ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 5, May 2014

Recognizing voice commands for robot using MFCC and DTW

Nidhi Desai [1], Prof.Kinnal Dhameliya[2], Prof. Vijayendra Desai[3]

M.Tech. Research Student, Department Of Electronics and Communication Engineering, Chhotubhai Gopalbhai Patel Institute of Technology (CGPIT) Bardoli, India[1]

Assistant Professor, Department Of Electronics and Communication Engineering, Chhotubhai Gopalbhai Patel Institute of Technology (CGPIT) Bardoli, India[2]

Assistant Professor, Department Of Electronics and Communication Engineering, C.K.P.C.E.T., Surat, India[3]

Abstract: This paper proposes an approach to recognize English words corresponding to control Robot in an isolated way by different male and female speakers. The aim is to focuses on recognizing voice using Mel-frequency cepstral coefficients (MFCCs) and Dynamic Time Warping (DTW) introduced by Sakoe Chiba [3]. MFCC are the coefficients that collectively represent the short-term power spectrum of a sound, deploy on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Computation of Short Time Energy (STE), Zero Crossing Rate (ZCR), start point and endpoint detection, Mel Frequency Cepstral Coefficient (MFCC) and DTW algorithm are used to process speech samples to accomplish the recognition. The algorithm is tested on speech samples. The system is then applied to recognition of isolated word in English language that is used to control robot for specify application. The algorithm is tested on speech samples that are recorded. The results show that the algorithm conducted to recognize almost 75.19% of all recorded words using four different methods applied on MFCC computation and likewise their comparison is observed.

Keywords: DTW, Start point, End point, MFCC, Recognition Accuracy, STE, ZCR

I. INTRODUCTION

Speech Recognition is the ability of machine or program to identify words and phrase from spoken language and convert them in to machine readable format. It is also known as Automatic Speech Recognition or computer speech recognition and speech to text conversion. The main intension of speech recognition area is to evolve techniques and system for speech input to machine. Several methods such as Liner Predictive Coding (LPC), Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Artificial Neural Network (ANN) and etc are evaluated with a view to identify a straight forward and effective method for voice signal.

This paper propose an effective method for voice recognition using MFCC(mel frequency cepstral coefficient) are utilize as feature extraction and Dynamic Time Warping(DTW) as feature matching technique that detect the nearest recorded word. Firstly, human voice is converted into digital signal form to produce digital data representing each level of signal at every discrete time step.

After that digitized Speech samples are processed using combination of features like STE, ZCR, start point, end point and MFCC to produce voice features. After that, these voice features can go through DTW to select the pattern that matches the database and find correlation between each reference database and test input file in order to minimize the resulting error between them on MFCC feature. WER or recognition accuracy is computed to measure the performance of the system. Figure 1 shows the overall process of speech recognition system.

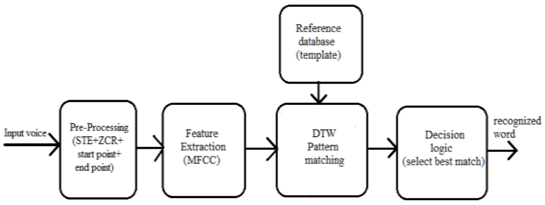
Reference database (template)
Input voice
Pre-Processing (STE+ZCR+ start point+ end point)
Feature Extraction (MFCC)
DTW Pattern matching
Decision logic (select best match)
recognized word

Fig.1 Block diagram of Speech Recognition System

The remaining of this paper is organized as follows: proposed approach in section 2, feature extraction (MFCC) in section 3, feature matching (DTW) in section 4, methodology in section 5 and finally results and discussion with conclusion in section 6.

II. PROPOSED APPROACH

This paper proposes an approach to recognize automatically English words from audio signals generated by different individuals in a 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). A Dynamic Time warping (DTW) is used to detect the nearest recorded word from database.

A. Short Time Energy(STE)

The energy content of a set of samples is related by the sum of the square of the samples. To calculate STE the speech signal is sampled using a rectangular window function of width ω samples, where $\omega << n$. Within each window, energy is computed as follows [7]:

$$e=\sum_{i=1}^{\omega}{}^{2}_{x}i \qquad (1)$$

B. *Zero Crossing Rate(ZCR)*

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 [7]. The audio signal is divided into temporal segments by the rectangular window function as represented above and zero crossing rate for each segment is computed as below, where *sgn(xi)* indicates the sign of the ith sample and can have three possible values: +1, 0, -1 depending on whether the sample is positive, zero or negative[7].

$$z=\sum_{i=1}^{\omega}\frac{|\operatorname{sgn}(xi)-\operatorname{sgn}(xi-1)|}{2} \qquad (2)$$

C. *Start Point End Point Detection*

Computation of these points is more beneficial as they are used to remove background noise and made voice signal better than previous signal. Start point of any voice signal provide the exact starting location of voice sample based of STE and ZCR values, so that all previous unwanted samples would be removed and new voice signal would created. This same process is applied to detect End points of any voice signal.

D. *Start Point End Point Detection based on ZCR*

A threshold value by several observations on the signal is found. The part of the signal from start to the start-point found by end-point detection and that from the end point to the end of the signal is checked for the zero-crossing rate. After comparing the zero-crossings with the threshold, the part of the frame is selected and start-point and end-point is changed. This is done according to the following conditions [14]:

- 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.

III. FEATURE EXTRACTION(MFCC)

MFCC is the most axiomatic and popular feature extraction technique for speech recognition. It approximates the human system response more closely than any other system because frequency bands are placed logarithmically here. The overall process of the MFCC is shown in Figure 2.

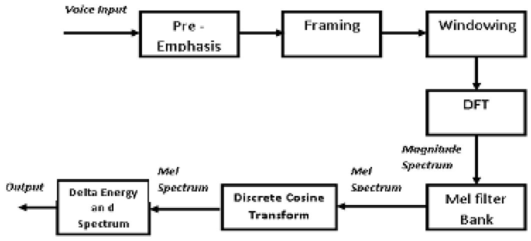

Fig. 2 Steps involve in MFCC Feature extraction [4]

Step-1: Pre-emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies [4]. This process will increase the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\, X[n-1] \qquad (3)$$

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

Step-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 [4]. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step-3: Windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and incorporates all the closest frequency lines. Equation of Hamming window is given as: 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:

$$Y(n)=X(n)\times W(n) \qquad (4)$$

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right)0\le n\le N-1 \qquad (5)$$

Step-4: Fast Fourier Transform

To disciple each frame of N samples 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. This statement footing the equation below:

$$Y(w)=FFT[h(t)*X(t)]=H(w)*X(w) \qquad (6)$$

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

Step-5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very ample and voice signal does not follow the linear scale. The bank of filters corresponding to Mel scale as shown in figure 3 is then performed [4].

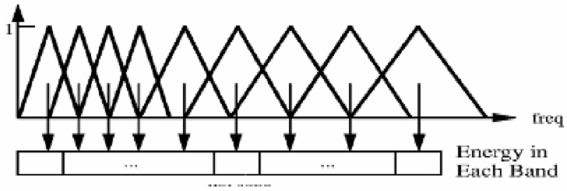

Fig. 3. Mel scale filter bank [4]

This figure depicts 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.

Magnitude frequency response of each filter triangular in shape and equal to unity at the Centre frequency and decline linearly to zero at centre frequency of two adjacent filters [7]. Then, each filter output is the sum of its filtered spectral components. After that the given equation is used to compute the Mel for given frequency f in HZ [4]:

$$\text{Mel}(f) = 2595 * \log10(1 + f/700) \qquad (7)$$

Step 6: Discrete Cosine Transform
This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the 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 [4].

Step-7: Energy and Spectrum:
As speech signals are random, so there is a requirement to add features related to the change in cepstral features over time. For this purpose, here, energy and spectrum features are computed over small interval of frame of speech signals. Mathematically, the energy within a frame for a signal x in a window from time sample t1 to time sample t2, is constituted as [4]:

$$Energy = \sum X^2[t] \qquad (8)$$

IV. DYNAMIC TIME WARPING(DTW)

DTW algorithm is based on measuring similarity among two time series which may vary in time or speed. The similarity is evaluated in terms of alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis [7-9]. This warping between two time series can then be used to find analogues regions between two time series or to determine similarity between the two time series. Mathematically, the DTW contrast two dynamic patterns and evaluates similarity by calculating a minimum distance between them. To realize this, consider two time series Q and C, which has length n and m respectively, where,

Q = q1, q2, qi,.... qn
C = c1, c2, cj,.... cm

To align two sequences using DTW, an n -by- m matrix where the (ith, jth) element of the matrix involves the distance d (qi, cj) among the two points qi and cj is established. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation [4]:

$$d(qi,cj) = (qi - cj)^2 \qquad (9)$$

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

$$D(i,j) = \min(D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \qquad (10)$$

This is done as follows in this paper [10]:

1. Start with the calculation of g (1, 1) = d (1, 1). Calculate the first row g (i, 1) =g (i–1, 1) + d (i, 1).
2. Calculate the first column g (1, j) =g (1, j) +d (1, j).
3. Move to the second row g(i, 2) = min(g(i, 1), g(I 1, 1), g(i – 1, 2)) + d(i, 2). Book keep for each cell the index of this neighbouring cell, which contributes the minimum score (red arrows).
4. Carry on from left to right and from bottom to top with the rest of the grid g (i, j) = min (g (i, j–1), g (i–1, j–1), g (i – 1, j)) + d (i, j).
5. Trace back the best path through the grid starting from g (n, m) and moving towards g (1,1) by following the red arrows. Hence the path which gives minimum distance after testing with the feature vectors stored in the database is the identified speaker.

V. METHODOLOGY

Speech signal during training and testing session can be greatly distinct due to frequent factors such as people voice varies with time, health condition (e.g. the speaker has a cold), speaking rate and acoustical noise and variation of recording habitat via microphone. Table I gives detail intelligence of recording and training session of whole voice recognition system.

TABLE I
TRAINING REQUIREMENT

Process	Description
1) Speaker	Seven male Seven female
2) Tools	Mono Microphone
3) Environment	Laboratory
4) Utterance	Twice each of the following words: Ready Forward Backward Left Right Stop Move
5) Sampling Frequency, Fs	8000Hz
6) Feature computational	12 MFCC coefficients with STE and ZCR

A. *Decision Logic*

The decision is made based on DTW matching technique to select best matching between reference file and test file. There are two criteria based on that decision is taken. These are minimum distance and maximum correlation between two sequences. It is clear from the figure that those reference MFCC vectors selects which have minimum MFCC vector or maximum correlation.

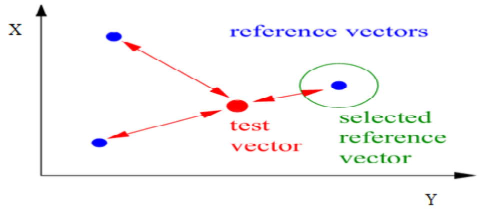

Fig. 4 Decision logic based on minimum distance

VI. RESULT AND DISCUSSION

According to equation 10 Recognition Accuracy [9] is calculated for four sessions which is shown in table 4.1

$$\mathit{Re\,cognitionAccuracy} = \frac{\mathit{Correctly\ Re\,cognizedWord}}{\mathit{Total\ Re\,cognizedWord}} \times 100 \qquad (11)$$

TABLE II
RESULT OF RECOGNITION ACCURACY

Number of Session	**Recognition Accuracy**
1	76%
2	75%
3	71%
4	78.75%
Average	**75.19%**

Proposed system can also be applied for ANN classifier to get higher recognition accuracy. Several other techniques such as Linear Predictive Coding (LPC), Hidden Markov Model (HMM) are currently being investigated.

VII. CONCLUSION

The DTW technique was able to authenticate the particular speaker based on the individual information that was included in the voice signal. The results outlines that this techniques could use effectively for voice recognition purposes. The recognition accuracy obtains in [7], using Euclidian Distance was 57.5% and proposed scheme suggest the recognition accuracy to be 75.19 %. The reason for the improvement is due to accurate start-point and end-point detection by using the combined concept of energy finding and zero-crossing rate. 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 [15]. Thus, combining the above concepts and using DTW and correlation for finding the best match proves to be an effective method for speech recognition that produces considerably better results.

REFERENCES

[1] MA Anusuya and S. K. Katti, “Speech Recognition by Machine,” International Journal of Computer Science and Information security, Vol.6, No.3, 2009J.

[2] SJ.Arora and RP.Singh, “Automatic Speech Recognition: A Review, “International Journal of Computer Applications, vol 60-No.9, December 2012.

[3] SK Gaikwad, Bharti W.Gawali and Pravin Yannawar “A Review on Speech Recognition Technique,” International Journal of Computer Applications, vol 10,No 3, November 2010.

[4] 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.

[5] Nidhi Srivastava and Dr.Harsh Dev “Speech Recognition using MFCC and Neural Networks”, International Journal of Modern Engineering Research (IJMER), March 2007.

[6] Dr.R.L.K.Venkateswarlu, Dr.R.Vasantha Kumari and A.K.V.Nagavya, “Efficient Speech Recognition by Using Modular Neural Network”, International Journal of Computer Technology. Appl., Vol 2 (3)

[7] 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.

[8] Om Prakash Prabhakar and Navneet Kumar Sahu, “A Survey On: Voice Command Recognition Technique,” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013

[9] MU Nemade and Prof. Satish K. Shah, “Survey of Soft Computing based Speech Recognition Techniques for Speech Enhancement in Multimedia Applications”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013.

[10] 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.

[11] MarutiLimkar, RamaRao and VidyaSagvekar, “Isolated Digit Recognition Using MFCC and DTW”, International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012.

[12] Elena Tsiporkova,“Dynamic Time Warping Algorithm for Gene Expression Time Series”.

[13] Jan Cernocky, “Speech Recognition – Introduction and DTW".

[14] Digital Processing of Speech Signals by Lawrence R. Rabinar and Ronald W.Schafer.

[15] Speech Coding algorithms Foundation and Evolution of Standardized Coders (wiley) by Wai C. Chu.