Analysis and Implementation of Speech Recognition System using ARM7 Processor

Ms. Nishiya Vijayan
Dept. of ECE,Sree Narayana Gurukulam College of Engineering, Kadayiruppu, Ernakulum, India
E-mail: nishiyavijayan27@gmail.com

Abstract

This paper introduces implementation and analysis of speech recognition system. Speech Recognition is the process of automatically recognizing a certain word spoken by a particular speaker based on individual information included in speech waves. This paper presents one of the techniques to extract the feature set from a speech signal, which can be used in speech recognition systems and an analysis study has been performed. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC),and others. Studies and experiments show that MFCC provides better results than LPC. Here vector quantization is used to increase speech recognition accuracy. Experiments shows that as the no. of MFCC coefficients increases get better accuracy, code book size also affects accuracy. The MFCC and VQ algorithm, for speech recognition have been implemented in MATLAB 7.7(R2008b) version on Windows7 platform. The control circuitry has been implemented in Keil µVision3; the supporting hardware setup is being implemented.

Keywords: Speech Recognition; MFCC; Vector Quantization; LPC

1. Introduction

Speech recognition has been an interesting research field for the last decades, which still yields a number of unsolved problems. Speech recognition is the process by which a computer identifies spoken words. Basically, it means talking to computer, and having it correctly recognize what you are saying. The principle is to extract certain key features from the uttered speech and then treat those features as the key to recognizing the word when it is uttered again. In this paper, the issue of speech recognition has been studied and a speech recognition system is developed for real time command recognition using MFCC and Vector quantization model to control dynamic devices.

The goal of this paper is a real time speech recognition system, which consists of comparing a speech signal from a registered speaker to a database of known speaker's speech. The system will operate in two modes: A training mode, a recognition mode. The training mode will allow the user to record voice commands and make a feature model of each voice commands. The recognition mode will use the information that the user has provided in the training mode and attempt to isolate and identify the voice command. The Mel Frequency Cepstral Coefficients and the Vector Quantization algorithms are used to implement the paper using the programming language Matlab, and KEIL software is used to drive the hardware section. The analysis has been performed with MFCC and LPC, experimental results shows that MFCC and VQ provide better recognition than LPC and VQ.

2. Feature Extraction

The speech feature extraction is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. So we need feature extraction. The resonance frequencies of the vocal tract tube are called formant frequencies or simply formants, which depends on the shape and dimensions of the vocal tract. Feature extraction means extract this formant frequency, which is unique to each person.

2.1 Mel Frequency Cepstral Coefficients

Mel frequency Cepstral Coefficients (MFCC) based on short time spectral analysis, and are commonly used feature vectors extraction method for speech recognition. A block diagram explanation of an MFCC process is given in Figure 1. The speech input is recorded at a sampling rate of 22050Hz. This sampling frequency is chosen to minimize the effects of aliasing in the analog-to-digital conversion process.

2.2.1 Framing

Due to physical constraints, the vocal tract shape generally changes fairly slowly with time and tends to be fairly constant over short intervals (around 10–20 ms).So Fourier transform cannot be directly applied because speech signal cannot be considered stationary due to constant changes in the articulator system within each speech utterance. To solve these problems, speech signal is split into a sequence of short segments in such a way that each one is short enough to be considered pseudo-stationary. Finally, a feature vector will be extracted from the short-time spectrum in each window. The whole process, known as short-term Spectral analysis.

2.2.2 Windowing

Prior to any frequency analysis, each section of signal is multiplied by a tapered window. This type of windowing is necessary to reduce any discontinuities at the edges of the selected region, which would otherwise cause problems for the subsequent frequency analysis by introducing spurious high-frequency components into the spectrum. The length of each analysis window must be short enough to give the required time resolution, but on the other hand it cannot be too short if it is to provide adequate frequency resolution.

The use for hamming windows is due to the fact that it removes the frequency domain high frequency components in each frame due to such abrupt slicing of the signal.

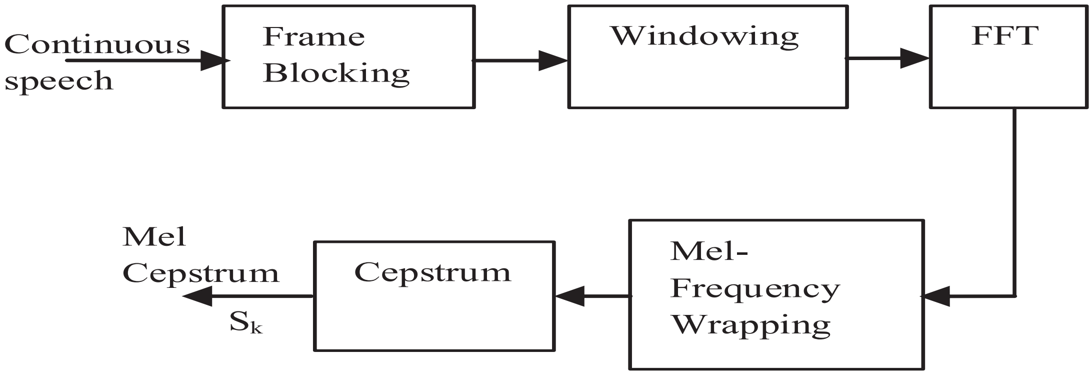

Figure 1. Block Diagram of MFCC Process

2.2.3 FFT

In this step, FFT converts each frame of N samples from the time domain into the frequency domain. Framing enables the non-stationary speech signal to be segmented into quasi stationary frames, and enables Fourier transformation of the speech signal. The single Fourier transform of the entire speech signal cannot capture the time varying frequency content due to the nonstationary behavior of the speech signal. Therefore, Fourier transform is performed on each segment separately

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,...,N-1 \tag{1}$$

The FFT is a fast algorithm to implement the Discrete Fourier Transform, which is defined on the set of N samples $\{x_n\}$, as shown in the above equation (1). The basis of performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain.

2.2.4 Mel Frequency Wrapping

The speech signal consists of tones with different frequencies. For each tone with an actual Frequency f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The mel-frequency scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. The following formula can used to compute the Mels for given frequency f, in Hz.

$$\mathrm{mel}(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

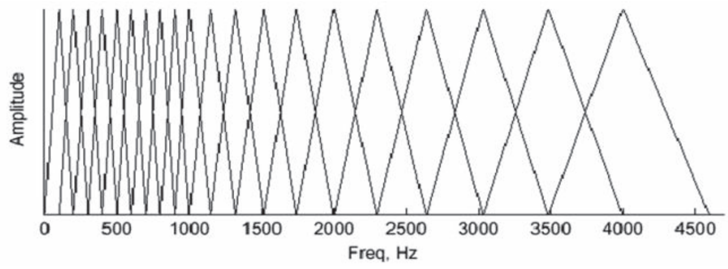

Fig 2. The Mel filter bank

The Mel filter bank consists of 20 triangular shaped band pass filter which are centred on equally spaced frequencies in Mel domain. The output of Mel filter bank is an array of filtered values, typically called Mel spectrum, each corresponding to the result of filtering the input spectrum through an individual filter.

2.2.5 Cepstrum

The output of the equation is log mel spectrum, it has to be converted back into time. The result is called the Mel frequency cepstrum coefficients (MFCCs. In this paper DCT is used, since the signal is real with mirror symmetry. So no complex operation on DCT does not exist. The DCT implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. The DCT is computationally more efficient.

The MFCCs can be calculated using this equation 3,

$$C_n^{\sim} = \sum_{k=1}^{K} (\log \tilde{S}_k) \left[\left(k - \frac{1}{2}\right) n \frac{\pi}{K}\right] \tag{3}$$

Where n=1, 2,…....K

These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker. The number of Mel cepstrum coefficients, K, is typically chosen as 12-32. The first component, c_0, is excluded from the DCT since it represents the mean value of the input signal which carries little speaker specific information.

3. Feature Matching

3.1 Vector Quantization

Vector Quantization is a lossy data compression method based on the principle of block coding and is the generalization of Scalar Quantization to groups of n pixels called vectors. A Vector Quantizer needs only a codebook and a distortion measure usually MSE. It is a fixed-to-fixed length algorithm. In 1980, Linde, Buzo, and Gray (LBG) proposed VQ design algorithm.

VQ is a process of mapping vectors of a large vector space to a finite number of regions in that space. Each region is called a cluster and is represented by its centre. A collection of all the centroid makes up a codebook. The amount of data is significantly less, since the number of centroid is at least ten times smaller than the number of vectors in the original sample. This will reduce the amount of computations needed for comparison in later stages. An element in a finite set of spectra in a codebook is called a code vector. The codebooks are used to generate indices or discrete symbols. This paper uses the LBG algorithm, also known as the binary split algorithm to estimate code book.

3.2 Matching

In the recognition phase an unknown speaker, represented by a sequence of feature vectors $\{x_1, x_2,…., x_T\}$, is compared with the codebooks in the database. For each codebook a distortion measure is computed, and the speaker with the lowest distortion is chosen. The well known distance measures are Euclidean, bhattacharya and Mahalanobis. In this paper Euclidean distance is used.

3.2.1 Euclidean distance

In this paper Euclidean distance is used as the distortion measure, which is the sum of squared distances between vector and its representative (centroid).

The formula used to calculate the Euclidean distance can be defined as following: The Euclidean distance between two points P = $(p_1, p_2…p_n)$ and Q = $(q_1, q_2...q_n)$,

$$\sqrt{(p_1-q_1)^2 + (p_2-q_2)^2+..+(p_n-q_n)^2} = \sqrt{\textstyle\sum_{i=1}^{n}(p_i - q} \tag{4}$$

The Euclidean distance is calculated as

$$s(X, C_i) = \frac{1}{T}\sum_{t=1}^{T} d\left(x_t, c_{min}^{i,t}\right) \tag{5}$$

Where c_{min} denotes the nearest codeword x_t in the codebook C_i and d(-,-) is the Euclidean distance. Thus, each feature vector in the sequence X is compared with all the codebooks, and the codebook with the minimized average distance is chosen to be the best.

4. Experimental Setup

Figure 3 showing the complete experimental setup for DC motor drive through speech recognition. Here PC denotes the personal computer. Microphone is used to pick the sound. According to the recognized command the LPC 2388 microcontroller starts receiving those characters which carried out throw the signal from the MAX232 from PC and the processing operation will begin depending on the program which has been written in Embedded C language and downloaded to the board using Flash Magic or JTAG debugger.

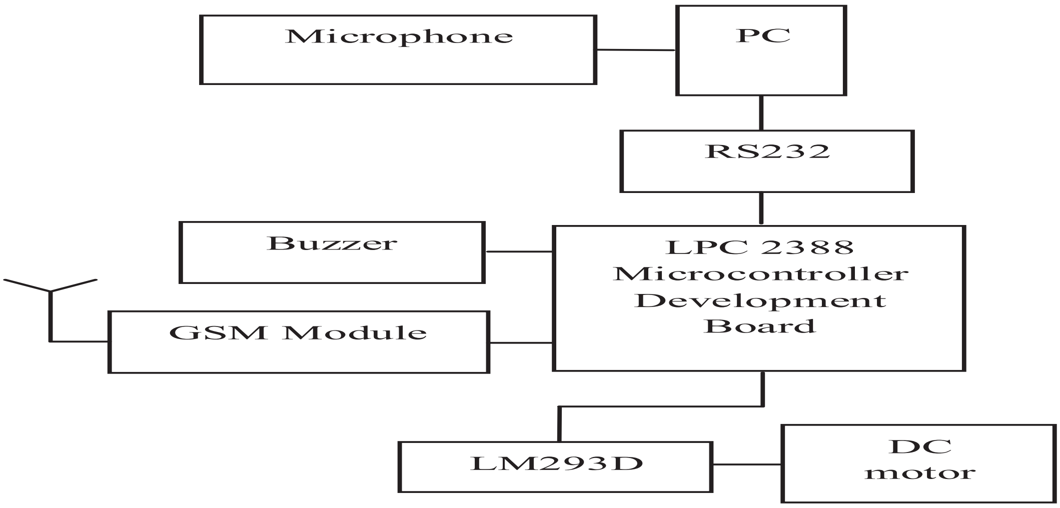

Figure 3. Hardware Block Diagram

Speech recognition was implemented in Matlab 7.7.0 version.

4.1 ARM7 Processor

MCB 2300 populated with LPC 2388 used for driving the control device, DC motor. MCB 2300 is a USB powered device. LPC2388 is a device with advanced ADC, DAC and USB capabilities. A 12.0 MHz crystal provides the clock signal for the CPU.

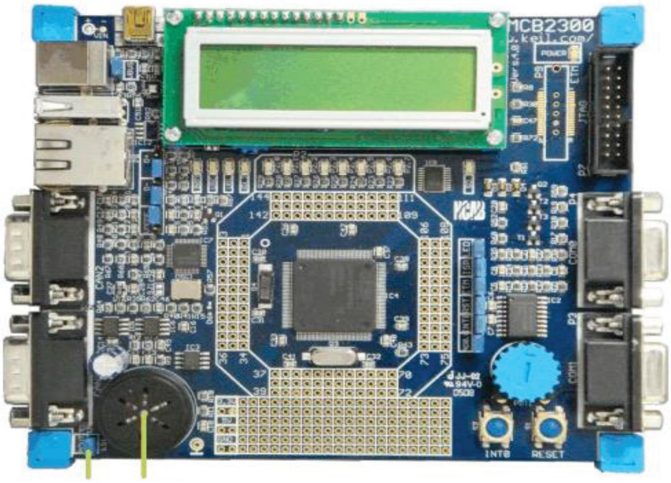

Fig 4. MCB 2300 Evaluation Board

5. Results and Analysis

Here record the words Forward, Backward, stop during training and store it in the database. Then compare with the speech recorded in testing phase. In this paper simulation is done at 3 level (1) Window type (2) Mel Frequency Cepstral Coefficient size (3) Codebook size. The following sections show the effect of changing the parameters associated to Window type, MFCC, Code book size

Table 1: comparison of different window performance (in terms of % efficiency)

Code Book Size	Triangular	Rectangular	Hamming
1	33.33	26.67	36.67
2	46.67	33.33	53.33
4	53.33	40	66.67
8	56.67	46.67	80
16	66.67	53.33	93.33
32	80	60	93.33

5.1 Window Type

The system has been implemented in Matlab7.7 on windows 7 platform. The result of the study has been presented in Table 1 and table 2. The testing phase has been repeated for 30 times. Here, identification rate is defined as the ratio of the number of times the speech identified to the total number of times tested.

The Table 1 shows identification rate when triangular, or rectangular, or hamming window is used for framing in a Mel scale. The table clearly shows that as codebook size increases, the identification rate for each of the three cases increases, and when codebook size is 32, identification rate is 93.33% for hamming windows and for the triangular window it must be 90%. When a codebook size is of 4 and hamming window is used it provides 83.33% identification rate is obtained. Therefore in speech recognition, the most commonly used window shape is the hamming window. The study reveals that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has to be increased as the number of speech command increases.

From the table 1, it is obvious that increasing the code book centroid results in increasing the identification rate, but the computational time will also increase.

5.2 Why MFCC

Several feature extraction algorithms are used to this task such as, linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC)**.**

MFCC is best compared to LPC for the following reasons

- Shows high accuracy results for clean speech .
- MFCC based on human ear auditory variations.
- Experiments show that the parameterization of the MFC coefficients which is best for discriminating speakers is different from the one usually used for speech recognition applications.
- The most common algorithm that used for speaker recognition system.

6. Conclusion

Speech recognition is a truly amazing human capacity and is one of the advanced areas. Many research works has been taking place under this domain to implement new and enhanced approaches. This paper implemented real time speech recognition system and perform its analysis. The feature extraction is done using MFCC and the speaker was modeled using Vector Quantization technique. Using the extracted features a codebook from each speaker was build clustering the feature vectors using the VQ algorithm. In this paper recognized speech command is used to control the DC motor drive. If we increase the number of samples as well as the number iterations (training), it can produce a good recognition result. The paper study reveals that as the number of centroid increases, the identification rate of the system increases. Also, the number of centroid has to be increased as the number of speech commands increases.

References

Punit Kumar Sharma, Dr. B.R. Lakshmikantha and K. Shanmukha Sundar," Real Time Control of DC Motor Drive using Speech Recognition", 978-1-4244-7882-8/11/$26.00 ©2011 IEEE.

Vibha Tiwari "MFCC and its applications in speaker recognition" (Received 5 Nov., 2009, Accepted 10 Feb., 2010) International Journal on EmergingTechnologies 1(1): 19-22(2010).

Wang Chen□Miao Zhenjiang , "Differential MFCC and Vector Quantization used for Real-Time Speaker Recognition System", 2008 IEEE Congress on Image and Signal Processing.

Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker Identification Using MEL Frequency Cepstral Coefficient",. 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh

F. K. Soong A. E. RosenbergL, R. Rabiner B. H. Juang AT&T Bell Laboratories Murray Hill, New Jersey 07974, "A Vector Quantization Approach to Speaker Recognition ".

"A Speaker Identification System using MFCC Features with VQ Technique, 2009 Third International Symposium on Intelligent Information Technology Application

A Robotic Arm Design for Stroke Patients, 2009 3rd International Conference on Power Electronics Systems and Applications Digital Reference: K210509126.

Dr. H. B. Kekre et. al., "Speaker Identification by using Vector Quantization", International Journal of Engineering Science and Technology Vol. 2(5), 2010, 1325-1331.