$

Formula from Mel scale to Frequency:

$$M-1 = 700(\exp(m/1125)) - 1 \qquad (5)$$

Cepstrum:

Cepstrum can be obtained by the cosine transformation of the log of the unwrapped phase of Fourier transform. Once we have filter energies, we have to take the logarithm of them. This step simply converts the multiplication of the magnitude of the Fourier transform into addition referred to as signal's logarithm Mel spectrum.

Logarithm is used for wide range.

IDFT: –

It is applied to the log spectral energy vector resulting in the group of Mel frequency Cepstral coefficients. DCT is mostly used due to its energy compaction, which results in its coefficients being more concentrated at lower indices than the DFT. This property allows approximating the speech signal with fewer coefficients [3]. The Mel filter bank output is given to IDCT by Logarithm compression which results in the group of coefficients called Mel frequency Cepstral coefficients. The formula for DCT is,

$$C_i = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} E_m \; cos\left(\frac{\pi i}{M}\left(m - \frac{1}{2}\right)\right) \qquad (6)$$

Where

i =1,2,3,4……..M

M is the number of Mel-scale Cepstral coefficients.

C. Feature Matching (DTW):

The features of the test samples and reference samples which are calculated in the previous stages are compared with each other.

$$PSNR = 10 \log 10^{\frac{MAX^2}{MSE}} \qquad (7)$$

Where, MAX is the maximum possible power of image and MSE is the mean square error.

The similarity degree between the two series is generally represented by defining a specific distance between two series called similarity distance.

Minkowski distance is one of the method use frequently and is defined as follows:

$$d(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \qquad (8)$$

When p=2, the distance between two series is called Euclidean Distance.

The problem with Minkowski method for complex series is to maintain the two series of same length, point to point correspondence and weight of each pair difference is equal. To solve this problem the Dynamic Time Warping (DTW) method is often used. Any data that can be turned into a linear sequence can be analyzed with DTW [2].

Distance is designated to draw the greatest similarity between series by calculating the minimum distance between them, which is defined as follows.

Let X(x1,x2,…..xn) and Y(y1,y2,…..ym) be two series with the length of n and m , respectively, and n*m an matrix M can be defined to represent the point-to-point correspondence relationship between X and Y , where the element Mij indicates the distance d(xi,yi) between xi and yj. Then the point-to-point alignment and matching relationship between X and Y can be represented by a time warping path W=(w1,w2,…..wk), max(m,n)<=K<m+n-1, where the element $w_k = (i, j)$ indicates the alignment and matching

relationship between xi and yj. If a path is the lowest cost path between two series, the corresponding dynamic time warping distance is required to meet

$$DTW(X,Y) = \min_{w}\{\sum_{k=1}^{K} d_k, W = \langle w1, w2, \ldots . wk\rangle\} \quad (9)$$

Where $d_k = d(x_i, y_j)$ indicates the distance represented as $w_k = (i, j)$ on the path W .

Then the formal definition of dynamic time warping distance between two series is described as

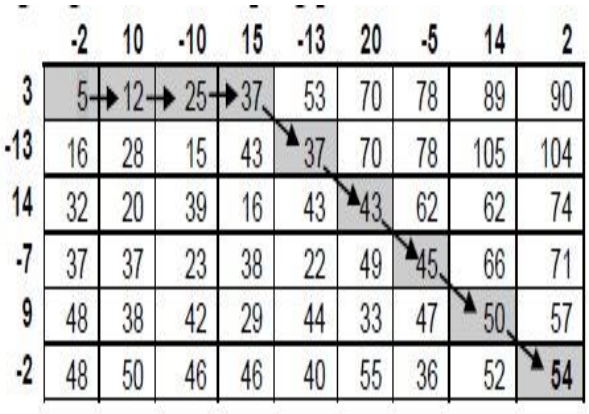
$$DTW(<>,<>) = 0; \quad DTW(X,\langle\ \rangle) = DTW(\langle\ \rangle, Y) = \infty \quad (10)$$

$$DTW(X,Y) = d(x_i - y_j) + \min\begin{cases} DTW(X, Y[2:-]); \\ DTW(X[2:-], Y); \\ DTW(X[2:-], Y[2:-]); \end{cases} \quad (11)$$

where < > indicates empty series [2:-], indicates a sub array whose elements include the second element to the final element in an one-dimension array d(xi,yj), indicates the distance between points xi and yj which can be represented by the different distance measurements, for example, Euclidean Distance

.

Dynamic Time Warping for two voice samples is illustrated in below figure.



The Minimum distance between the samples is given by:

$$r(i,j) = d(x_i, y_j) + min\{r(i-1,j), r(i,j-1), r(i-1,j-1)\} \quad (12)$$

3. RESULTS

The distances while comparing similar words and different words using different windows is shown in table 1. With similar words the distance is below 100 using all windows. As DTW calculate possible alignment between two vector paths, when two same sequences are compared, the obtained distance should be 0.

Table 1: Comparison of Response of different windows in MFCC

S. no	Name of the Window	Min distance between the reference sample and test sample			
		Speech1	Speech2	Speech3	Speech4
1.	Hamming	22.1434	195.4865	322.4250	174.4436
2.	Hanning	36.0666	183.2792	351.1506	155.9772
3.	Kaiser	32.4864	276.9533	470.0916	226.9273
4.	Triangle	19.6827	197.0051	289.8412	188.4864
5.	Rectangle	35.0730	278.7917	474.4544	226.7722
6.	Bartlett	20.0520	196.4836	290.9326	188.2754

4. CONCLUSION

Isolated word detection system is generated with MFCC and DTW using MATLAB. From the above observation we can conclude that DTW distance between identical words is less than 100 and distance between different words is greater than 280. So setting a threshold of 150 we can filter the word uttered by the user from the other words.

REFERENCES

[1] B.P. Das, R. Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research, Vol. 2, No. 3,June 2012, pp. 854-858.

[2] L. Muda, M. Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, No. 3,March 2010, pp. 138-143.

[3] "MFCC and its applications in speaker recognition" Vibha Tiwari, Dept. of Electronics Eng., Gyan Ganga Institute of Technology and Management, Bhopal, (MP) INDIA (Received 5 Nov., 2009, Accepted 10 Feb., 2010).

[4] Molau, Sriko, Michael Pitz, Ralf Schluter, and Herman Ney. "Computing Mel Frequency Cepstral Coeeficients on the power spectrum. In Acoustics, speech, and signal processing, proceedings. (ICASSP'01), IEEE International Conference on , Vol.1,pp.73-76.IEEE,2001.

[5] E. Karpov, "Real Time Speaker Identification," Master`s Thesis, Department of Computer Science, University of Joensuu, 2003.