A Noval Approach for Speech Recognition

K.Rajasekhar

Assistant Professor, Department of ECE

BABA Institute of technolgy and sciences

PVJ Rajkumar

Assistant Professor, Department of ECE

BABA Institute of technolgy and sciences

T.Kiran kumar

Assistant Professor , Department of ECE

VIZAG Institute of technolgy

Abstract— This paper describes an approach of software and hardware control using a robust speech recognition system. Speech Recognition is the process of automatically recognizing the spoken words of a person with unique characters. The Software and hardware in this system can be controlled by using isolated speech recognition. The major steps in speech recognition system design are Feature detection, Feature extraction and Feature matching. For Feature detection and extraction, the algorithm used was Mel Frequency Cepstral coefficients (MFCC) and Feature matching was done by using the algorithm called Dynamic Time Wrapping (DTW). The external hardware can be controlled by interfacing with speech recognition system through a programmable logic device like Arduino board. The main difficulty in any recognition system design is large database size, here in this proposed system the database size is reduced by detection of region of interest, other than processing the entire speech signal.

Keywords— MFCC, DTW, PLD, Region of Interest (ROI).

I. INTRODUCTION :SPEECH RECOGNITION SYSTEM:

Speech recognition refers to the study of speech signals and its processing methods. Speech processing usually processed in digital representation. User gives a predefined voice instruction to the system and the system understand this command and execute the required function [1].Most the speech recognition systems are classified as Isolated and Continuous systems. Isolated speech having only one word or utterance where continuous speech allows user to speak naturally. Continuous speech means a continuous utterance without any pause between the utterances [2].Speech Recognition systems can be further classified as Speaker dependent or Speaker independent systems. Speaker dependent systems recognizes speech from only one particular person, where as Speaker independent systems recognizes speech from any person [2]. Speech Recognition is done by training the data base with fixed number of isolated words as a different classes and each class will be trained with unique word with different utterances.

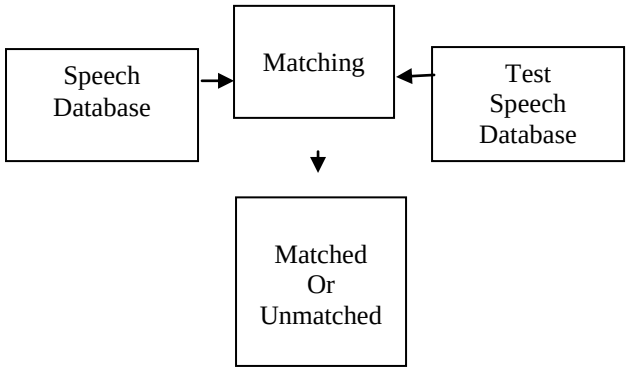


Fig 1: Basic Speech Recognition System

II. IMPORTANCE:

The main purpose of the Speech recognition system is to make the digital systems human friendly.

III. Elements to design speech recognition systems:

A. Feature extraction*:*

Feature extraction was done by train the data base with the limited number of isolated words by using different algorithms and find their features.

B. Feature matching:

Feature matching was done by finding similarities between the trained data base features with the features of test speech by using specified algorithm.

IV. FEATURE EXTRACTION ALGORITHMS:

A. Linear predictive coding coefficients:

LPCC analysis is an effective method to estimate the main parameters of speech signals. The conclusion extracted was that an all-pole filter, H(z), is a good approximation to estimate the speech signals. Its transfer function was described. In this way, from the filter parameters the speech samples could be synthesized by a difference equation. Thus, the speech signals resulting can be seen as linear combination of the previous p samples. Therefore, the speech production model can be often called linear prediction model, or the autoregressive model [3].

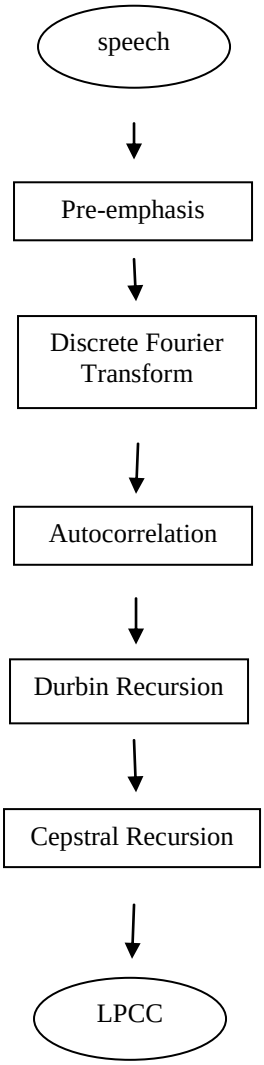


Fig 2: LPC coefficients extraction process

B. Mel Frequency Cepstrum Coefficients:

MFCC is a beneficial approach for speech recognition. Figure 3 illustrates the complete process to extract the MFFC vectors from the speech signal. It is to be emphasized that the process of MFCC extraction is applied over each frame of speech signal independent.

The difference between the cepstrum and the Mel frequency cepstrum is that in the MFC, 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 obtained directly from the FFT or DCT [4].

- Smooth the magnitude spectrum such that the pitch of a speech signals is generally not presented in MFCCs.
- Reduce the size of the features involved.

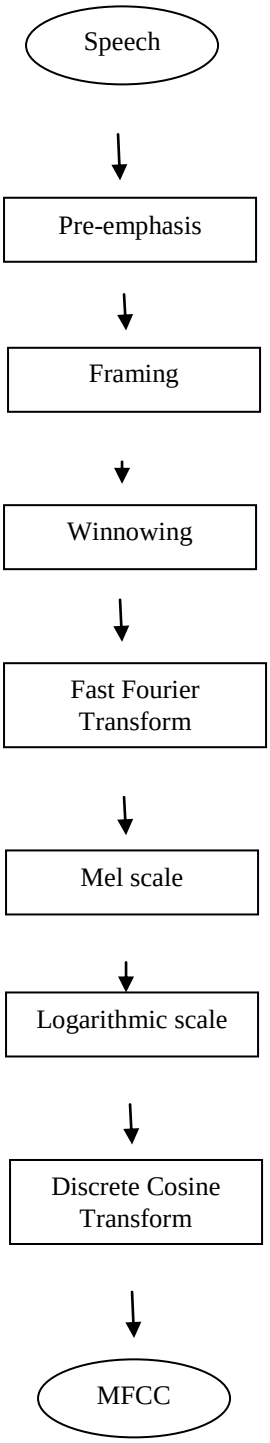


Fig 3: MFC coefficients extraction process

V. FEATURE MATCHING:

A. Linear Time Warping (LTW) *:*

Linear Time Warping is the method of calculating Euclidean distance. Euclidean distance or Euclidean metric means ordinary i.e. straight-line distance between two points in Euclidean space [5]. The process of finding Euclidean distance can be shown as

Distance(D)= (| x-y|)

x= [1,1,4,2,5] & y= [2,1,3,2,2,4,6]

x= 1 1 4 2 5

y= 1 2 3 2 2 4 6

d= 0 1 1 0 3

Here x can be taken as test samples and y can be taken as data samples. Data samples are more than the test samples hence it is difficult to find the distance in this technique. For accurate distance measurement we are approaching DTW technique.

B. Dynamic Time Warping:

Dynamic time warping is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. It is an optimal match optimal match between two sequences with certain restrictions [6]. For example, similarities in walking patterns can be measured by using DTW.

Distance(D)= (| x-y|)

x = (1,1,4,2,5) and y= (1,2,3,2,2,4,6)

6	5	5	2	4	1
4	3	3	0	2	1
2	1	1	2	0	3
2	1	1	2	0	3
3	2	2	1	1	2
2	1	0	3	1	4
1	0	1	2	0	3
	1	1	4	2	5

Fig 4: Tabular representation of Distance between x & y

In the above figure we can find the shortest distance between two samples of different length which cannot be calculated by the Euclidean distance.

VI. Control Design

Here after extracting the features of trained data base of different classes of speech signal matching of features should be done with the features of test speech to find the distance. After finding distance of shortest path software and hardware devices such as computer, mobile and PLD's such as Arduino, FPGA, DSP boards etc.…

VII. Implementation of speech system:

A. Block diagram:

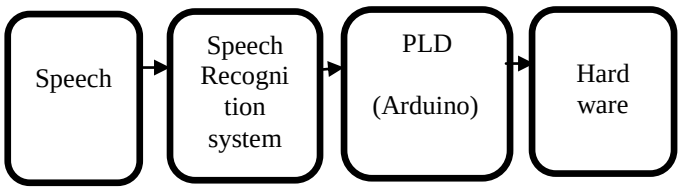


Fig 5: Speech Recognition System

1) Design Aspects:

a) Type of Input Speech.

Continuous Speech.

Continuous speech allows user to speak naturally. Continuous speech means a continuous utterance without any pause between the utterances.

Isolated Speech.

Isolated speech means having only one utterance

b) Separation of voiced and unvoiced portions.

By assigning a threshold value to the input isolated speech we can separate the voiced and unvoiced portions. After assigning the envelop value the voiced and unvoiced portions can be separated by finding the absolute value or magnitude value of the input isolated speech. The value less than the threshold can be treated as unvoiced portion and the signal value above the threshold can be treated as voiced portion of the signal.

c) Feature extraction.

Feature extraction can be extracted by using MFCC algorithm, this can be implemented by the following steps

Pre emphasis:

The digitalized speech has a dynamic range and suffering from additive noise. In order reduce this range and spectrally flatten the speech signal, pre emphasis is applied.

Frame blocking:

Audio signals are continuously changing so the speech signal can be split into small frames such that each frame can be analyzed in short time instead of analyzing the entire signal at once. The frame size is of 0-20 ms. Overlapping is applied on each frame because windowing applied on each frame.

Windowing:

Windowing is used to avoid discontinuities in the speech signal and distortion in underlying spectrum [7]. In speech recognition the most commonly used window shape is the hamming window. The choice window is the trade of between different factors [8].

$$W(n)=0.54-0.46\cos(2n\pi/N-1)$$

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

Fast Fourier Transform:

FFT is used to convert each frame of N samples from the time domain into frequency domain. FFT is a fast algorithm to implement the Discrete Fourier Transform(DFT) [9].

$$X(k)=\sum_{n=0}^{N-1} x(n)\, e-\frac{j2\pi nK}{N}, \quad n=0,1,2\ldots N-1$$

Mel-frequency scale:

Mel frequency scale relates to the perceived frequency than the normal frequency. Human ear perception of frequency contents of sounds doesn't follow a linear scale. Therefore, for each tone with an actual frequency f, measured in Hz, a particular pitch is measured on a scale is called Mel scale. Mel scale is linear spacing below 1000 Hz and logarithmic spacing above 1000 Hz. To compute mels for given frequency f in Hz, can be calculated as

$$Mel(f)=S(k)=2595*\log10(1+f/700)$$

By using Mel frequency scales we can smoothen the magnitude spectrum and reduces the size of features involved.

Logarithm:

Log compress the featured values to match more closely to human audible range. Log can improve the robustness than mel scale by allows us to use cepstral mean subtraction.

Cepstrum:

This is the final step in the feature extraction, in this we can convert the log mel spectrum back into time domain. This result is called as Mel Frequency Cepstral 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) [10]

$$C(n) = \sum_{k=1}^{K} (\log S(k)) \cos[n(K - \frac{1}{2})\frac{\pi}{K}], \quad n = 1,2 \ldots . K$$

d) Feature matching.

Feature matching is done by using Dynamic Time Warping algorithm. It is used to find the shortest path between the MFCC coefficients of different classes of speech signal with the MFCC coefficients of test speech [11].

Unlike Linear Time Warping (LTW) which compares two-time series based on linear mapping of the two temporal dimensions, Dynamic Time Warping (DTW) allows a nonlinear warping alignment of one signal to another by minimizing the distance between the two as shown in Fig 6.

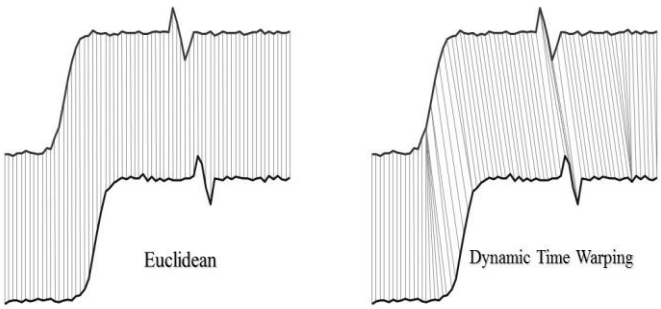
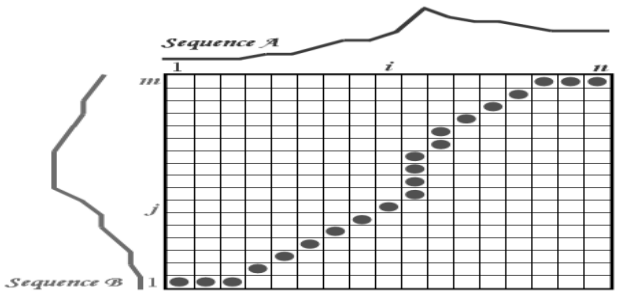


Fig 6: comparison between LTW and DTW



To find the distance between two samples of A= {x_1, x_2….} and B= {y_1, y_2….} can be give as….

Distance (D)= (| x_1-y_1|, |x_2-y_2|…) as shown in the fig…

Fig 7: Graphical representation to calculate shortest path using DTW

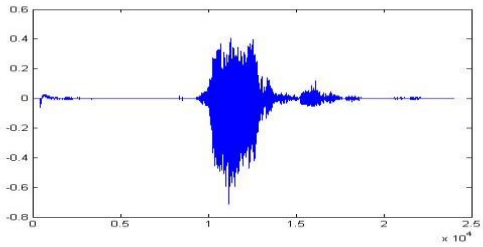
e) Control design

By using the feature matching technique, shortest distance can be calculated by using DTW algorithm, then we can assign some software control commands by using isolated speech to the software devices to control them and the hardware devices were

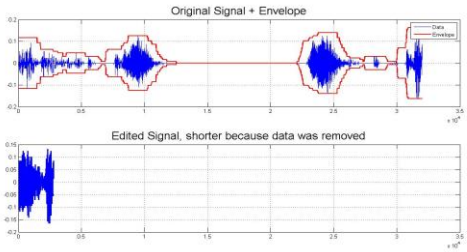
interfaced with the Speech Recognition system by using interfacing devices like RS232 etc.… and they can be controlled by using a programmer by dumping the program into the hardware devices to operate required functions in the hardware by the isolated words.

VIII. Result Analysis:

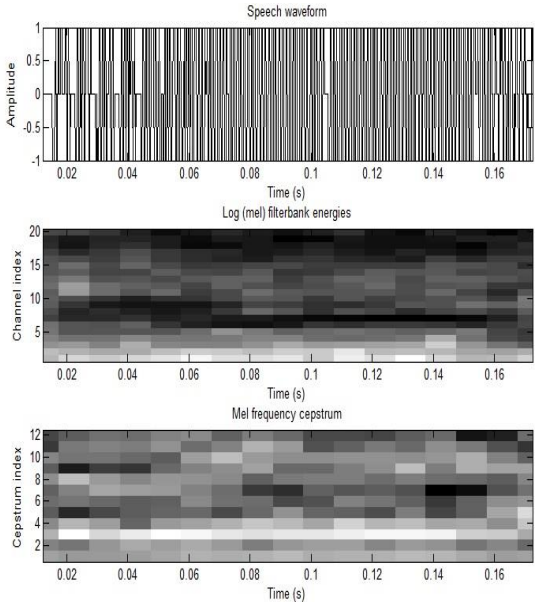
The results which are obtained after the design of software controlled Speech recognition system can be shown below as



Plot 1: Isolated speech



Plot 2: separation of voiced & unvoiced portions



Plot 3: MFCC extraction plots

IX. Conclusion

The main aim of this paper was to recognize isolated speech using MFCC and DTW techniques. The feature extraction was done by MFCC and the feature matching was done with the help of DTW technique. The extracted features were stored in a '. mat' file using MFCC algorithm. The experimental results were analyzed with the help of MATLAB and it is proved that the results are efficient.

References

[1]. Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk "Speech Recognition using MFCC" International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)

[2]. http://in.mathworks.com/company/newsletters/articles/developing-an-isolated-word-recognition-system-in-matlab

[3]. Gray Jr., A.H. & Markel, J. D. (1976), "Distance Measures for Speech Processing", IEEE Transactions on Acoustics, Speech and Signal Processing, issue 5, pp. 380-391, Oct 1976.

[4]. Brookes, M., Voicebox: Speech Processing Toolbox for Matlab [on line], Imperial College, London, available on the World Wide Web: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebo x/voicebox.

[5]. https://en.wikipedia.org/wiki/Euclidean_distance

[6]. https://en.wikipedia.org/wiki/Dynamic_time_warping

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

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

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

[10]. "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).

[11]. J. Deller, J. Proakis, and J. Hansen, Discrete Time Processing of Speech Signals, Prentice Hall, NJ, USA, 1993