Speaker Recognition Technique for Web Browser using MFCC Algorithm and RGB Colour Detection for Mouse Curser Movement

Pushpa Rani MK[1]
[1]Senior Asst. Professor,
MITE, Moodabidri 574225.

Dr. D Vinod Kumar[2]
[2]Professor ,
VMKVEC,Salem.

Dr. Abdulla gubbi[3]
[3] Professor,
BIT, Mangalore,

Dr. Dattatraya[4]
[4] Professor,
AIET, Mangalore.

***Abstract*— Speech processing is the boon given by science and technology to human kind. The purpose of the speech is communication. The area of speech processing is just developing and shows tremendous potentialities for widespread use in the future. In this paper we have processed the speech signal with the help of the digital signal processing technique. The speech signal is given as the input will be verified using speech recognition technique using matlab. We have used Mel Frequency Cepstral Coefficient (MFCC) along with Vector Quantization (VQ) and Euclidian Distance to identify different characters.**

Keywords—Quantization; Euclidian Distance; MFCC; matlab

1 INTRODUCTION

A human can easily recognize a familiar voice however; getting a computer to distinguish a particular voice among other is a more difficult task. Immediately, several problems arise when trying to write a voice recognition algorithm. The majority of these difficulties are due to the fact that it is almost impossible to say a word exactly the same way on two different occasions. Some factors that continuously change in human speech are how fast the word is spoken, emphasizing different parts of the word, etc., furthermore. Suppose a word could in fact be said the same way on different occasions, then we would still be left with another major dilemma. Namely, in order to analyze two sound files in time domain, the recordings would have to be aligned just right so that both recordings would begin at precisely the same moment. Hence the analysis is usually done in frequency domain. We perform Voice recognition, an extremely complex visual task, almost instantaneously and our own recognition ability is far more robust than any computer's software can hope to be. We are able to recognize the voice of several thousand individuals whom we have met during our lifetime.

This current research is focused towards developing a sort of unsupervised pattern recognition scheme that does not depend on excessive geometry and computations like deformable templates. Neural network approach seemed to be an adequate method to be used for recognition due to its simplicity, speed and learning capability, also it was chosen because it has proved to be highly robust in pattern recognition tasks and because it is relatively simple to implement. This paper is based on text independent speaker recognition system and makes use of mel frequency cepstrum coefficients to process the input signal and vector quantization approach to identify the speaker. The above task is implemented using MATLAB.

2ANATOMY OF SPEECH PRODUCTION

The main aspect of speech production system is the vocal tract shape which is showed in figure 1. It is generally considered as speech production organ above the vocal folds. The vocal folds consists of laryngeal pharynx, oral pharynx, oral cavity, nasal pharynx, nasal cavity. The waves of human speech will be modified by the vocal tract, as it passes through it, therefore producing the speech. It is common in speaker identification system to make use of feature derived from the vocal tract. Thus the amplitude of the speech wave of individual is different from the other, so the MFCC is used to represent those amplitudes.

3 OBJECTIVE

The core objective of the presented system is to recognize the user's voice which is then converted into system recognizable commands. The system connects to the Internet and uses the commands to navigate through the Internet. Also control of mouse cursor movement and click events of mouse using hand movement. Hand movements were acquired by web camera based on RGB colour detection.

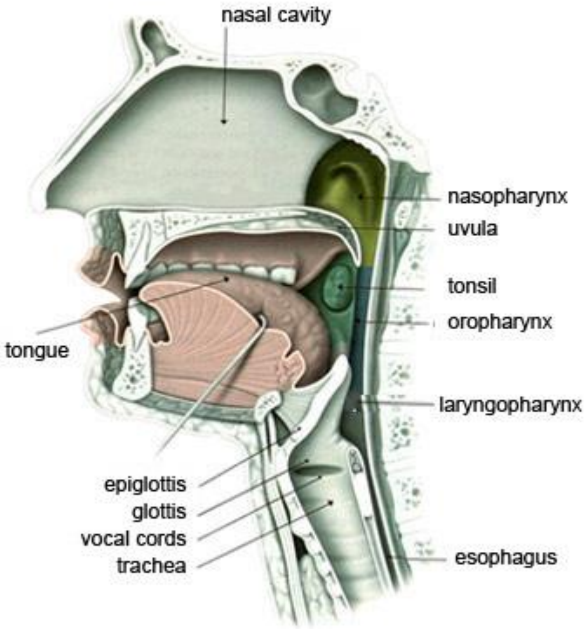

Figure 1 Vocal tract of speech production system.

4 LITERATURE SURVEY

The demography of the world population shows at rend that the elderly population world wide is increasing rapidly as a result of the increase of the average life expectancy of people. Caring for and supporting this growing population is a concern for governments and nations around the globe. Speaker identification is one of the major growing aspects that can change people lifestyle towards security and authentication.

In the proposed system [1] gives detail about speaker identification using mel frequency cepstral coefficient (MFCC) algorithm for feature extraction and for classification is done with the help of vector quantization for compression of feature and finally the speaker identification is done with the help of Euclidean distance measurement. Mel Frequency Cepstral Coefficient (MFCC) algorithm is used to recognize the speech of speaker and to extract features of speech. The K-mean clustering algorithm is used for clustering of feature vectors.

In proposed system [2] gives the design of Automation system using wireless communication and speaker recognition using MATLAB code. The speech recognition centers on recognition of speech commands stored in database of MATLAB and it is matched with in coming voice command of speaker. Feature extraction is done by using Mel Frequency Cepstral Coefficient (MFCC). It uses low-power RF Zig Bee transceiver wireless communication modules which are relatively cheap. This automation system is in tended to control lights, fan sand other electrical appliances in a home or office using speech commands like Light, Fan etc.

The proposed system [3] gives a brief details about the review of different algorithms and applications of speech recognition systems.Speech signal or the audio signal has to be properly represented. Different representation formats are MP3, WAV, AIFF, AU, RA etc. Based on the application proper format will be chosen. This representation follows feature extraction from the speech signal where only the necessary features will be extracted. Some applications like Speaker recognition systems require pattern matching (or Classification) to be done for the extracted features. Both feature extraction and pattern matching (or Classification) is necessary in many applications.

The proposed system [4] gives knowledge about the controlling of devices using input speech signals. To provide controlled access services like speech biometric, database access services, voice based dialing etc. using MATLAB tool. Recognition of voice is done with the help of different algorithms like RCC, LPC, MFCC, LPCC for feature extraction and identification of speech.

5 PROPOSED SYSTEM

There are two types in speaker recognition method, they are text dependent speaker recognition and text independent speaker recognition methods. In text dependent method, the speaker has to say key words or sentences having the same text for both training and recognition trials. Whereas, the text independent meaning that the system can identify the speaker regardless of what is being said. The goal of this paper is a real time text-independent speaker identification, which consists of comparing a speech signal from an unknown speaker to a database of known speakers. The system will operates in two modes: A training mode, a recognition mode. The training mode will allow the user to record voice and make a feature model of that voice. The recognition mode will use the information that the user has provided in the training mode and attempt to isolate and identify the speaker. The Mel Frequency Cepstral Coefficients and the Vector Quantization algorithms are used to implement the paper using the programming language Matlab.

6 METHODOLOGY

The input speech signals from different users are taken initially, then these speech signals are pre-processed to enhance the accuracy and efficiency. Then feature extraction is done with the help of Mel Frequency Cepstral Coefficient (MFCC) and a database is created for these extracted feature vectors, this process is done in training period.

Then in testing period, input for the user is taken and pre-processed, feature extraction is done and identification of unknown user is done by using K-mean clustering method and minimum distance between centroid and the feature vector is calculated. If the condition satisfies, then it is identity of the speaker. Once the match is done the given command is processed further by the system. This overall paper is done in MATLAB with the use of GUI tool.

7 FEATURE EXTRACTION AND CLASSIFICATION SPEAKER RECOGNITION SYSTEM OVERVIEW

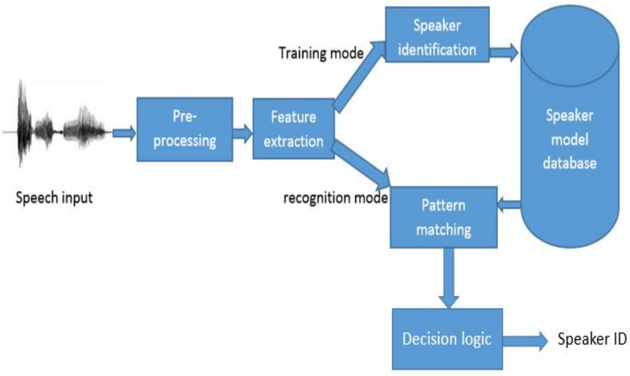

Figure 2 Schematic diagram of the closed-set speaker identification system.

Referring to the above diagram, we can see that the input speech will pass through two major stages in order to get the speaker identity. They are,
1- Feature extraction.
2- Classification and Feature matching.

FEATURE EXTRACTION

Feature extraction is a special form of dimensionality reduction, and here in this paper we need to do dimensionality reduction for the input speech ; we will do that by extracting a specific features from the speech, these features carry the characteristics of the speech which are different from one speaker to another, so these features will play the major role in paper, as our mission is to identify the speaker and make a decision that highly depends on how much we were successful in extracting a useful information from the speech in a way enables our system to differentiate between speakers and identify them according to their features matching.

On focus a little at these features and see: what are they represent, what the specific characteristics they should carry, what the algorithms that we can use to extract them.

- Feature vectors to be extracted from the input speech.
- These feature vectors represents Formants, which carry the identity of the speech.

To identify these formants here a brief discussion to the speech production in the human body is given:

The three main cavities of the speech production system are nasal, oral, and pharyngeal forming the main acoustic filter. The form and shape of the vocal and nasal tracts change continuously with time, creating an acoustic filter with time-varying frequency response. As air from the lungs travels through the tracts, the frequency spectrum is shaped by the frequency selectivity of these tracts. The resonance frequencies of the vocal tract tube are called formant frequencies or simply formants, which depend on the shape and dimensions of the vocal tract. The speed by which the cords open and close is unique for each individual and define the feature and personality of the particular voice.

So we want to extract these formants (the amplitudes of the speech wave form) and a smooth curve connecting them (the envelope of the speech waveform).The following figure shows the formants and the envelope connected them,

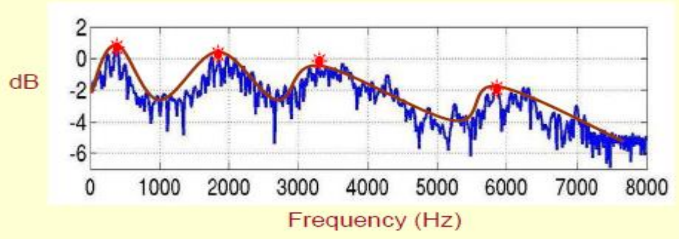

Figure 3 The envelop connecting the Formants.

These extracted features should carry some specific characteristics, they are
1. Easily measurable.
2. Vary much as possible among the speakers, but be consistent within each speaker.
3. Not change over time or be affected by the speaker's health.
4. Not be affected by background noise nor depend on the specific transmission medium.
5. Not be modifiable by conscious effort of the speaker or at least, be unlikely to affect by attempts to disguise the voice and should occur naturally and frequently in speech.

A) Pre-Processing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing covers digital filtering and speech signal detection. Filtering includes pre-emphasis filter and filtering out any surrounding noise using several algorithms of digital filtering.

Pre-emphasis filter: In general, the digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range, pre-emphasis is applied. This pre-emphasis is done by using a first-order Finite Impulse Response (FIR) high-pass filter.

In the time domain, with input x[n] and $0.9 \leq a \leq 1.0$, the filter equation

$$y[n] = x[n]-a \cdot x[n-1].$$

And the transfer function of the FIR filter in z-domain is:
$H(Z) = (1 - \alpha).(z - 1)$, $0.9 \leq \alpha \leq 1.0$
Where α is the pre-emphasis parameter.

The pre-emphasizer is implemented as a fixed coefficient filter or as an adaptive one, where the coefficient a, is adjusted with time according to the auto-correlation values of the speech. The aim of this stage is to boost the amount of energy in the high frequencies. The drop in energy across frequencies (which is called spectral tilt) is caused by the nature of the glottal pulse. Boosting the high frequency energy makes information from these higher formants available to the acoustic model. The pre-emphasis filter is applied on the input signal before windowing.

B) Framing and Windowing

First we split the signal up into several frames such that we are analysing each frame in the short time instead of analysing the entire signal at once, at the range (10-30) ms the speech signal is for the most part stationary.

Also an overlapping is applied to frames. Here we will have something called the Hop Size. In most cases half of the frame size is used for the hop size. The reason for this is because on each individual frame, we will also be applying a hamming window which will get rid of some of the information at the beginning and end of each frame. Overlapping will then reincorporate this information back into our extracted features.

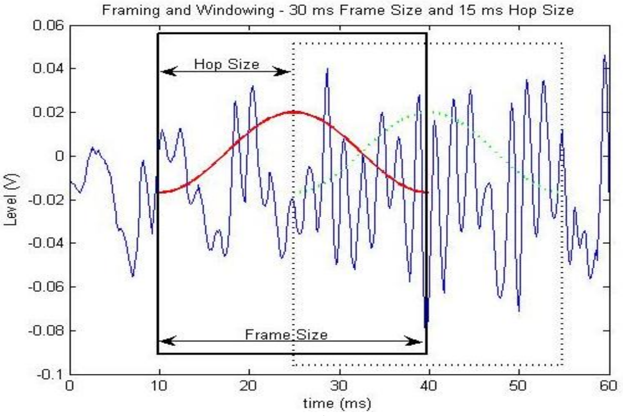

Figure 4 Representation of Framing and Windowing.

Windowing,It is necessary to work with short term or frames of the signal. This is to select a portion of the signal that can reasonably be assumed stationary. Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. The choice of the window is a tradeoff between several factors. In speaker recognition, the most commonly used window shape is the hamming window.

The multiplication of the speech wave by the window function has two effects:-

1-It gradually attenuates the amplitude at both ends of extraction interval to prevent an abrupt change at the endpoints.

2-It produces the convolution for the Fourier transform of the window function and the speech spectrum.

Actually there are many types of windows such as: Rectangular window, Hamming window, Hann window, Cosine window, Lanczos window, Bartlett window (zero valued end-points) , Triangular window (non-zero end-points), Gauss windows etc. we used hamming window the most common one that being used in speaker recognition system.

Hamming window equation is given as:

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

Then the result of windowing signal is shown below:

Y(n)=X(n).W(n)

Where, N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window

$W(n)=0.54-0.46\cos[2\pi n/(N-1)]$;$0\leq n\leq N-1$

The use for hamming windows is due to the fact that mfcc will be used which involves the frequency domain

C) Fast Fourier Transform

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 follow:

$$X_k= \sum_{n=0}^{N-1} Xne^{-j2\pi nk},$$

k=0,1,2,…,N-1

The result after this step is often referred to as spectrum.

The FFT behaves as though it was a bank of narrow-band filters followed by a bank of corresponding detectors that calculate the vector sum of all the signal components that each filter passes.

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 supports the equation below:

Y (w)=FFT [h (t)*X (t)]=H (w)X (w)

Where, X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t).

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame.

D) Mel Scaled Filter Bank

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 a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.

We can use the following formula to compute the mels for a given frequency f in Hz:

mel(f)= 2595*log10(1+f/700).

One approach to simulating the subjective spectrum is to use a filter bank, one filter for each desired Mel frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.

Mel frequency analysis,

1. Mel-Frequency analysis of speech is based on human perception experiments.

2. Human ears, for frequencies lower than 1 kHz, hears tones with a linear scale instead of logarithmic scale for the frequencies higher than 1 kHz.

The information carried by low frequency components of the speech signal is more important compared to the high frequency components. In order to place more emphasize on the low frequency components, mel scaling is performed.Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions.

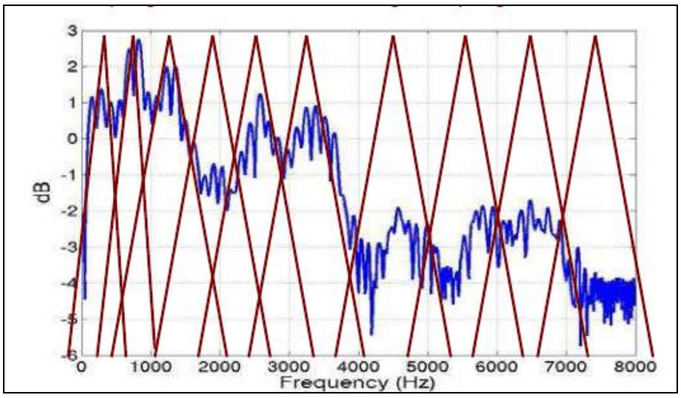

Figure 5 Mel Scale Filter Bank.

So after having the spectrum (fft for the windowed signal) we applied mel filter banks, the signal processed in such away like that of human ear response:

$$\tilde{S}(l) = \sum_{k=0}^{N/2} S(k) M_l(k)$$

Where,
S(l) :Mel spectrum, S(K) :Original spectrum, M(K) :Mel filterbank, L=0 ,1 , ..., L-1, Where L is the total number of mel filterbanks, N/2 = Half FFT size.

E) Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers(and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT).

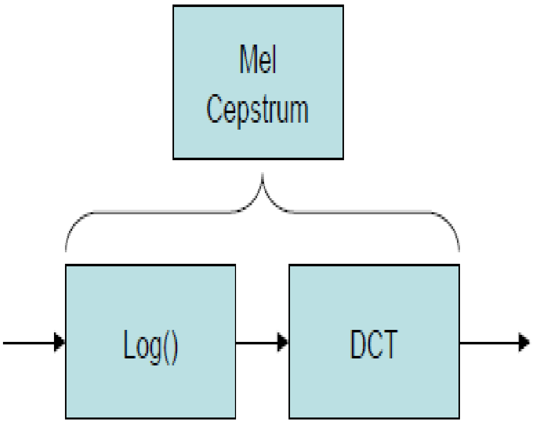

Figure 6 Mel Cepstrum Coefficient.

Since the speech signal represented as a convolution between slowly varying vocal tract impulse response (filter) and quickly varying glottal pulse (source), so, the speech spectrum consists of the spectral envelop(low frequency) and the spectral details(high frequency).

Now, our goal is to separate the spectral envelope and spectral details from the spectrum. It is known that the logarithm has the effect of changing multiplication into addition. Therefore we can simply converts the multiplication of the magnitude of the Fourier transform into addition. Then, by taking the inverse FFT or DCT of the logarithm of the magnitude spectrum, the glottal pulse and the impulse response can be separated.

8 FLOW DIAGRAM FOR RGB COLOR DETECTION FOR MOUSE TRACKING

Initially the capturing of real time video using web camera is done. Then processing of the individual image frames. After this the flipping of these each image frame. Conversion of each frame to a gray-scale image is done after flipping of the image. Colour detection and extraction of different colour should be performed, after this the images are then converted into binary images. Centroid region of this image is calculated. Tracking of mouse pointer using coordinates of centroid. According to the coding arrangement for the colour identification the movement of mouse cursor is performed through interface with the web camera [7].

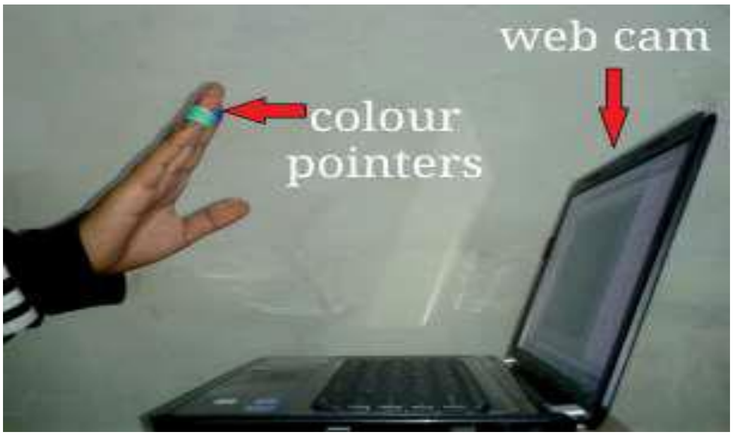

Figure 7 Capturing the video.

10 CONCLUSION AND FUTURE WORKS

The goal of this paper was to implement a text-independent speaker identification system. The feature extraction is done using Mel Frequency Cepstral Coefficients {MFCC} and the speakers was modeled using Vector Quantization technique. Using the extracted features a codebook from each speaker was build clustering the feature vectors using the K-means algorithm. Codebooks from all the speakers was collected in a database. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speaker with the speaker database.

The study reveals that as the number of centroids increases, the identification rate of the system increases. Also, the number of centroids has to be increased as the number of speakers increases. The study shows that as the number of filters in the filter-bank increases, the identification rate increases.

Our experiments in the communication lab. Environment showed that reducing the test shot lengths reduced the recognition accuracy. In order to obtain satisfactory result for real time application, the test data usually needs to be more than ten seconds long.

All in all, during this paper we have found that VQ based clustering is an efficient and simple way to do speaker identification. Our system is 90-95% accurate in identifying the correct speaker when using 30 seconds for training session and several ten seconds long for testing session.

FUTURE WORK:

There are a number of ways that this paper could be extended, these are a few:

- A statistical approach such as Hidden Markov Modeling or Dynamic Time Warping could be used for speaker modeling instead of vector quantization technique in order to improve the efficiency of the system.
- The system could be improved so that it can works satisfactorily in different training and testing environments.
- Noise is a really big deal, it can increase the error rate of speaker identification system. So, use of noise cancellation and normalization techniques to reduce the channel and the environment effects is recommended.
- Also, voice activity detection should be done. All of these can improve the recognition accuracy.
- Moreover, the system could be developed to perform the verification task. And so, an unknown speaker who is not registered in the database should be rejected.
- Speaker verification will be challenging because of highly variant nature of input speech signals. Speech signals in training and testing sessions can be greatly different due to many facts such as :
  1) People voice change with time.
  2) Health conditions.
  3) Speaking rates.
  4) Variations in recording environments play a major role.
- Therefore, threshold estimation is not an easy issue. In addition, use of impostor's data to estimate the threshold creates difficulties in real applications.
- Use of Linear Predictive Coefficient (LPC), Linear Predictive Cepstrum Coefficient (LPCC), Human Factor Cepstrum Coefficient (HFCC) for feature extraction process. So that better accuracy and efficiency for feature vectors can be achieved.

REFERENCES

[1] Aseem Saxena,Amit Kumar Sinha, Shashank Chakravarti, Surabhi Charu, "*Speaker Recognition using MFCC and Vector Quantization Model*". International Journal of Advances in Computer Science and Cloud Computing, ISSN: 2321-4058, Volume-1, ISSUE-2, NOV-2013.

[2] S. R. Suralkar, AmolC.Wani, Prabhakar V Mhadse, "*Speech Recognized Automation System Using Speaker Identification through Wireless Communication*" IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p-ISSN: 2278-8735. Volume 6, Issue 1 (May. - Jun. 2013), PP 11-18.

[3] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "*Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC), Dynamic Time Warping (DTW) Techniques and Linear Predictive Coding (LPC)*", JOURNAL OF COMPUTING, pp 138-143, VOLUME 2, ISSUE 3, MARCH 2010.

[4] E. Darren Ellis Department of Computer and Electrical Engineering – University of Tennessee, Knoxville Tennessee 37996,"*Controlling of Device through Voice Recognition Using Matlab1*" ISSN NO: 2250-3536 VOLUME 2, ISSUE 2, MARCH 2012.

[5] A. Khan, et al., "Speech Recognition: Increasing Efficiency of Support Vector Machines," International Journal of Computer Applications vol. 35, dec 2011.

[6] Rashmi.C.R, Assistant Professor, Department of CSE,CIGT,Gubbi, Tumakur, Karnataka, India, "*Review of Algorithms and Applications in Speech Recognition System*". International Journal of computer Science and Information Technologies, Volume-5(4), ISSN: 0975-9646, NOV-2014.

[7] Abhik Banerjee, Abhirup Ghosh, Koustuvmoni Bharadwaj, Hemanta Saikia, Department of Electronics & Communication Engineering, Sikkim Manipal Institute of Technology East Sikkim, India "*Mouse Control using a Web Camera based on Colour Detection*". International Journal of Computer Trends and Technology (IJCTT) – volume 9 number 1– Mar 2014.

[8] Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur, "*The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans: A Review*", IJARCSSE, Volume 2, Issue 2, February 2012.

[9] W. Astuti, A.M Salma, A.M. Aibinu, R. Akmeliawati, Momoh Jimoh E.Salami, "Automatic Arabic Recognition System based on Support Vector Machines (SVMs)", IEEE, 2011.