ISSN (Print) : 2320 – 3765
ISSN(Online) : 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2014

Hardware Implementation of Speech Recognition Using MFCC and Euclidean Distance

UmaraniJ.Suryawanshi[1], Prof. Dr. S. R. Ganorkar[2]

Department of E&TC Engineering, Sinhgad College of Engineering, Pune, India[1]

Professor, Department of E&TC Engineering, Sinhgad College of Engineering, Pune, India[2]

ABSTRACT:This paper suggests Digital Signal processor (DSP) based speech recognition system with improved performance in terms of recognition accuracies and computational cost. The comprehensive surrey of various approaches of feature extraction like Mel filter banks with Mel Frequency Cepstrum Coefficients (MFCC). This paper describes an approach of isolated speech recognition by Digital Signal Processor TMS320C6713 using Mel scale Frequency Cepstral Coefficients and Euclidean distance. Several features are extracted from speech signal of spoken words. An experiments database of total five speakers, speaking 5-10 words each is collected under acoustically controlled room is taken. MFCC are extracted from speech signal of spoken words. To compare inter speaking differences Euclidean distance is used

KEYWORDS:Speech Recognition, Feature Extraction MFCC, Pattern Recognition, Euclidean distance.

I.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 popular spectral based parameter used in recognition approach is the Mel Freqency Cepstral Coefficients called MFCC, MFCC's are coefficients, which represents 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.

II.LITERATURE SURVEY

The system consists of microphone through which input in the form of speech signal is applied. The data acquisition system of speech processor acquires the output from the microphone and then itdetects the exact word spoken If such a system is installed in a motor car, then by using several start, stop,forward,backwardetc.; we can drive a carwithout even out hands. [1].Dynamic Time warping (DTW) algorithm is one of the most popular mathematical models in the field of speech recognition.

It is widely applied to some special fields such as resource poor embedded system for its simple and effective algorithm.DTW-based application of portable value added tax calculator with speaker dependent connected word and isolated word speech recognition abilities built in.[2]Automatic speech recognition is an interesting task but it requires a lot of effort. With technical development, speech recognition system achieved excellent results, still it possess some major limitations. Especially, recognition system with hidden markov models (HMM) as major elements are suitable for many applications, but do not suffer from major restrictions that make speech recognition system unsuitable for real time applications [2].

An isolated word, speaker dependent speech recognition system capable of recognizing spoken words at sufficiently high accuracy. The system has been tested and verified on MATLAB as well as TMS320C6713 DSK with

10.15662/ijareeie.2014.0308042

an overall accuracy exceeding 90%. An isolated word speech recognition system requires the user to pause after each utterance. There are two phases in the system: training and recognition. During the training phase, a training vector is generated for each spoken by the user. The training vector extracts the special features for separating different classes of words. Each training vector can serve as a template for a single word or a word class. During the recognition phase, the user speaks any word for which the system was trained. A test pattern is generated for that word and the corresponding text string is displayed as the output using a pattern comparison technique. [3]

Automatic Speech Recognition (ASR) is one of the most developing fields of the modern science having a wide range of applications. The statistical methods for speech recognition by extracting formants of the speech and analyzing their behaviour. It is a novel way to speech recognition. The whole analysis for speech recognition is based on upon five formants of the speech. The method has been tested for Urdu speech and the result obtained is of high accuracy. The effects of the formants are seen in the spectrum pattern of a speech sound. It is because of spectrum is strongly affected by resonance of vocal tract. When the effects of vocal resonance are apparent in the spectrum of speech sound, spectrum peaks may be called formants of the speech sound. Formants are the acoustic properties of vocal tract that produce the spectrum. Formants of a speech sound are numbered in order to their frequencies [4].

In speech recognition, Linear predictive coefficients (LPC) cepstrum based on LPC or Mel frequency Cepstral Coefficients (MFCC) are based on Mel-frequency filter bank are widely used as feature extraction that determines the performance. The complex parameter are converted to LPCC's and MFCC as a feature vector of HTK(HMM tool kit) in order to realize the HMM speech recognition. Through continuous speech recognition experiments with the converted LPCCs and MFCCs, it was found that the complex speech analysis method would not perform well than real one [5].

Automatic speech processing system are employed more and more often in real environments. However they are confronted with high ambient noise levels and their performance degrades drastically. Thus, there is strong need to improve the performance of these systems in noisy conditions. Adaption of the speech models to include the effect of noise. Methods using this approach including biascompensation algorithm and parallel model combination [6]. The speech recognition technique the test data is converted to templates. The recognition process then consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The lowest distance measure is based upon dynamic programming. This is called a Dynamic Time Warping (DTW) word recogniser. In order to understand DTW the two concepts need to be dealt with, features and distance [7].

III.SPEECH RECOGNITION

1.Recognition Algorithm

A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. A different levels, different operation are performed on the input signal such as Pre-emphasis, Framing, Windowing, Mel Cepstrum analysis and matching(recognition) of the spoken word.

2. Feature Extraction

The extraction of best parametric representation of acoustic signals is an important task to produce a better recognition performance. The MFCC algorithm is used to extract the features. The MFCC chosen for the following reason –

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

10.15662/ijareeie.2014.0308042

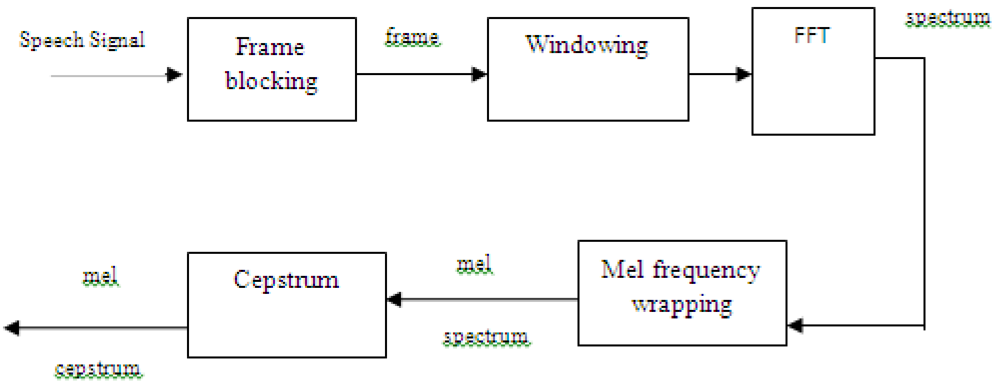

Fig-1 Block diagram of MFCC

As shown in figure 1 MFCC consists of computational steps. Each step has its function and mathematical approaches as discussed briefly in the following-

Step 1:Pre-emphasis

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

Step 2: Framing

This process of segmentation 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. 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 :Hamming window

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The hamming window equation is given as –

Y(n)=X(n) × W(n)

$$Y(\omega)=0.54-0.46\cos\left[\frac{2\pi n}{N}-1\right] 0 \leq n \leq N-1$$

If the window is defined as W(n), $0\leq n \leq N-1$

where

N=number of samples in each frame

Y(n)= output signal

X(n)= input signal

W(n)= Hamming window

Step 4: Fast Fourier Transform

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

$$Y(\omega)=FFT[h(t)*X(t)]=H(\omega)X(\omega)$$

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

Step 5: Mel Filter Bank Processing

The frequencies range in FET spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 2.

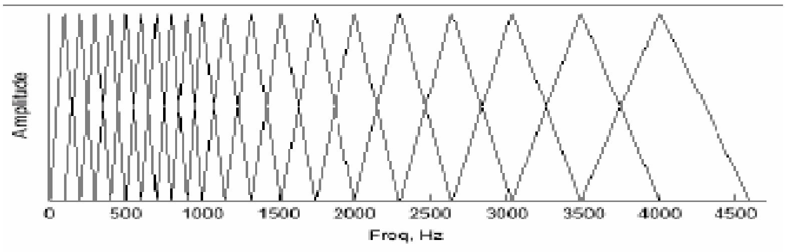

Fig- 2 Mel scale Filter Bank

This figure 2 show a set of triangular filters that are used to compute a weighted sum of filters 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 utility at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in Hz:

$$M=1127.01048\log_e(1+ f / 700)$$

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 conversion is called Mel Frequency Cepstrum Coefficients. The set of Coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

3. Vector Quantization

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called Cluster and can be represented by its center called a centroid. The collection of all code words is called a codebook.

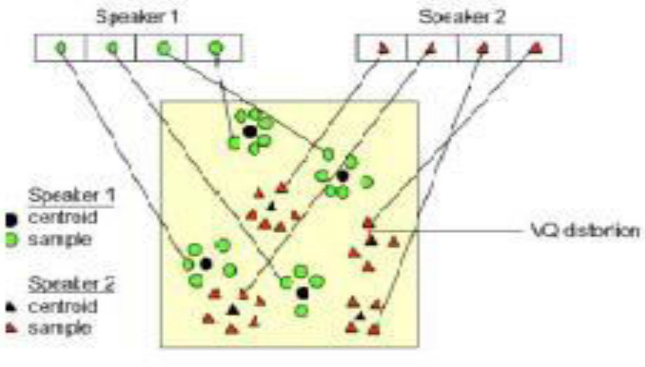

Fig- 3 Vector Quantization

One speaker can be discriminated from another based of the location of centroid figure 3 shows a conceptual diagram to illustrate this recognition process. In the figure, only two speakers and two dimensions of the acoustic vectors from the show. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker- specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result code words (centroid) are shown in figure 3 by black circles and black triangles for speaker 1 and 2 respectively. The distance from a vector to the closest code wordof a codebook is called a VQ-distortion. In the

10.15662/ijareeie.2014.0308042

recognition phase an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total VQ codebook with smallest total distortion is identified.

4.Feature Matching

There are many feature-matching technique used in speaker recognition such as Dynamic Time warping(DTW). DTW technique is used for feature matching.

Dynamic Time Warping (DTW)

The time alignment of different utterance is the core problem for distance measurement in speech recognition. A small shift leads to incorrect identification. DTW is efficient method to solve the time alignment problem. DTW algorithm aim at aligning two sequences of feature 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 n dimensional space.

$x=[x1,x2............x_n]$

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.

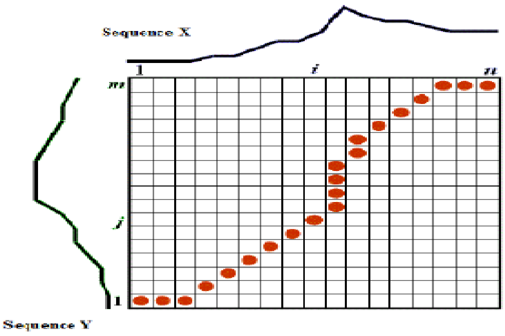

Fig. 4 Global Distance Grid

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

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

The best match or alignment between these two sequences is 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 distance (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 path 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)

IV. DATABASE

Database consists of two groups of speech samples records in an environment controlled recording room to have all possibly less acoustic interferes to the quality of sound sample during recording time. The first group comprises of total five speakers, speaking e words each from of same sound samples. All speech signals are recorded under most similar setting conditions such as the same length of recording time, the level of sound amplitude. In training, code composer program named 'train'. In testing phase when a code composer program named 'test' is executed at postulates to the user to choose any speech sample from the test group that are pre-recorded in the database. MFCC at the back end extracts the features of the chosen speech sample. Euclidean distance measures the minimum distance and its result of the test program that shows the correct spoken word in the command window of code composer program.

V.RESULT

In the following table different performance parameters are given for standard and real time recognition.

Table 1- Performance evaluations parameter for method MFCC and Euclidean distance for speaker and speech recognition

Sr.No.	Database	Accuracy	Precision	Sensitivity
1	Standard	85.00%	70.00%	40.00%
2	Real Time	76.67%	80.91%	30.00%

Recognition (standard database)
Total words=12

- Same person same word= 7
- Different word different person= 2
- Same word different person= 2
- Different word same person= 1

Recognition (Real Time)
Total words=15

- Same person same word= 9
- Different word different person= 3
- Same word different person= 2
- Different word same person= 1

V. CONCLUSION

The main aim of the paper was to recognize isolated speech using MFCC and DTW. The feature extraction was done using MFCC and the feature matching was done using DTW technique. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speech signal with database. The experimental results were analysed with the help of code composer of TMS620C6713 and it is proved results are efficient.Also implemented same speech recognition using matlab.This process can be extended for n number of speakers. This papershow that the DTW is nonlinear feature matching technique in speech identification, with minimum error rates and fast computing speed.

REFERENCES

[1] Mohammad Salman Haleem, "Voice Controlled Automation System", Proceedings of the 12th IEEE International Multitopic Conference, December 23-24,2008,, 2008.

[2] Chun wan, LiliLiu,"Research and Improvement on Embedded system Application of DTW-based Speech Recognition ", IEEE,2008.

[3] Talal Bin Amin, ItekharMahmood, "Speech Recognition using Dynamic Time Warping", 2nd International Conference on Advances in space

Technologies, Islamabad,29th -30th Nov.2008, vol2, pp.74-79.

[4] AhmadAli,SafiullahBhatti, Dr.MuhammadSleemMilan,"Formants based Analysis for Speech Recognition", IEEE 1-4244-0457-6/06$20.00,2006.

[5] Tasuhiko Kinjo and Keiichi Funaki,"On HMM Speech Recognition based Complex Speech Analysis", IEEE 1-4244-0136-4/06/$20.00,2006,,pp.3477-3480.

[6] Ye Tian ,Ji Wu, Zuoying Wang and Daijn Lu," Robust Noisy Speech Recognition with adaptive Frequency bank selection", Proc. of the Fourth IEEE International Conference of Mutimodal Interfaces,2002.

[7] Belgace Ben Mosbah, " Speech Recognition for Disabilities People", IEEE 0-7803- 9521-2/06/$20.00,2006,pp.864-869.

[8] Heungsuk Chin, J.Kim,I.Kim, Y.Kwon, K.Lee, Sung-il Yang, "Realization of Speech Recognition using DSP", ISIE 2001, Pysan , Korea,2001,pp.508-512.

[9] SwapnilD.Daphal, SonalJ.Jagtap, "DSP based Improved Speech Recognition System",International Conference on Communication, Information& Computing Technology, octo.19-20,Mumbai, India,2012.

[10] Jaime Zabalz, JinchangRen, Carmine Clemente, Gaetano Di Caterina and John Soragha,"Embedded SVM on TMS320C6713 For Signal Prediction In Classification And Regression Applications", 5th Education and Research conference,13th -14th sept.2012,pp.90-94.

[11] Xiu-Qing Zhang, Shu-Wang Chen," Speech Recognition System Based on DSP and SVM", Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010,pp.2313-2316.

[12] Yoshiaki Kitazume, EijiOhira, TakeyukiEndo,"LSI Implementation of a pattern Recognition Algorithm for Speech Recognition",IEEE Transactions of Acoustics, Speech and Signal Processing, vol. Assp-33,No.1,Febrauary 1985.

10.15662/ijareeie.2014.0308042