International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 4, April - 2014

Interactive Home Automation System using Speech Recognition

Shaishav Pandya
Student, Electronics and Telecommunication Department
KJSCE, Mumbai, India

Om Kale
Student, Electronics and Telecommunication Department
KJSCE, Mumbai, India

Nazim Shaikh
Student, Electronics and Telecommunication Department
KJSCE, Mumbai, India

Skanda Vishwanath
Student, Electronics and Telecommunication Department
KJSCE, Mumbai, India

***Abstract*— In this paper we intend to describe a home automation system which uses speech commands uttered by a user to control various home appliances. The main goal of our speech recognition system is to analyze, extract and characterize information from the command uttered by the human entity. This proposed system combines both the matlab software for speech recognition and the Arduino micro-controller that is used to provide control signals to the relays, which will control various devices which are part of the home automation system. Our system will work more efficiently depending upon how well it is trained. The proposed system uses Mel Frequency Cepstrum Coefficients for feature extraction and uses vector quantization (VQ) output of those MFFCs for feature matching. The proposed system has been tested with and without known speakers as well as with and without addition of noise. The proposed system displayed an average efficiency of 75.835% for the recognition of the speech command.**

Keywords—Speech Recognition, Home automation, Mel Frequncy Cepstral Coefficient, feature extraction, Vector Quantization, feature matching.

I. INTRODUCTION

The manual buttons have always been an extremely popular medium for turning on or off various machines. But there exists numerous problems with this button technology, if considered from handicapped or senior citizen point of view. This galvanized the engineers to come up with solutions to solve the problem. New types of switching devices were created like remote controlled switches and Bluetooth controlled switches. But again, they come with a drawback for physically challenged people as well as the elderly class of the society as they find it extremely difficult to use such remote control devices. Thus, we propose of developing a home automation system which uses speech commands which will be uttered by a user. Speech is a natural mode of communication for all humans. The fact that speech gives "the human aspect" to communication with machines is without any denial. Hence we believe that speech brings about solutions to most of the problems mentioned. Moving forth, the current scenario in this industry is that home automation systems are based on strong algorithms that work independently. The user does not have any control over it once designed and implemented. Even if there is control, most of it is based on continuously changing instructions in real time and rewriting the code. Also these mechanical aspects do not give a feeling of interaction that today is what everyone craves for in this isolated lifestyle. Research in this automation area using speech and there successful implementations have also been made. However the design constraints are still complex and these systems are sometimes unreliable as pitch changes from person to person. Our main objective would thus be in reinventing this technology by making it more simple and interactive. Our system will work more efficiently depending upon how well it is trained. .It is based on Mel frequency cepstrum coefficients as feature vectors and Vector quantization for feature matching by finding the speech input's Euclidean distance from samples present in the database. The output of which will give a command that will be used by Matlab to instruct the microcontroller to perform the corresponding action. The proposed system was developed and tested on Matlab 2013a environment.

II. PROPOSED SYSTEM OVERVIEW

As mentioned previously, the home automation system that we are trying to build is based on speech recognition technique to control the various appliances used in home.

The block diagram is shown in figure 1.The heart of the system is the computer, which will be used for the speech processing i.e. speech recognition purpose. Initially the user will give a command which will be read by the computer using a microphone. The system will process the command and notify the user about the same whether it is an invalid command or valid one. After that the computer will give a signal to the microcontroller to perform the action corresponding to the command. To take an example, let's say the user gave a command to turn on a fan. The microcontroller will not only turn it on, but it will set the speed of the fan depending upon the ambient temperature which it will sense using an appropriate sensor. Now the requirement of fan speed, and for that matter of any other device, varies from person to person. So the proposed system will ask the user

whether the speed is adequate or not. Depending on the user's new command, the system will increase, decrease or do nothing to the fan's speed. Thus the device is controlled by the sensor output shown in dark black line while the users

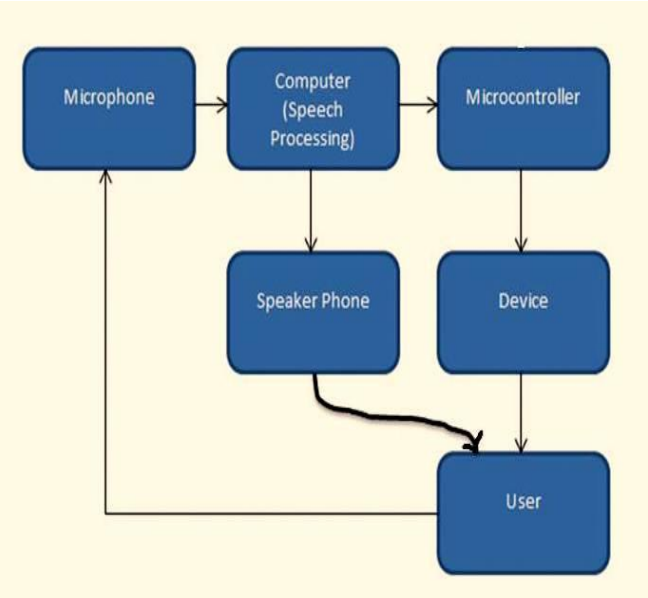

Fig. 1. Block diagram of proposed Home automation system.

command depends on what the computer is asking the user, again shown in dark black line in fig 1.

III. VOICE ACTIVITY DETECTOR

Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected. The main uses of VAD are in speech coding and speech recognition. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session: it can avoid unnecessary coding/transmission of silence packets in Voice over Internet Protocol applications, saving on computation and on network bandwidth.[1]

VAD is an important enabling technology for a variety of speech-based applications. Therefore various VAD algorithms have been developed that provide varying features and compromises between latency, sensitivity, accuracy and computational cost. For our project, we are using algorithm for isolated word detection by [2]. The algorithm uses short-time energy as basic parameter for reliable islands detection and zero-crossing rate for refinement. During the reliable islands searching procedure two thresholds are used. The refinement procedure is designed as backward searching after the reliable islands detection procedure and according to its results. Calculation of thresholds is made according to the statistics of noise: maximum and minimum of energy on the non-speech period (first 100 msec. of the record) and mean and standard deviation of zero-crossing rate.

IV. MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC)

The Mel frequency is a short term representation of power spectrum of a sound. The MFCC are the coefficients that collectively make up the Mel frequency cepstrum. The main difference between cepstrum and Mel frequency cepstrum is that, the frequency bands are equally spaced on the Mel scale which approximates the human auditory scale more accurately and closely than linearly spaced frequency bands used in normal cepstrum. The cepstrum is a common transform used to gain information from a person's speech signal. It can be used to separate the excitation signal (which contains the words and the pitch) and the transfer function (which contains the voice quality). It is the result of taking Fourier transform of decibel spectrum as if it were a signal.[3-8]

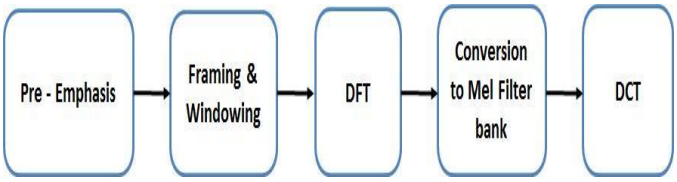

Fig. 2. Block diagram of steps involved in MFCC calculation

Mathematically,

Cepstrum of signal = IFT[log{FT(the windowed signal)}]

The Procedure, shown in figure 2, is described stepwise below:

A. Pre-Emphasis

The speech signal may have frequency components that fall of at high frequencies. The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.[5]

The speech signal y(n) is sent to a high-pass filter:

$y(n) = x(n) - a*x(n-1)$

where y(n) is the output signal and the value of a is usually between 0.9 and 1.0.

B. Framing and Windowing

An audio signal is constantly changing, so to simplify things we assume that on short timescales the audio signal doesn't change much (when we say it doesn't change, we mean statistically i.e. statistically stationary, obviously the samples are constantly changing on even short time scales). This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.[7]

Windowing is a technique used to shape the time portion of your measurement data. If we consider a rectangular it produces an abrupt change in the signal, which usually distorts the analysis. Hence, it is desirable to use a tapered window such as Hamming. The other reasons being as follows:

- To decrease the spectral distortion created by the overlap

- To minimize errors produced by FFT

The Hamming window also improves the sharpness of harmonics and removes discontinuities on the edges.

Hamming window is defined by:

$w(n) = 0.54 - 0.46 \cos(2pn/(N-1)), \quad 0 \leqq n \leqq N-1$
N= No. of samples in each frame

C. Fast Fourier Transform

After the windowing, Fast Fourier Transformation (FFT) is calculated for each frame to extract frequency components of a signal in the time-domain. FFT is used to speed up the processing. It uses Radix 2 algorithm. It reduces the number of calculations as compared to the conventional DFT.[6]

The Fast Fourier Transform converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples $\{x_n\}$ as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\left(\frac{2\pi}{N}\right)nk} \quad (k = 0,1,\ldots,N-1)$$

For further processing, we are interested in the power spectrum of the signal, and can compute the squares of the absolute values, |X(k)|^2. Due to periodicity and symmetry of X(k), only values |X(0)|^2...|X(N/2)|^2 are used for further processing, giving a total number of N/2+1 values. It should be noted that |X(0)| contains only the DC offset of the signal and therefore provides no useful information for our speech recognition task.

D. Mel Filter Bank

The human ear has high frequency resolution in low frequency parts of the spectrum and low frequency resolution in the high frequency parts of the spectrum. Here in this case the coefficients of power spectrum are transformed to reflect frequency resolution of the human ear. The One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale.[8,9]

The frequency response calculation expression of Mel triangle filter group $H_l(k)$ as follows :

$$H_l(k) = \begin{cases} \dfrac{k - o(l)}{c(l) - o(l)}, & o(l) \le k \le c(l) \\ \dfrac{h(l) - k}{h(l) - c(l)}, & c(l) \le k \le h(l) \\ 0 & others \end{cases}$$

The filter bank (see figure 3) has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The number of mel spectrum coefficients, K, is typically chosen as 12 or 20. This filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle-shape windows in the figure on the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

We have used triangular band pass filters because of the following reasons:

1. Spectrum can be smoothed to eliminate harmonic effect and show voice formant characteristics.

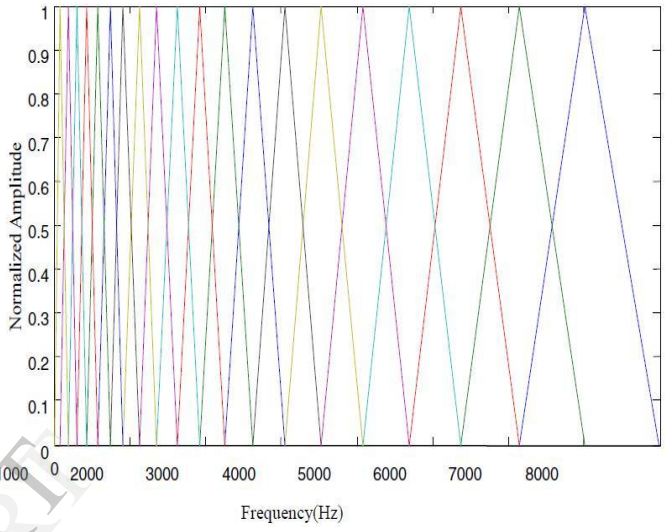

Fig. 3. Mel Frequency Scaling Filter Groups

2. We can get a higher concentration of characteristic parameters.

E. Cepstrum

The parameters that thus obtained in the last step, i.e. the log Mel spectrum is in frequency domain. In this step, the same is then converted to time domain. The result is called Mel frequency cepstrum coefficients. The Mel spectrum coefficients are real numbers. Hence they are actually converted to time domain using discrete cosine transform given by the equation below:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \qquad k = 0,\ldots,N-1.$$

V. VECTOR QUANTIZATION AND FEATURE MATCHING

The Vector Quantization (VQ) is the fundamental technique used in speech coding, image Coding and speaker recognition. These techniques are applied firstly in the analysis of speech where a large vector space is mapped into a finite number of regions in the space [12].

In VQ, an ordered set of signal samples or parameters can be efficiently coded by matching the input vector to a similar pattern or code vector (code word) in a predefined codebook.

VQ involves the process of taking a large set of feature vectors of a particular user and producing a smaller set of vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. The feature vector may represent a number of different possible speech coding parameters including linear predictive coding (LPC) coefficients, Mel Frequency cepstrum coefficients. It is used since it would be highly impractical to represent every single feature vector in feature space that we generate from the training utterance of the corresponding speaker. While the VQ algorithm does take time in computation, but it saves a lot of time while we compare feature vectors to distinguish the commands.

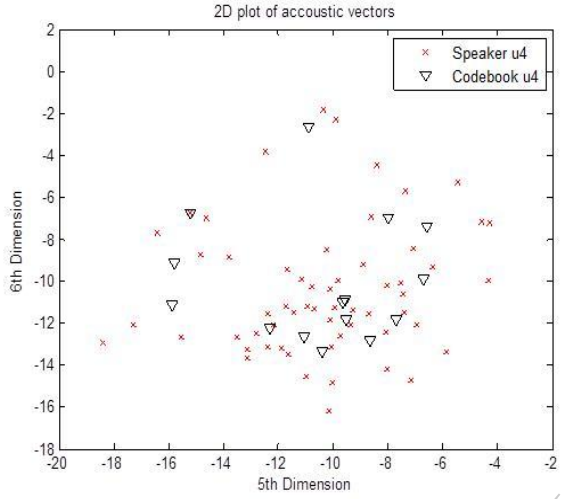

Fig. 4. Plot of trained vectors for the command 'YES'

A. LBG Algorithm

Vector quantization itself is implemented using the Linde-Buzo-Gray algorithm [10][14]. The MFCC after vector quantization is used for feature matching purpose. The working of the algorithm can be explained as follows [11, 12, and 14].

Step1: Design a 1 vector codebook; this is the centroid of the entire of the entire set of training vectors.

Step2: Increase the size of the codebook twice by splitting each current codebook y_n according to rule:

$y_n^+=(1+e)y_n$

$y_n^-=(1-e)y_n$,

Where n varies from 1 to the current size of the codebook, and 'e' is a splitting parameter (we choose e = 0.01).

Step3: Nearest neighbour search: for each training vector find the code word in the current codebook that is the closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest code word)

Step4: Centroid update: Update the code word in each cell using the centroid of the training vectors assigned to that cell

Step5: Iteration 1: Repeat steps 3 and 4 until the average distance falls below a present threshold

Step6: Iteration 2: Repeat steps 2, 3, and 4 until a codebook size of M is designed.

Intuitively, the LBG algorithm generates an M vector codebook iteratively. It starts first by producing a 1-vector code book, then uses a splitting technique on the code word to initialize the search for a 2 vector codebook and continuous the splitting process until the desired M vector codebook is obtained.

In this proposed system, we are vector quantizing the MFCC coefficient to a 16 vector codebook, which is shown in figure 4 for two (5^{th} and 6^{th}) dimensions of MFCCs. The database of sample voice Vector quantized MFCCs is generated. When the user input voice command, its vector quantized MFCCs are compared with that available in the database. This is done by measuring the distortion distance of two vector sets based on the Euclidean distance. The formula used to calculate the Euclidean distance can be defined as following:[13]

The Euclidean distance between two points P = (p1,p2…pn) and Q = (q1, q2...qn),

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

The database vector, whose Euclidean distance is least from the input vector, is selected as the detected word.

VI. RESULT AND CONCLUSION

The recognition rate for a user's input command is tabulated in Table I. The result is for both, users whose sample was taken for training purpose and making the database (known speaker) as well as the other ones (unknown speaker). The testing was carried out in a noise free environment as well as noisy environment where man-made noises (e.g. through music system) were added.

TABLE I

Recognition rate for commands doe different users (with and without external noise).

Speaker	No. of Spoken Words	No. of recognized words	Recognition Rate (%)
Known Speaker without Noise	30	26	86.67
Known Speaker with Noise	30	22	73.33
Unknown Speaker without Noise	30	23	76.67
Unknown Speaker with Noise	30	20	66.67

After first stage development of the proposed work, due to the absence of the voice activity detector several noise signals

along with the actual voice command were also picked up by the micro phone. The inclusion of the voice activity detector reduced the noise in the system. Further experiments conducted by practical evaluation method helped us understand and fix a particular zero crossing frequency and threshold amplitude for noise signals. This helped to get rid of about 90% noise in the system.

The application of the speech recognition technology, on actual home appliances showed, that greater the number of pre-recorded voice samples, higher is the system's efficiency. False alarm as well as misjudging of the input command can be avoided if the number of samples recorded and processed is high. The importance of inclusion of both male and female voices in the database was also learnt whilst the system was tested multiple times with different people. The designed system's efficiency was found to be 75.835% calculated by taking mean of the different recognition rates from Table I. In other words, nearly, three out of four times, the machine successfully recognized and implemented the voice commands that were given to it.

VII. REFERENCES

1. J. Ramirez, J. M. Gorriz and J. C. Segura (2007). Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.), InTech, Available from: http://www.intechopen.com/books/robust_speech_recognition_and_understanding/voice_activity_detection__fundamentals_and_speech_recognition_system_robustness
2. Bachu R.G, Kopparthi S, Adapa B, Barkana B.D, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal" , Electrical Engineering Department ,School of Engineering, University of Bridgeport..
3. Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", IJARET, Volume 1, Issue VI, July 2013.
4. Hongyu Xu, Xia Zhang and Liang Jia,"The Extraction and Simulation of Mel Frequency Cepstrum Speech Parameters", 2012 International Conference on Systems and Informatics.
5. Kavitha K J, "An automatic speaker recognition system using MATLAB", World Journal of Science and Technology 2012, 2(10):36-38.
6. Alan V Oppenheim, "Speech Spectograms Using Fast Fourier Transform", IEEE Volume 7, No. 8, 1970.
7. Mahdi Shaneh and Azizollah Taheri, " Voice Command Recognition System Based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology 33,2009
8. Prof. Ch.Srinivasa Kumar and Dr. P. Mallikarjuna Rao "Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization and LBG algorithm", International Journal on Computer Science and Engineering (IJCSE), August 2011
9. Lawrence Rabiner and B.H. Juang, " Fundamentals of Speech Recognition", Pearson Publication, 2009.
10. Balwant A. Sonkamble, D.D.Doye, "Speech Recognition Using Vector Quantization through Modified K-means LBG Algorithm", IISTE Vol 3, No.7, 2012.
11. Khalid Sayood ," Introduction to data compression', Morgan Kauffman Publishers, fourth edition.
12. Vincent FONTAINE, Henri LEICH and Jean HENNEBERT, "Influence of Vector Quantization on Isolated Word Recognition", Faculté Polytechnique de Mons, 31 Boulevard Dolez, B-7000 MONS, Belgium and Ecole Polytechnique Federale de Lausanne, CH-1015 LAUSANNE, Switzerland.
13. Akanksha Singh Thakur, Namrata Sahayam, "Speech Recognition Using Euclidean Distance",IJETAE, Volume 3, Issue 3, March 2013.
14. Yoseph Linde, Andres Buzo, Robert M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions On Communications, Vol. Com-28, No. 1, January 1980.