International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Performance Evolution of Combine Approach of MFCC & DTW Technique for Speech Recognition

Nilesh Patel[1], D. G. Agrawal[2]

[1]Shri Sant Gadge Baba College of Engineering & Technology, Bhusawal, Maharashtra, India

[2]Professor, Shri Sant Gadge Baba College of Engineering & Technology, Bhusawal, Maharashtra, India

Abstract: ***This paper proposes on speech recognition for English word using Mel Frequency Cepstral Coefficients (MFCCs) and Dynamic Time Warping (DTW) introduced by Sakoe Chiba. MFCC are the coefficients which collectively represent the short-term power spectrum of a sound. That Power spectrum based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Start point, Short time Energy and Zero crossing detector and end point detection using Mel Frequency Cepstral Coefficients and Dynamic Time Warping algorithms used for recognition of samples English word which would be by different male and female speakers. The algorithm is test for sample speech signal which are recorded. This system is used for different application for like security system and also for controlling of robot. The result shows that accuracy is all different English word is above 61.66% and also shows a comparison for different method applied for MFCC.***

Keywords: STE, ZCR, MFCC, DTW

1. Introduction

One of the most common ways of communication is speech for Human. Speech recognition is a technique to collect the speech information in computer and take action on it. It is also called as Automatic Speech Recognition, Speech to text converter. From the many decades of the years there are many technique used for the speech recognition like Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Artificial Neural Network (ANN) and etc. The main application of speech recognition is to speech to text converter, automatic data entry, voice recognition password etc.

Speech recognition done by two techniques, Feature extraction and Feature matching. In this propose system feature extraction is done using MFCC (Mel Frequency Cepstral Coefficients) and feature matching is done using DTW (Dynamic Time Warping).Voice is collected by the microphone and convert in analog signal to digital signal. Now this signal is ready to process on it.

2. Proposed Approach

In this paper, proposed an approach to recognition isolated English word automatically from input audio signal generated by different male and female speaker in controlled environment. It uses a combination of features based on Short Time Energy (STE), Zero Crossing Rate (ZCR), Start point End point detection, Mel Frequency Cepstral Coefficient (MFCC) [1]. DTW is used as a feature extraction to detect nearest word which is recorded in database.

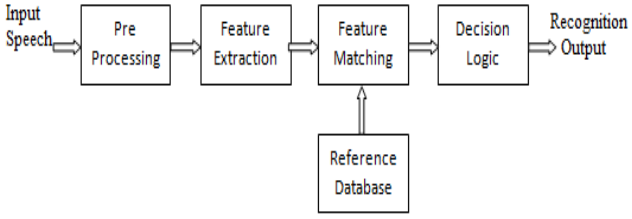

Figure 1: Basic block diagram of speech recognition system

3. Pre Processing

Each signal generated by individual is represented as a collection of sample value. Before extracting the relevant features, the voice signal should be pre processed to eliminate the problems that may arise in feature extraction [2]. It consists of Short Time Energy (STE), Zero Crossing Rate (ZCR) and Start point End point detection.

3.1 STE (Short Time Energy*)*

The energy content set of sample which is generated by speaker is sum of sample square of the sample. To calculate STE the speech signal is sampled using a rectangular window function of width ω samples, where ω << n. Within each window, energy e is computed as follows [1]:

$$e = \sum_{i=1}^{\omega} X_i^2 \quad (1)$$

The energy for each window is collected to generate the STE feature vector having $W = \frac{n}{\omega}$ element [1]

$$E = \bigcup_{j=1}^{w} e_j \quad (2)$$

Licensed Under Creative Commons Attribution CC BY

3.2 ZCR (Zero Crossing Rate)

ZCR of an audio signal is a consistent of the number of times the signal crosses the zero amplitude line by passage from a positive to negative or vice versa[1]. The Speech signal is divided into temporal segments by the rectangular window function as represented above and zero crossing rate for each segment is computed as given below, where *sgn*(x_i) indicates the sign of the i^{th} sample. It has three possible values depending on whether the sample is positive, zero or negative: -1, 0 and 1.

$$z = \sum_{i=1}^{\omega} \frac{|\,\mathrm{sgn}(x_i) - \mathrm{sgn}(x_{i-1})\,|}{2} \quad (3)$$

The value for each window is collected to generate the ZCR feature vector having W=n/w element [1].

$$z = \bigcup_{j=1}^{w} z_j \quad (4)$$

3.3 Start point end point detection

To making noise free and better speech signal as compared to generated speech signal this computation of these point is much beneficial. On the base of the STE and ZCR value, it gives the starting location of the speech value. So that previous unwanted sample would be removed and new speech signal is generated. Same procedure is applying for the end point detection.

3.4 Start Point End Point Detection based on ZCR

Threshold value of signal is found by several observations on speech signal. Some start point is found by the end point detection and end of the signal is checked for the zero crossing rate. Start point and end point is changed after the comparing the threshold value with zero crossing rate on particular part of frame. This is done according to the following conditions [6]

- If ZCR > 3*(threshold), then start-point shifts one frame left.
- If ZCR > 3*(threshold), then end-point shifts one frame right, provided that the previous end-point is not in the last frame.

4. Feature Extraction

Many techniques available for feature extraction like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) etc. MFCC is most popular technique for feature extraction in speech recognition. This may be attributed because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better [4]. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [5]. There are two type of filter in MFCC , spaced linearly for low frequencies below 1KHz and logarithmic spaced above the 1KHz. Block diagram of MFCC is given below figure 2 [5].

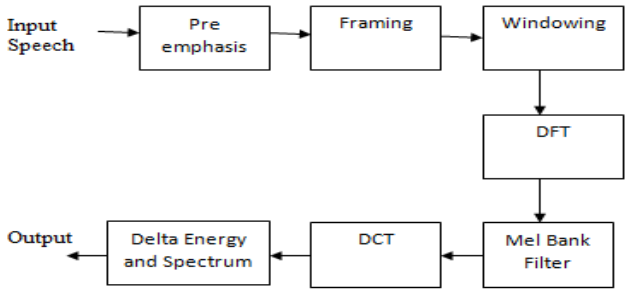

Figure 2: Block diagram of MFCC

4.1 Pre emphasis

In this process a filter passed signal which have emphasizes higher frequencies. This step increases the energy at higher frequencies.

$$Y[n] = X[n] - 0.95X[n-1] \quad (5)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample [5].

4.2 Framing

The technique of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec [5]. The speech signal is divided into the frame; each has size of N samples. Adjacent frames are being separated by M (M<N).

4.3 Windowing

For windowing use the hamming windowing technique. Window as a shape by considering the next block in feature extraction processing chain and incorporates all the closest frequency lines. Hamming window equation is given below.
If the window is defined as W (n), $0 \le n \le N-1$, Where N = number of samples in each frame, Y[n] = Output signal, X (n) = input signal, W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n) \quad (6)$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{n-1}\right) \quad (7)$$

Where $0 \le n \le N-1$

4.4 Fast Fourier transform

Each frame of N samples convert from time domain into frequency domain. The Fourier Transform is to translate the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain [5]. This statement footing the equation below:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \quad (8)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

Licensed Under Creative Commons Attribution CC BY

4.5 Mel filter bank

FFT spectrum has very wide frequency range and speech signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 3 is then performed [5].

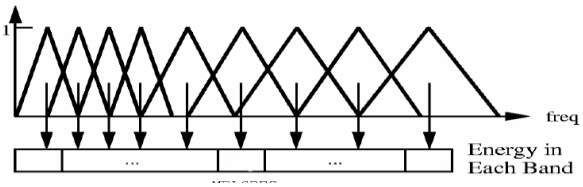

Figure 3: Mel Scale filter bank

As per shown in above figure a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. 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 [5]. Output of each filter is sum of its filtered spectral component. Equation of the compute the Mel for given frequency f is given below.

$$F(Mel) = [2595 * \log 10[1 + f]700] \quad (9)$$

4.6 Discrete cosine transform

In this process the log Mel spectrum convert into time domain using Discrete Cosine Transform (DCT). The result of reconstruction is called Mel-Frequency Cepstrum Coefficient. The collection of coefficient is called acoustic vectors. So that, each input utterance is transformed into a sequence of acoustic vector [5].

4.6 Delta energy and delta spectrum

Speech signal is varying with the time so frame is also change with the time, so there is a need to add features related to the change in cepstral features over time. For this energy and spectrum features computed over on small interval of frame of speech signal. Mathematically, the energy within a frame for a signal x in a window from time sample t1 to time sample t2, is constituted as [5].

$$Energy = \sum x^2[t] \quad (10)$$

5. Feature Matching

Input Speech signal is comparing with the reference database signal for the recognition of the word using the pattern recognition. So we created the reference database of speech signal.

5.1 Reference Database

Record the words from the different male and female speaker and create the speech database for the comparing with test speech signal for the recognition of the word. Each recorded word has become reference pattern to keep for the matching process.

5.2 Recognition using DTW

DTW algorithm used for the comparing the test signal and reference pattern available in database. Dynamic time warping algorithm works on measuring similarities between two time series which is vary in speed and time. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis [5]. This warping can used to find corresponding regions or similarities between two time series. Below figure 4 shows example of the warping of the two time series

.

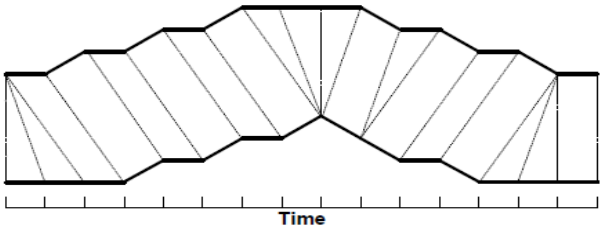

Figure 4: A warping between two series [5]

DTW contrast two dynamic patterns and evaluates similarity by calculating a minimum distance between them [4]. Let us take two time domain series Q and C which has length of n and m respectively.

$$Q = q1, q2...., qi....qn \quad (11)$$

$$C = c1, c2...., cj....cm \quad (12)$$

To align two series using DTW, an *n*-by-*m* matrix where the (i^{th}, j^{th}) element of the matrix contains the distance (*q*i, *cj*) between the two points *q*i and *c*j is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation [5].

$$d(qi, cj) = (qi - cj)^2 \quad (13)$$

Each matrix element (i, j) corresponds to the alignment between the points qi and cj. Then, accumulated distance is obtained by [5].

$$D(i, j) = \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \quad (14)$$

5.3 Decision logic

DTW algorithm matched the input speech and database speech signal using maximum correlation and minimum MFCC distance.

6. Methodology

Voice signal can be varied during training and testing session due to some factors like voice varies with time, health of the individual, rate of speaking, acoustical noise etc. Table gives

Licensed Under Creative Commons Attribution CC BY

details of recording and training session of speech recognition system.

Table 1: Recording Requirement

Process	*Descriptions*
Speaker	Male and Female Speaker
Tools	Mono Microphone
Sampling Frequency	8KHz
Feature Computational	12 MFCC coefficients with STE and ZCR
Environment	Laboratory

7. Results

Recognition accuracy is calculated by below equation

$$Recognition Accuracy = \frac{Correctly\ Recognition Word}{Total\ Recognition Word} \times 100 \quad (15)$$

Table 2: Recognition Accuracy

Session	*Accuracy*
1	61%
2	60%
3	64%
Average	61.66%

8. Conclusions

Speech recognition system works on information contain in the speech signal, DTW technique is extract that information and give the effective result. The recognition accuracy obtains in [1], using Euclidian Distance was 57.5% and proposed scheme suggest the recognition accuracy to be 61.66%. Improvement of the result is due to find accurate start point and end point detection using STE and ZCR. This is because energy finding removes the noise and silent period present in the signal and zero-crossing is used to detect the weak fricatives and weak plosives [7]. This system is also applied to ANN classifier for the improving accuracy.

References

[1] BP Das and Ranjan Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research (IJMER) , Vol.2, Issue.3, May-June 2012.
[2] P. G. N. Priyadarshani, N. G. J. Dias, Amal Punchihewa, "Dynamic Time Warping Based Speech Recognition for Isolated Sinhala Words", Circuit and System(MWSCAS), IEEE, Aug-2012
[3] SJ.Arora and RP.Singh, "Automatic Speech Recognition: A Review, "International Journal of Computer Applications, vol 60-No.9, December 2012.
[4] Nidhi Srivastava and Dr.Harsh Dev "Speech Recognition using MFCC and Neural Networks", International Journal of Modern Engineering Research (IJMER), March 2007.
[5] Lindasalwa Muda, Mumtaj Begam and I.Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques ", Journal of Computing, Volume 2, Issue3, March 2010.
[6] Digital Processing of Speech Signals by Lawrence R. Rabinar and Ronald W.Schafer.
[7] Speech coding algorithms Foundation and Evolution of Standardized Coders (wiley) by Wai C. Chu.
[8] Sahil Verma, Tarun Gulati and Rohit Lamba, "Recognizing Voice For Numerics Using MFCC And DTW", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 5, May 2013.
[9] Elena Tsiporkova, "Dynamic Time Warping Algorithm for Gene Expression Time Series".
[10] Jan Cernocky, "Speech Recognition – Introduction and DTW".

Licensed Under Creative Commons Attribution CC BY