International Journal of Engineering Research and General Science Volume 4, Issue 4, July-August, 2016
ISSN 2091-2730

Isolated Word Recognition Using MFCC and Vector Quantization

Manish Kumar Sharma

Department of Electrical Engineering, VJTI Mumbai

Email:-manishkumarsharma0911@gmail.com

Mob. NO. 8380902226

Abstract— Automatic Speech Recognition (ASR) technology is a way to interface with computer. In this paper we describe speech recognition technique using multiple codebooks of MFCC derived features. The proposed algorithm is useful in detecting isolated words of speech. In this algorithm we first create database i.e. codebook by calculating mel frequency cepstral coefficient first and then codeword for each word. Vector quantization method is used to obtain codeword. Different codewords make a codebook. We calculate distance between uttered speech codeword and codewords which are stored in matrix form to detect the particular word. Codebook reduces processing time. The efficiency obtained by MFCC is much more than the other available techniques.

Keywords— MFCC, ASR, Vector quantization, DFT, FFT, Pattern Recognition, Discrete cosine transform**.**

INTRODUCTION

Recently, there have been a number of successful commercial ASR products. However, still many problems exist in real-world and real-time speech recognition applications. The recognition accuracy of a system is, in most of the cases, very far from that of a human listener, and its performance would degrade drastically with small modification of speech signals or speaking environment [1]. To overcome this problem care must be taken about the signal to noise ratio of speech signal [1] [11].

There are three major types of feature extraction techniques, like, perceptual linear prediction (PLP) Mel frequency cepstrum coefficient (MFCC) and linear predictive coding (LPC). MFCC and PLP are the most commonly used feature extraction techniques in modern ASR systems [2]. In our proposed system we have used MFCC for feature extraction purpose because it gives best results among all the available techniques.

ALGORITHM FOR SPEECH RECOGNITION

In this proposed system we first do the preprocessing on the speech signal and then calculate mel frequency cepstrum coefficient. Using these coefficients we make a codeword for each and every word in our database by using vector quantization method. Different codewords make a codebook. These codewords are stored in the matrix form. Then for speech recognition we calculate Euclidian distance between test sample and with each and every sample which is already stored in the database and that particular stored sample which gives minimum distance will be our output. The same procedure is explained by using figure 1.

A. Data Acquisition

The speech waveform, sampled at 11025 Hz is used as an input to the feature extraction module. Sampling frequency is chosen in such a way that we should have a good quality signal but care should be taken about the size of the speech waveform. Because if we choose higher sampling frequency we may have good quality signal but time taken for processing will increase. The acoustic model of the environment in which we are recording may affect the performance of the system [12].

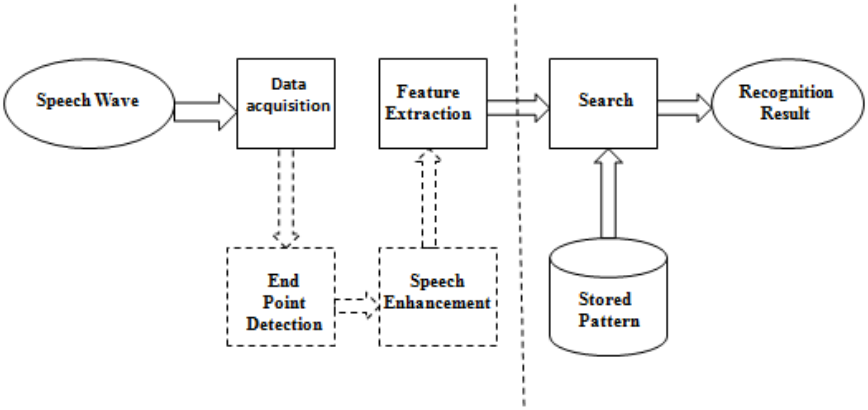

Fig.1. Block diagram for Speech Recognition

B. Endpoints detection

In order to detect an endpoint in the signal, a signal recording has been processed by the Endpoint Detector Algorithm which uses energy and zero crossing rate of signal to detect endpoints [9].

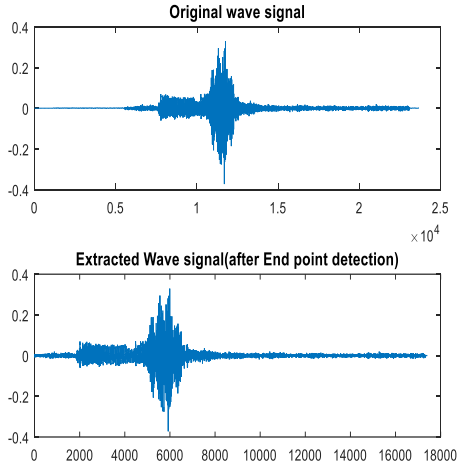

Fig.2. End point detection of signal INDIA

FEATURE EXTRACTION USING MFCC

For feature extraction purpose MFCC is used and its algorithm has following steps:

A. Pre-emphasis

Noise is an unwanted signal and it has a greater effect on the higher modulating frequencies than the lower ones. Hence, to reduce the effect of noise higher frequencies are artificially boosted to increase the signal-to-noise ratio. Pre-emphasis process is used for spectral flattening using a first order finite impulse response (FIR) filter [2]. Equation (1) represents first order FIR filter.

$$H(z) = 1 - \alpha z^{-1}, 0.9 \leq \alpha \leq 1 \quad (1)$$

B. Frame blocking

Speech is a non-stationary signal but for the analysis and feature extraction purpose we always consider short duration frames and we assume that for that short duration signal is stationary. Frame duration should not be too long because it may affect time resolution and frame duration should not be too short because it may affect frequency resolution [2]. In the proposed system frame of length 256 samples is used and frame overlapping length is 128.

C. Windowing

There may be discontinuities at the beginning and end of the frame which are likely to introduce undesirable effects in the frequency response. Hence, to reduce discontinuities each row is multiplied by a window function. A window alters the signal, tapering it to almost zero at the beginning and the end. In the proposed system Hamming window is used as it introduces the least amount of distortion. Our proposed system uses Hamming window of length 256 [5].

$$h(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\Pi n}{N}\right), 0 \le n \le N \\ 0, otherwise \end{cases} \quad (2)$$

D. Spectral magnitude of DFT

DFT is used to convert time domain signal into frequency domain signal. Time domain data is converted into frequency domain to obtain the spectral information [5]. By spectral information we mean the energy levels at each and every frequency in the given window. Time domain data is converted to frequency domain by applying Discrete Fourier Transform (DFT) on it.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\Pi kn}{N}}, 0 \le k \le N-1 \quad (3)$$

For each frame we find its DFT. We have used FFT algorithm to find the DFT of the signal. FFT output is a set of complex numbers i.e. real and imaginary parts. But speech recognition systems deal with real data. Hence, we avoid the use of complex value. Let us assume the real and imaginary parts of X(k) as Re(X(k)) and Im(X(k)), then the magnitude of the speech signal can be obtained by using equation (4).

$$X(k) = \sqrt{(Re(X(k)))^2 + (Im(X(k)))^2} \quad (4)$$

E. Mel frequency filter bank

The best thing about the Mel-frequency analysis of speech is that, it is based on human perception experiments. How our ears recognize the speech same is used in Mel frequency analysis and therefore it is proven to be best technology for speech recognition. It has also been proved that ears of human are more sensitive and have higher resolution to lower frequencies compared to higher frequencies. Therefore, the filter bank is designed in such a way that it emphasizes the low frequency over the high frequency [2].

Also the speech signal does not follow the linear frequency scale used in FFT. Hence, a perceptual scale of pitches equal in distance, namely Mel scale is used for extraction of features. Mel scale frequency is quite similar to the logarithm of the linear frequency, reflecting the human perception. We use log because our ears work in decibels.

$$Mel(f) = 2595\log_{10}(1 + \frac{f}{700}) \quad (5)$$

Use of triangular band pass filter is to extract the spectral envelope, which is formed by dominant frequencies present in the speech signal. Therefore, Mel-frequency filters are triangular band pass filters which are non-uniformly spaced on the linear frequency axis but uniformly spaced on the Mel frequency axis and numbers of filters are more in the lower frequency region and less number of filters in the higher frequency region [2].

Below equation (6) denotes the filter bank with M (m = 1, 2, 3….M) filters, where m is the number of triangular filter in the filter bank then Hm(K) will be as below. In the implementation we have used 13 triangular band pass filters and the range of frequency over which we are operating is 133Hz to 2143Hz.

$$Hm(k) = \begin{cases} 0, for\ k < f(m-1) \\ k - \frac{f(m-1)}{f(m)-f(m-1)}, for\ f(m-1) \le k \le f(m) \\ f(m+1) - \frac{k}{f(m+1)-f(m)}, for\ f(m) \le k \le f(m+1) \\ 0, for\ k > f(m+1) \end{cases} \quad (6)$$

F. Logarithm of filter energies

Human ears smooth the spectrum and use the logarithmic scale approximately. We use equation (7) to compute the log-energy i.e. logarithm of the sum of filtered components for each filter.

$$S(m) = \log_{10}[\sum_{k=0}^{N-1}|X(k)|^2 . Hm(k)], 0 \le m \le N \quad (7)$$

G. Discrete cosine transform

The discrete cosine transform (DCT) converts the log power spectrum (Mel frequency domain) into time domains. DCT gathers most of the information of the signal to its lower order coefficients, resulting in significant reduction in computational cost. Equation (8) represents the discrete cosine transform.

$$C(k) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi k\left(m+\frac{1}{2}\right)}{M}\right),\ 0 \le k \le K \quad (8) \quad (8)$$

The number of coefficients obtained for each frame is 13 and equal to number of triangular band pass filter. MFCC features will be a matrix and whose rows are equal to number of triangular bandpass filter and columns are equal to number of frames in a word. Number of frames varies for each word so MFCC matrix size will be different for each word.

VECTOR QUANTIZATION

The Vector Quantization (VQ) technique is used for mapping vectors from a large number of vector space to a finite number of regions in that space. We call each region a cluster and can be represented by its centre called a codeword. Accuracy of the system will be defined by inter-cluster variance and intra-cluster variance. For good accuracy inter-cluster variance should be large and intra-cluster variance should be less. The collection of all codeword is called a codebook. After extracting features, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors [4]. The problem of speech recognition belongs to a much broader topic in engineering and scientific research and is called pattern recognition. The goal of pattern recognition is to classify objects of importance into one of a number of classes or categories [3]. The clusters here refer to isolated word.

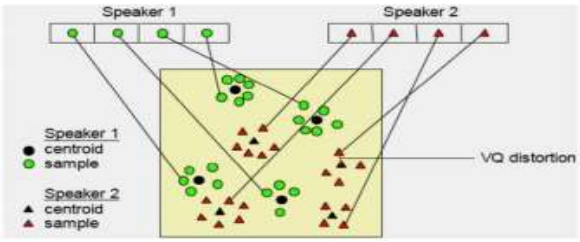

Fig.3. clusters of different codeword

The algorithm which we have used for vector quantization is Lin-Buzo-Gray algorithm. LBG algorithm is like a K-means clustering algorithm which takes a set of input vectors S = {xi ∈ Rd| i = 1, 2, . . .n} as input and generates a representative subset of vectors C = {cj ∈ Rd| j = 1, 2, . . . ,K} with a user specified K << n as output according to the similarity measure. In our implementation we have used k=16 and d=number of frames in speech waveform of a word [8].

Vector quantization can give different size of matrix as its output but we have chosen a matrix of 8×16 for each sample as it gives cluster which gives high accuracy. For recognition of test sample we calculate Euclidian distance between test sample and each sample stored in the database. Then we calculated average of the distance of a every cluster and compared all the distances. The minimum distance cluster is output.

RESULTS

In this work five words i.e. Invention, Technology, Science, Teacher and India are used to train and test the system. At first we saved the 10 patterns of these words and then one of the earlier speaker's samples is used to test the system. Test sample used is of word 'Science'. Distance of test sample with word Invention, Technology, Science, Teacher and India is 1.8265, 1.7584, 1.1494, 1.9367 and 1.6327 respectively. The result shows that the distance is minimum for the word science and is equal to 1.1494 hence spoken word is 'Science'. Below figure (4) shows the same result graphically.

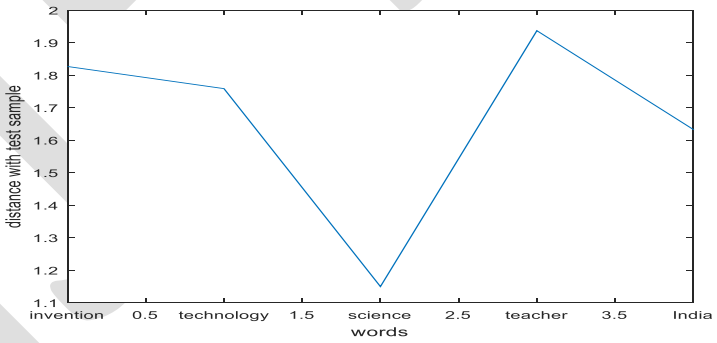

Fig 4. Result showing distance with test sample

ACKNOWLEDGMENT

Authors would like to express their sincere gratitude towards Dr. R N Awale for being a source of help and encouragement to put forth this project. Authors would also like to thank their Institute Veermata Jijabai Technological Institute, Mumbai for providing the necessary facilities for carrying out research work.

CONCLUSION

With MFCC and Vector Quantization, isolated word detection system is generated in MATLAB environment. System is trained by saving templates of five separate words. Results showed that saving ten templates for each word in training phase gives good results compared with five templates but generates delay.

REFERENCES:

[1] Endpoint Detector Algorithm for Speech Recognition Application E.A. Escoto-Sotelo, E. Escamilla-Hernandez, E. Garcia-Rios, H. M. Perez-Meana Instituto Politecnico Nacional SEPI ESIME Culhuacan

[2] Yuan Meng, Speech recognition on DSP: Algorithm optimization and performance analysis, The Chinese university of Hong Kong, July 2004, pp. 1-102

[3] Isolated Word Speech Recognition Using Vector Quantization (VQ) Dipmoy Gupta, Radha Mounima C. Navya Manjunath, Manoj PB Dept. of EC,AMCEC, Bangalore Volume 2, Issue 5, May 2012, pp 164-168

[4] Vector Quantization in Speech Coding. John Makhoul, Fellow ,IEEE, Salim Roucos, Member IEEE, And Herbert Gish ,Member, IEEE proceedings of the IEEE, Vol 73, No. 11, November 1985 pp. 1552-1586

[5] An Efficient MFCC Extraction Method in Speech Recognition Wei HAN, Cheong-Fat CHAN, Chiu-Sing CHOY and Kong-Pang PUN Department of Electronic Engineering, The Chinese University of Hong Kong ISCAS 2006 pp. 145-148

[6] Robust speech recognition using neural networks and hidden markov models Lin Cong, Saf Asghar and Bin Cong Advanced Micro Devices, 3625 Peterson Way, Santa Clara, CA, 95054, USA Dept. of Computer Science, California State University.

[7] A general approach to natural language conversion Md. Abu Nuser Musud', Md. Muntusir Mamun Joarder', Md. Turiq-UI-Azam

[8] A Robust Lin-Buzo-Gray Algorithm in Data Vector by Quantization Liu Jing , Wang Quan

[9] Speech Analysis for Automatic Speech recognition, Norwegian University of Science and Technology by Noelia Alcaraz Meseguer

[10] MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition Siddhant C. Joshi, Dr. A.N.Cheeran

[11] Fundamentals of speech by Lawrence Rabiner and Biing Huang Juang

[12] Vocal tract acoustics and speech synthesis by Shinji Maeda Ecole Nationale Supkrieure de TkltScommunication, Dipartement SIGNAL and Centre National de la Recherche Scienhifiqpe