DYNAMIC SPEECH RECOGNITION SYSTEM TO CONTROL HOME APPLIANCES

[1]Sachin R Jaybhaye, [2]Dr P.K.Srivastava
[1]PG Student, [2]Prof. Department of Electronics and Telecommunication
JSPM, PVPIT, Bavdhan Pune, Maharashtra, India.
Email: [1]sachinjay8812@gmail.com,[2]pankoo74@gmail.com

Abstract - Speech processing is one of the most important branches in digital signal processing. Here we present speech recognition by using the Mel-Scale Frequency Cepstrum Coefficients (MFCC) and Vector Quantization (VQ) on open source hardware platform (BegalBoard XM) using Matlab Simulink. There are different methods of speech recognition system. They generally are more accurate for the correct speaker, but much less accurate for other speakers. They assume that the speaker will speak in a consistent voice and tempo. Speaker independent systems are designed for a variety of speakers. Adaptive systems usually start as speaker independent systems and utilize training techniques to adapt to the speaker to increase their recognition accuracy. Here we use two sessions: first is for training session and another is for recognition session. Sound data will be record from audio device in training and speech recognition takes place by giving audio commands. Sample rate of audio signal is 48KHz and the output frame size is 4800 samples.

Keywords- HMM (Hidden Markov model), LPC(Linear predictive coding), MFCC(Mel-Frequency Cestrum Coefficients), VQ(Vector Quantization), Zero crossing feature .

I. INTRODUCTION

Speech recognition is a wide topic of interest and is looked upon as sophisticated problem. Speech recognition improves productivity, solve problems and changes approach we run our lives using modern features. However, present methods have been able to achieve impressive degree of accuracy. Speech Recognition is more innovative and active area of research for last many years. It is because of the voluminous range of its potential applications. The process of speech recognition involves several features like MFCC, VAD, LPC, HMM, Zero crossing features and various system models[1]. With the development and maturity of speech recognition technique, speech becomes an important part of man-machine interface. This speech inter-face is increasingly used in office automation, factory automation and home automation. Speech Recognition is also called automatic speech Recognition system [2].

Mel-Frequency Cestrum Coefficient (MFCC) due to its excellent performance compared with other methods in many applications. However, it needs a great deal of processing for digital signals to perform MFCC, which will result in great burdens on real-time system. In general, the design of embedded system has two requirements: one is the request of real-time, while the other is the simplification of memory usage. so the signal represented in time domain should be converted into frequency-domain representation by FFT[3].This paper proposes solution for providing speech interface to electronic devices in a home. The Open source hardware is used for implementing speaker dependent speech recognition system for capturing vocal commands for operating the appliances and the microcontroller serves as the

main interface between hardware and the control circuit handling the appliances [4].The MFCC algorithm is used to simulate feature extraction module. Using this algorithm the Cepstrum coefficients are calculated of Mel frequency scale. VQ (Vector Quantization) method will be used for reduction of amount of data to decrease computation time. In the feature matching stage Euclidean distance is applied as similarity criterion[5].In this paper we propose speech recognition system which will have more accuracy.

THE PROPOSED SYSTEM MODEL

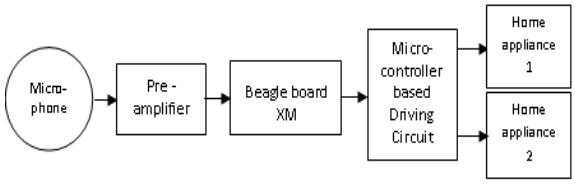

Fig.1. The proposed system model

Input signal is givens through Microphone and it will converts speech into electrical signal. The preamplifier extends the signal from source sound to signal the strength and convert the signal into high voltage. This signal gives to Beagalboard XM. Then through Beagalboard XM interface Arduino Uno and it will operates the appropriate relays scanning its digital input. It then processes and identifies the spoken word. These relays then in turn operate the corresponding appliances like; lamps, fan, dishwashers, clothes washers, dryers, microwaves, refrigerators, freezers, etc., switching them either ON or OFF.

II. THEORY AND FLOW DIAGRAM

The Mel-scale used in this work is to map between linear frequencies scales of speech signal to logarithmic scale for frequencies higher than 1 kHz. This makes the spectral frequency characteristics of signal closely corresponding to the human auditory perception. The Mel-scale frequency mapping is formulated as

$$mel(f) = 2595 \times log10(1 + fin/700)$$

(1) in which *mel*(f) is the perceived frequency and *fin* is the real linear frequency in speech signal.

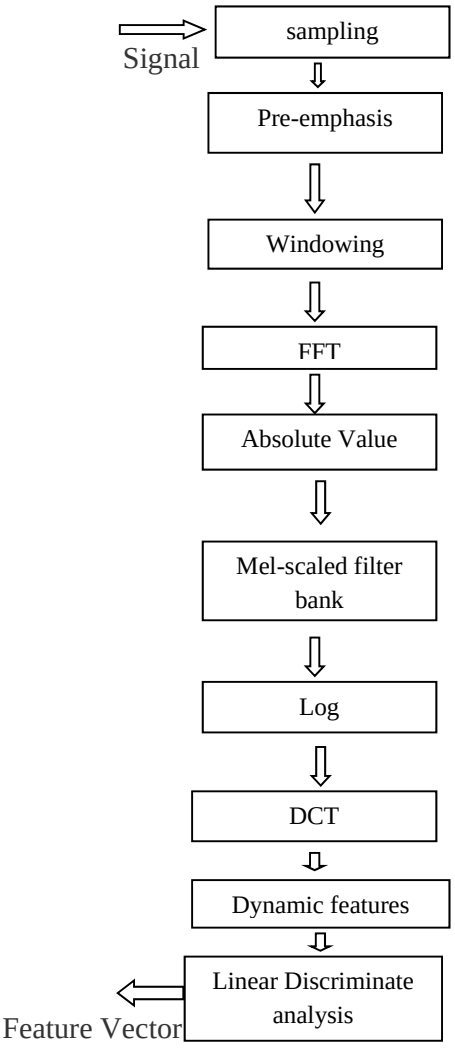

Fig2.Flow of MFCC

Steps to calculate the Coefficient

1. Frame the signal into short frames
2. For each frame calculate the period gram estimate of the power spectrum
3. Apply the Mel filter bank to the power spectra, sum the energy in each filter
4. Take the logarithm of all filter bank energies
5. Take the DCT of the log filter bank energies
6. Keep DCT coefficients 2-13, discard the rest

Speech Recognition Platform

After the feature of speech signal is retrieved, the speech signal can then be recognized in recognition session so that the home appliances are operated. The output of hardware and Arduino Uno kit will perform switching functionality. Before doing speech recognition on the recognitiosn platform, we need to do some processing of the speech signal, such as end-point detection, pre-emphasis, multiplication of Hamming window and retrieval of feature, which are called pre-processing for speech signal Since

the speech signal belongs to time-varying signal which is complicated and the signal need to be sliced into many small frames (audio frame).

III. MATHEMATICAL ANALYSIS AND CALCULATION

A Pre-emphasis

Pre-processing for Speech Signal Spreading via air, the magnitude of speech signal will reduce as the frequency rises. In order to compensate the attenuated speech signal, we put the signal through a high-pass filter to recover the signal. The difference equation governing high-pass filter is as follows:

$$S(n) = X(n) - 0.95X(n-1), 1 \le n \le L \quad (2)$$

In Equation (2), $S(n)$ represents the signal that has been processed with high-pass filter, while $X(n)$ represents the original signal, and L is the length (number of samplings) of each audio frame.

B *Hamming window*

The purpose of applying Hamming window to the signal is to prevent the non-continuity in the two ends of the audio frame, and to avoid the nuance of front and back audio frames when analyzed. The non-continuity of the audio frames can be eliminated by multiplying the Hamming window with the audio frames, because this process can make each audio frame more centered on the frequency spectrum. The following Equations (3) and (4) are the function of Hamming window and the result of the multiplication from Hamming window and audio frame, respectively.

$$W(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{L-1}\right), 0 \le n \le L - 1 \quad (3)$$

$$F(n) = W(n) \times S(n), \quad (4)$$

in which $S(n)$ is a frame of speech signal, $W(n)$ is the Hamming window, and $F(n)$ is the result of audio frame multiplied by Hamming window.

C *Retrieval of feature*

Suppose the speech sound signal is sampled at the rate of 8k, there are 8,000 sampling points to be processed per second. Such tremendous data and the complexity in speech signal make the process for the speech recognition hard. Therefore, the properties of a speech signal, i.e., the features of a speech signal, are extracted for the speech recognition. And then, the speech recognition can be performed according to these features. This will massively reduce the number of sampling point to be processed. In speech recognition, the methods commonly used for extracting the feature of speech signal are time-domain analysis and frequency-domain analysis. First, each audio frame is transformed to frequency domain by FFT. Due to masking effect in sound, the energy in each frequency domain will be multiplied by a triangle filter as follows:

$$B_m(k) = \begin{cases} 0, & k < f_{m-1}, \\ \frac{k - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} \le k \le f_m, \\ \frac{f_{m+1} - k}{f_{m+1} - f_m} & f_m \le k \le f_{m+1}, \\ 0, & f_{m+1} < k, \end{cases} \quad (5)$$

Where m is the number of the filters. After accumulating and applying the log function, we can get a energy function

$$Y(m) = \log\left[\sum_{k=f_{m-1}}^{f_{m+1}} |X(k)| B_m(k)\right] \quad (6)$$

Finally, the MFCC can be obtained by applying the Discrete Cosine Transform on m pieces of $Y(m)$ as:

$$c_x(n) = \frac{1}{M}\sum_{m=1}^{M} Y(m)\cos\left[\frac{\left(m-\frac{1}{2}\right)n\pi}{M}\right], \quad (7)$$

Where, $cx(n)$ is MFCC.

D Fast Fourier Transform (FFT)

If Discrete Fourier Transform (DFT) is used to transform time-domain signal to frequency domain signal in the calculation of MFCC, the burdens on computation time will be too huge to have a real-time application. So, we use FFT to increase the speed. However, due to the limitation of FFT, the sampling points of each audio frame should be limited in $2n$ times.

The transformation of the discrete signal from time domain to frequency domain by DFT is described as follow

$$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn}, \quad 0 \le k \le N-1, 0 \le n \le N-1 \quad (8)$$

Where $W_N = e^{\frac{-j2\pi}{N}}$, N is the number of sampling points in an audio frame.

For the calculation of DFT, a higher efficiency can be obtained by decomposing and computing above equation. So DFT has more number of computations. During this FFT, both the symmetric and periodic properties of the complex number index $W_N^{kn} = e^{-j\left(\frac{2\pi}{N}\right)kn}$ are used. The decomposition of the algorithm is

based on composing the sequence $x[n]$ into many small sequences. Hence, it is called Time Division Algorithm. The DFT in above Equation is decomposed as

$$X[k] = \sum_{n=0}^{\frac{N}{2}-1} f[n]\, W_{\frac{N}{2}}^{nk} + W_N^k \sum_{n=0}^{\frac{N}{2}-1} g[n] W_{\frac{N}{2}}^{nk} \quad (9)$$

$$= F[k] + W_N^k G[k]$$

in which $f[n] = x[2n]$ and $g[n] = x[2n+1]$ are the even sampling and odd sampling of $x[n]$ respectively.

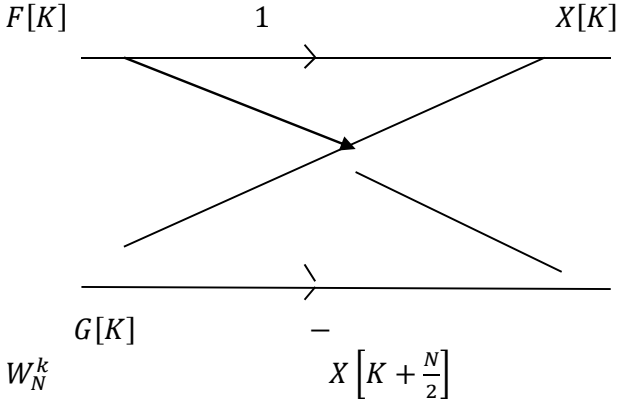

Fig 3.The time Division for FFT

Figure 3 shows the time division for FFT. The multiplication complexity $N \times N$ for original DFT can then be reduced to be $\frac{N}{2}\log_2 N$.

E Vector Quantization

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector generated from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the cancroids of the distribution. The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible. By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then used to make the recognition decision.

F Euclidean Distance

In the speaker recognition phase, an unknown speaker's voice is represented by a sequence of feature vector {x_1, x_2,…. x_i), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance. The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. The formula used to calculate the Euclidean distance can be defined as follows: The Euclidean distance between two points P = (p_1, p_2…p_n) and Q = (q_1, q_2...q_n), is given by

$$\sqrt{(p1-q1)2+(p2-q2)2} + \cdots (pn-qn)2$$

$$= \sqrt{\sum_{i=1}^{n}(pi-qi)2} \quad (10)$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person.

IV. Results

Methods adopted for implementation of speech recognition on Matlab Simulink. Here we have implemented this on PC same will be implement on Begalboard.

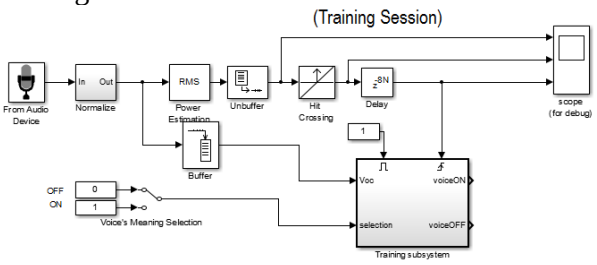

Fig 4. Training Session for ON and OFF

In the training session first we train kit by ON & OFF words. The recorded sound data from audio device is down sampled by an integer k. RMS block computes RMS value along specified

direction of input. At the same time the signal is applied to buffer which converts scalar sample to frame output at lower rate. RMS output is unbuffered i.e. it converts frame to scalar sample output at higher rate.

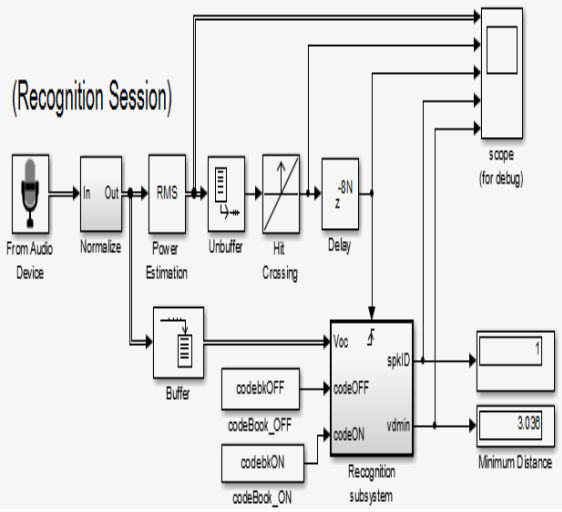

Fig 5.Recognition Session for ON

If the input signal to hit-cross block crosses offset value in specified direction the block output is 1 at the crossing time which is considered as voice else output is noise. The output from hit-cross is delayed and applied to training subsystem.

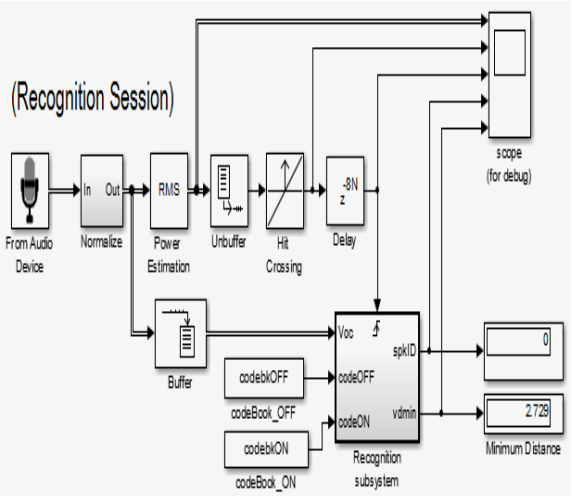

Fig 6. Recognition Session for OFF

Table 1:Result Table

Commands	Euclidean Distance	Display
ON	3036	1
OFF	2729	0
ON	3066	1
OFF	2792	0
ON	3074	1
OFF	2922	0
ON	3038	1
OFF	2732	0
ON	3036	1
OFF	2733	0

V. CONCLUSION

This paper addressed the principle and algorithm of MFCC and Vector Quantization. In the MFCC frame the signal is divided into short frame and each frame will calculate the power spectrum. Then the mel filterbank is applied to power spectrum. Finally the logarithm of all filterbank energies is taken and after taking DCT coefficients. The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients) and VQ. A VQ codebook is generated by clustering the training feature vectors. In Recognition stage, a distortion measure which is based on the minimizing the Euclidean distance used when matching an unknown speaker.

The paper also presents speech recognition algorithm using Matlab Simulink approach and using Begalboard XM. The simulations as well as experimental results of the hardware circuit are included. These results indicate that the home appliances can be operated reliably with voice commands. The proposed method also finds promising applications in robot control and helpful for industries where there is immense danger in operating the system manually.

REFERENCES

[1] Ms. Deepali . Y. Loni, “DSP Based Speech Operated Home Appliances Using Zero Crossing Features Signal Processing”, An International Journal (SPIJ), Volume (6) : Issue (2) : 2012, pp.52

[2] Bin Gao, "Wearable Audio Monitoring: Content-Based Processing Methodology and Implementation", *IEEE, and Wai LokWoo, Senior Member, IEEE* transactions on human-machine systems, vol. 44, no. 2, april 2014

[3] Shing-tai pan, Chih-Chin Lai, "International Journal of Innovative Computing, Information and the implementation of speech recognition systems on fpga-based embedded systems with soc architecture", Control ICIC International, c 2011 ISSN 1349-4198 Volume 7, November 2011, pp. 6161{6175}

[4] Sandipan Chakroborty, "Improved Closed Set Text-Independent Speaker Identification", combining MFCC with Evidence from Flipped Filter Banks International Journal of Signal Processing 4;2, © www.waset.org Spring 2008

[5] 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)

[6] Kashyap Patel and R.K. Prasad, "Speech Recognition and Verification Using MFCC & VQ" International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013

[7] The MathWorks, "Signal Processing Toolbox User's Guide", Version 6, 2003.

[8] Tiago Duarte, Rafael Prikladnicki, Fabio Calefato, and Filippo Lanubile, "Speech Recognition for Voice-Based Machin translation", Published by the IEEE Computer Society, Issue No. 01, Jan.-Feb. (2014 vol.31), pp: 26-31

[9] Mahdi Shaneh and Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ algorithms", World Academy of Science, Engineering and Technology, 2009.

.