ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Isolated Speech Recognition Using MFCC and DTW

P.P.S.Subhashini
Associate Professor,
Department of ECE,
RVR & JC College of Engineering.

Dr.M.Satya Sairam
Professor & HOD,
Department of ECE,
Chalapathi Institute of Engineering and Technology.

Dr. D.Srinivasa Rao
Professor,
Department of ECE,
JNTUH, Hyderabad.

ABSTRACT

This paper describes an approach of isolated speech recognition by using the Mel-Scale Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW). Several features are extracted from speech signal of spoken words. An experimental database of total five speakers, speaking 10 digits each is collected under acoustically controlled room is taken. MFCC are extracted from speech signal of spoken words. To cope with different speaking speeds in speech recognition Dynamic Time Warping (DTW) is used. DTW is an algorithm, which is used for measuring similarity between two sequences, which may vary in time or speed.

Keywords: ***MATLAB, Mel frequency cepstral coefficients (MFCC), Speech Recognition, Dynamic Time Warping (DTW).***

INTRODUCTION

SPEECH recognition is the process of automatically recognizing the spoken words of person based on information in speech signal. Each spoken word is created using the phonetic combination of a set of vowel semivowel and consonant speech sound units. The most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. MFCCs are coefficients, which represent audio, based on perception of human auditory systems. The basic difference between the operation of FFT/DCT and the MFCC is that in the MFCC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT.

Due to its advantage of less complexity in implementation of feature extraction algorithm, certain coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database.

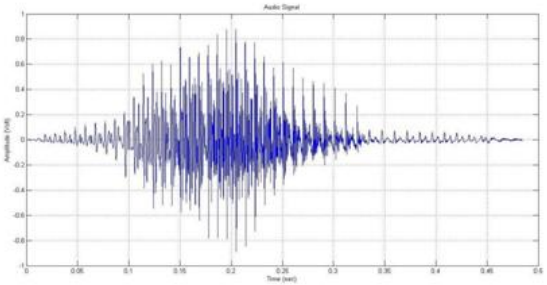

Figure 1. Speech signal representation in Matlab.

Two utterances of the same word by the same user can differ in time. For example, ***two*** can be pronounced as ***to*** or ***too***. DTW resolves this problem by aligning the words properly and calculating the minimum distance between two words. A local distance matrix is formed for all the segments in the sample word and template word.

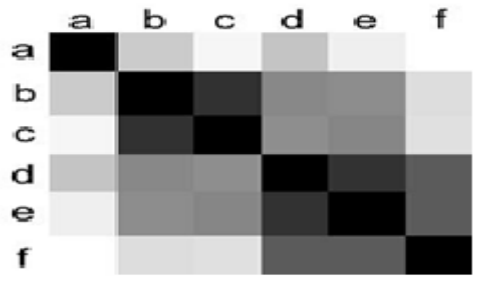

Figure 2. Local Distance Matrix

METHODOLOGY

Feature Extraction

Several feature extraction algorithms can be used to do this task, such as - Linear Predictive Coefficients

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

(LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), and Human Factor Cepstral Coefficient (HFCC). [2] The MFCC algorithm is used to extract the features. The functions used for feature extraction [x_cep, x_E, x_delta, x_acc]. MFCC are chosen for the following reasons:-

- MFCC are the most important features, which are required among various kinds of speech applications.
- It gives high accuracy results for clean speech.
- MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

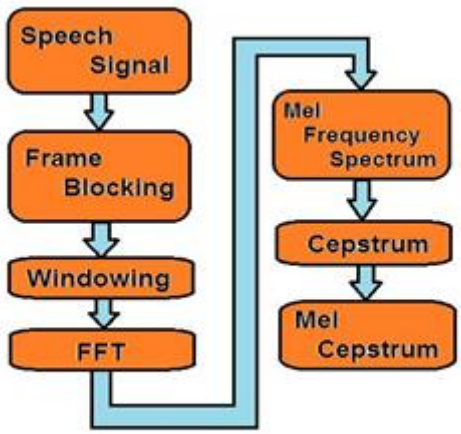

Figure 3. Block diagram for the feature extraction process applying MFCC algorithm

A. Preprocessing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. There are two steps in Pre-processing.

1. Pre-emphasization.
2. Voice Activation Detection (VAD).

1. Pre-emphasization

The digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range and spectrally flatten the speech signal, pre-emphasis is applied. First order high pass FIR filter is used to pre-emphasize the higher frequency components.

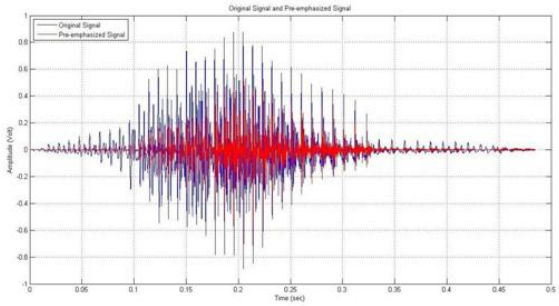

Figure 4. Pre-Emphasized Speech Signal

2. Voice Activation Detection (VAD)

VAD facilitates speech processing, and it is used to deactivate some processes during non-speech section of an audio sample. The speech sample is divided into non-overlapping blocks of 20ms. It differentiates the voice with silence and the voice without silence.

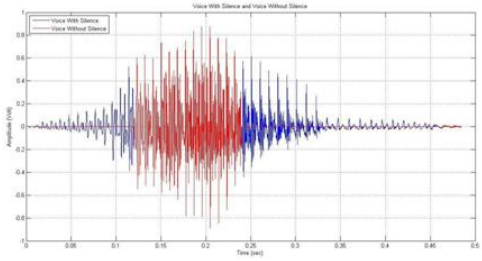

Figure 5. VAD Speech Signal

B. Frame Blocking

The speech signal is split into several frames such that each frame can be analysed in the short time instead of analysing the entire signal at once. The frame size is of the range 0-20 ms. Then overlapping is applied to frames. Overlapping is done because on each individual frame, hamming window is applied. Hamming window gets rid of some of the information at the beginning and end of each frame. Overlapping reincorporates this information back into our extracted features.

C. Windowing

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

the underlying spectrum. The choice of the window is a tradeoff between several factors. In speech recognition, the most commonly used window shape is the hamming window.

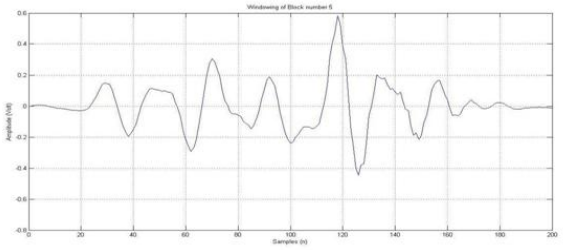

Figure 6. Windowing of the Speech Signal

D. Fast Fourier Transform

The basis of performing fast Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain [5]. Spectral analysis signify that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore, FFT is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage i.e. Mel Frequency Warping.

E. Mel-frequency warping

Human ear perception of frequency contents of sounds for speech signal does not follow a linear scale. Therefore, for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the „mel" scale.

The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz. To compute the mels for a given frequency f in Hz, a the following approximate formula is used.

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

The subjective spectrum is simulated with the use of a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

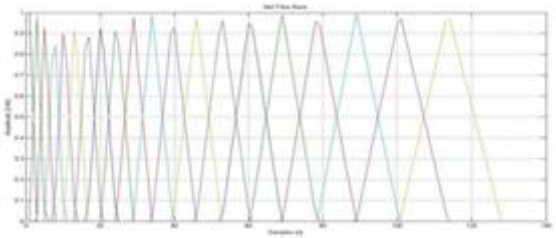

Figure 7. Mel Filter Banks

F. Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). By doing DCT, the contribution of the pitch is removed. In this final step Log Mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The discrete cosine transform is done for transforming the mel coefficients back to time domain.

$$\tilde{c_n} = \sum_{k=1}^{K} (\log \tilde{S}_k) \left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K} \right]$$

where n=1,2,....K

Whereas S_k, K = 1, 2, … K are the outputs of last step.

Feature Matching

There are many feature-matching techniques used in speaker recognition such as Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization. DTW technique is used for feature matching.

Dynamic Time Warping (DTW)

The time alignment of different utterances is the core problem for distance measurement in speech recognition. A small shift leads to incorrect identification. Dynamic Time Warping is an efficient method to solve the time alignment problem. DTW algorithm aims at aligning two sequences of feature

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

vectors by warping the time axis repetitively until an optimal match between the two sequences is found. This algorithm performs a piece wise linear mapping of the time axis to align both the signals.

Consider two sequences of feature vector in an n-dimensional space.

$x = [x_1, x_2, ..., x_n]$ and $y = [y_1, y_2, ..., y_n]$

The two sequences are aligned on the sides of a grid, with one on the top and other on the left hand side. Both sequences start on the bottom left of the grid.

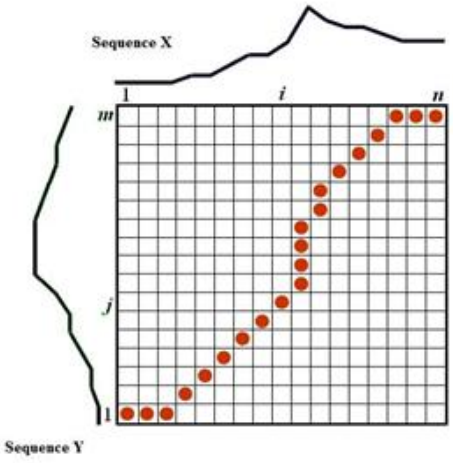

Figure 8. Global Distance Grid

In each cell, a distance measure is placed, comparing the corresponding elements of the two sequences. The distance between the two points is calculated via the Euclidean distance.

$$\text{Dist}(x, y) = |x - y| = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2]^{1/2}$$

The best match or alignment between these two sequences is the path through the grid, which minimizes the total distance between them, which is termed as Global distance. The overall distance (Global distance) is calculated by finding and going through all the possible routes through the grid, each one compute the overall distance.

The *global distance* is the minimum of the sum of the distances (Euclidean distance) between the individual elements on the path divided by the sum of the weighting function. For any considerably long sequences the number of possible paths through the grid will be very large. Global distance measure is obtained using a recursive formula.

$$GD_{xy} = LD_{xy} + \min (GD_{x-1\ y-1}, GD_{x-1\ y}, GD_{x\ y-1})$$

Here, GD = Global Distance (overall distance)

LD = Local Distance (Euclidean distance)

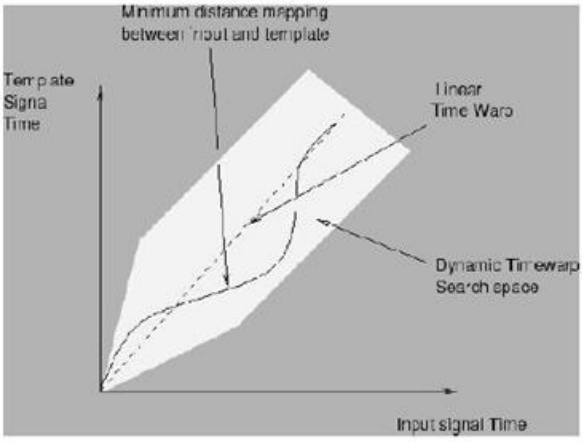

Figure 9. Dynamic Time Warping

CONCLUSION

The main aim of this project was to recognize isolated speech using MFCC and DTW techniques. The feature extraction was done using Mel Frequency Cepstral Coefficients {MFCC} and the feature matching was done with the help of Dynamic Time Warping (DTW) technique. The extracted features were stored in a .mat file using MFCC algorithm. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speech signal with the speech signal database. The experimental results were analysed with the help of MATLAB and it is proved that the results are efficient. This process can be extended for n number of speakers. The project shows that the DTW is the best nonlinear feature matching technique in speech identification, with minimal error rates and fast computing speed. DTW will receive the utmost importance for speech recognition in voice based Automatic Teller Machine.

REFERENCES

[1] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk "Speech Recognition using MFCC" International Conference on Computer

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)

[2] http://www.springerlink.com/content/n1fxnn5gpkuelu9k.

[3] B. Gold and N. Morgan, Speech and Audio Signal Processing, John Wiley and Sons, New York, NY, 2000.

[4] C. Becchetti and Lucio Prina Ricotti, Speech Recognition, John Wiley and Sons, England, 1999.

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

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