VACHANTAR LOKBHASHA A SPEECH TO TEXT CONVERSION FOR MARATHI

Ms.Archana V. Chechare
Computer Engineering
SRES College of Engineering, Kopargoan, Ahmednagar
Savitribai Phule Pune University
Archanachechare27@gmail.com

ABSTRACT

Speech processing has always been so important application area of digital signal processing. The various fields are available for researching in speech processing that are speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of Vachantar Lokbhasha, A Speech to text conversion for Marathi is to recognize speaker's speech and then convert it to text. This text will be saved to a file and that file can be further used. Automatic speaker recognition is to extract and recognize the information about speaker's speech. Feature extraction is the first step for this. Many algorithms are stated by the researchers for feature extraction. In this work, the Mel Frequency Cepstral Coefficient (MFCC) feature extraction algorithm has been used for designing this system. In this Artificial Neural Networks will be used for feature classification.

Keywords: *Feature extraction, Mel frequency cepstral coefficients (MFCC), Speaker recognition, Neural Network*

I. INTRODUCTION

In this"Vachantar-Lokbhasha:Speech to Text Conversion for Marathi", such a system is developed which recognizes words from the voice of the user and converts to text in marathi. Speech recognition [8] is the process of recognizing the spoken words of person on the basis of information in speech signal. The acoustical parameters of speech signal used in recognition tasks have been studied and investigated so far, and making it able to be categorized into two types of processing domain: One is spectral based parameters and another is dynamic time series. Out of them the most popular spectral based parameter used in recognition approach is the Mel Frequency Cepstral Coefficients called MFCC. In Speech recognition technique it is becoming easy for the speakers voice to be used in verifying their identity and providing access to services such as banking by telephone, telephone shopping, voice dialing, voice mail, database access services, information service, remote access and security control for the confidential information areas to computers.

The human speech is made up of various discriminative features that can be used to identify speakers [6]. Speech contains significant energy ranging from zero frequency to around 5 kHz. The property of speech signal changes as a function of time. For studying the spectral properties of speech signal the Fourier representation is used. However, the temporal properties of speech signal such as zero crossing, correlation, etc are assumed constant over a short period. That means its characteristics are short-time stationary or varying. Therefore, using hamming window, speech signal is framed into a number of blocks of short duration so that normal Fourier transform can be used. In this the MFCC features has been used for designing and developing a speech recognition system.

II. LITERATURE SURVEY

In the area of speech to text conversion following systems are already available,

1. First system is Microsoft Cortana [12], which is an intelligent personal assistant created by Microsoft for windows phone8.1. Cortana features include being able to set reminders, recognize natural voice without the user having to input a predefined series of commands and answer questions using information from Bing. But this system is limited for certain applications only.

2. Another system available is Google Voice Search [13]. Google voice search is a tool from Google labs that allowed people to use their phone to make a Google query. Google voice search or search by voice is a Google product that allows user to use

Google search by speaking on a mobile phones or computer, i.e. have the device search for data upon entering information on what to search into device by speaking. In this user don't have to submit query by typing it, rather it can be submitted by speaking.

3. Google Maps with Voice search [14] is also one of the available systems in which Google added voice search to the blackberry pearl version of Google maps for mobile allowing pearl users to speak their searches instead of typing them. It has features like in-navigation voice control, faster access to voice input from main maps screen and elevation change information for bicyclists.

4. Vachantar-Rajbhasha [15] is a speech to text translation system which uses speech recognition and machine translation. In this speech recognition engine takes English speech as an input through microphone and generates typed English text as output. This output becomes input to translation engine that translates it to Hindi.

III. IMPLEMENTATION DETAILS

A. System Overview

In the "Vachantar-lokbhasha: Speech to text conversion for Marathi", Speaker's voice will be processed and unique features from that will be extracted by using MFCC. Then those features will be further classified by using Neural Networks. So the whole process is divided into two sub steps those are as follows

1. Speech signal pre-processing by using MFCC

Pre-processing of a signal is applying any required form of processing to the signal in time domain before the feature extraction phase .In this stage the speech signal goes through several common processes including Analog to digital (A/D) conversion, enhancement, pre-emphasis filtering and usually for SR applications silence removal or end point detection (EPD). The A/D process converts a sound wave into its digital form. There are three steps in the A/D Conversion process which are sampling, quantization and coding. The final outcome of this process is a digital version of the speech signal that can be used as well as processed by the computers. In speech recognition and processing in general, speech enhancement is conducted to suppress unwanted noise from the speech signal. For SR application removing noise increases the accuracy of the recognizer. In almost all SR application a pre-emphasis filtering step is conducted to the speech signal. The pre emphasis filter is used to emphasis the speech spectrum above 1 kHz which contains important aspects of the speech signal and equalizes the speech propagation trough air [7].

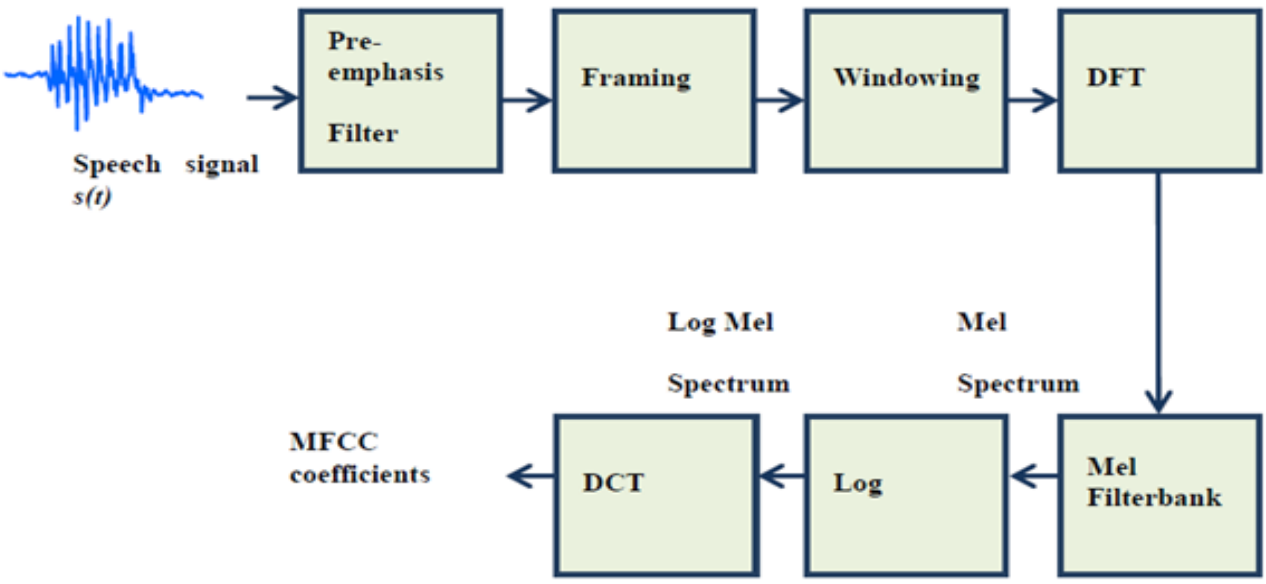


Fig. 1 Feature Extraction by using MFCC

The Mel-frequency Cepstrum Coefficient (MFCC) technique is used for creating the fingerprint of the sound files. The MFCC is based on the known variation of the human ears Critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. It is scientifically proven that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each audio signal with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. The basic formula developed for computing the Mels for a particular frequency is: mel (f) = 2595*log10 (1+f / 700). MFCC processes block diagram is shown in above Figure 1. The speech waveform is cropped to remove acoustical interference that is present in the beginning or end of the sound file. The windowing minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The Fast Fourier Transform converts each frame from the time domain to the frequency domain. In the Mel-frequency wrapping block, the signal is plotted against the Mel spectrum to mimic human hearing. In the last step, the Cepstrum, the Mel-spectrum scale is converted back to standard frequency scale. This Mel-spectrum provides a good representation of the spectral properties of the signal which is a key for recognizing & representing characteristics of the speaker. After the fingerprint is created, we will also going to create an acoustic vector. This vector can be stored as a reference in the database [7].

2. Classification By using Neural Networks

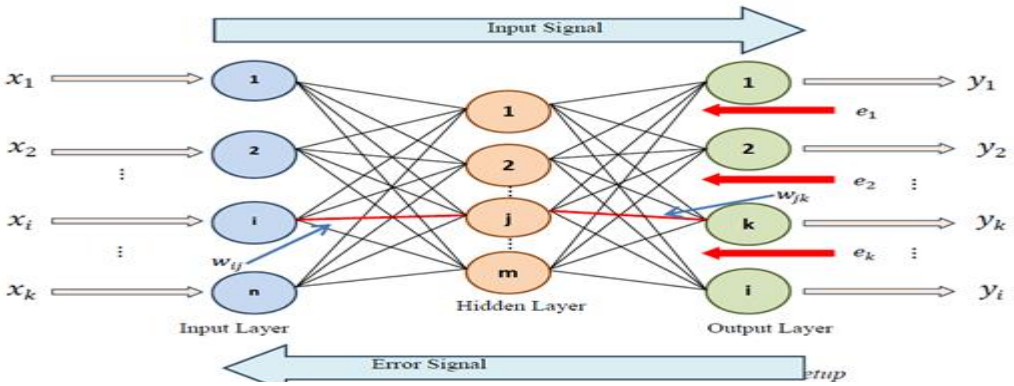


Fig.2 Classification by using Neural Networks

The second most important step in an Automatic Speech Recognition system is the classification stage. This stage includes classifying the input speech to determine whether the input speech uttered matches the desired targeted speech or not. Categories of classification schemes are statistical and artificial intelligence approaches. In this, we chose neural networks (NN) as one of the artificial intelligence approach. Neural networks (NN) are parallel distributed information processing structure with processing elements connected through unidirectional signal channels called connections. ANNs consist of simple interconnected processing elements that are called neurons that perform weighted summation of inputs [7].

B. MFCC Algorithm

Stepwise representation of MFCC Algorithm

1. Start
2. Divide audio signal into frames
3. Then apply windowing function, aim here is to model small sections of the signals that are satisfactorily stationary. The window function removes edge effect.
4. Obtain Amplitude Spectrum: DFT of each frame will be taken, we will obtain logarithm of amplitude spectrum. We discard phase information because perceptual studies have shown us that the amplitude of the spectrum is much more important than phase.
5. Take Logarithm: We take the logarithm of amplitude spectrum because the perceived loudness of a signal has been found to be approximately logarithmic.
6. Convert to Mel spectrum:
7. Take DCT: Take Discrete Cosine Transform to smooth the signal, now we will generate feature vectors
8. Do linear discriminate analysis

9. Generate feature vectors
10. Stop

Graphical representation of MFCC algorithm

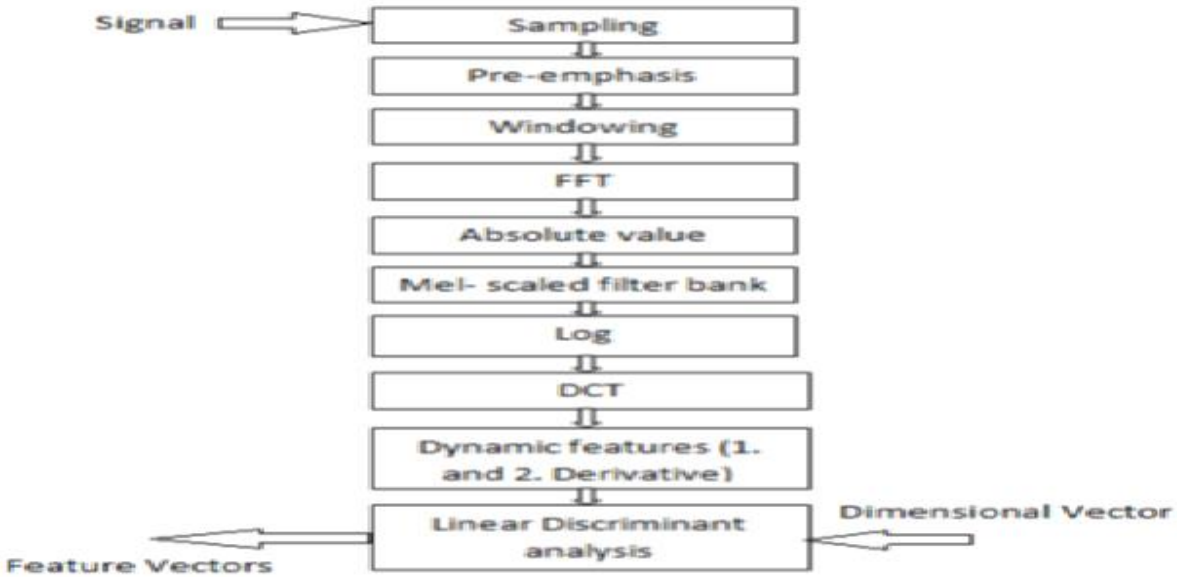


Fig. 3 MFCC algorithm

C. Mathematical Model

In general, total set of system S= I , P , R , O. where

- I is an input set
- P is a process set
- R is a rules set
- 0 is a outcome or output set

I = I1

I1 audio signal from speaker

Process Set It is the set of process. S= P1 , P2 , P3 , P4 , P5 , P6. Where

- Process P1 =analyze signal
- Process P2 =framing data
- Process P3 =windowing data
- Process P4 =generating ceptral coefficients
- Process P5 = analyze mfcc and compare
- Process P6 =recognize words

Output Set There is two output sets.

The first is, intermediate output set is denoted by:

IO= IO1, IO2, IO3, IO4, IO5, IO6.

Where

- IO1 =signal analyzed
- IO2 =signal framed
- IO3 =signal windowed
- IO4 =cesptral coefficients generated

- IO5 =MFCC generated and compared
- IO6 =words recognized

The second is final output set is denoted by O= O1.
Where O1=converted to text in Marathi

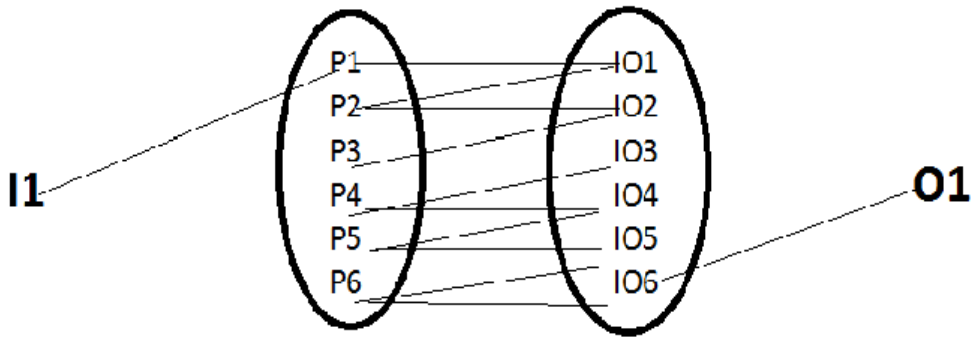


Fig. 4 Venn Diagram

D. Experimental Setup

The system is being built using Java framework (version jdk 6) on Windows platform. The JCreator/ Net beans (version 6.9) are alternatively used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application. Sometimes system's detachable microphones are required for recording voice.

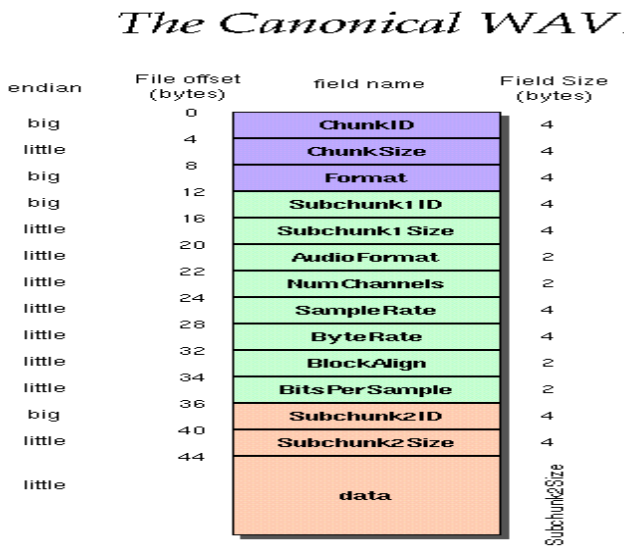
IV. RESULTS AND DISCUSSION

A.Experimental Analysis

For the experiment purpose we modeled a system with a microphone for recording voice, then primarily we provided user with certain predefined inputs and performed analysis of speech. On the basis of extracted features final classifications done and decisions are taken.

B. Module Description

Canonical format of Waveform



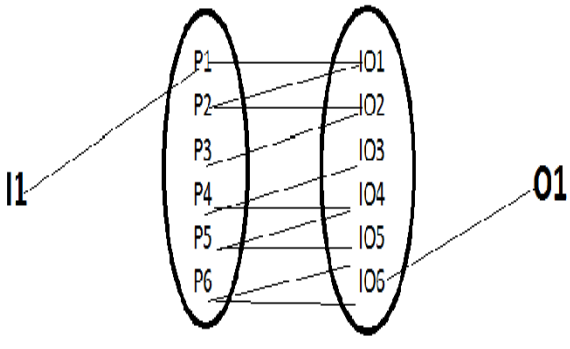


Fig. 5 Canonical form of Wave file (.wav)

As shown in above Fig. 4 our recorded audio in .wav format has above structure, where each bit has some value. So for decoding an audio each and every bit will be considered and analyzed. Further operations will be done on the bit pattern that will be given by incoming audio. Following is the representation of audio bit pattern [16]

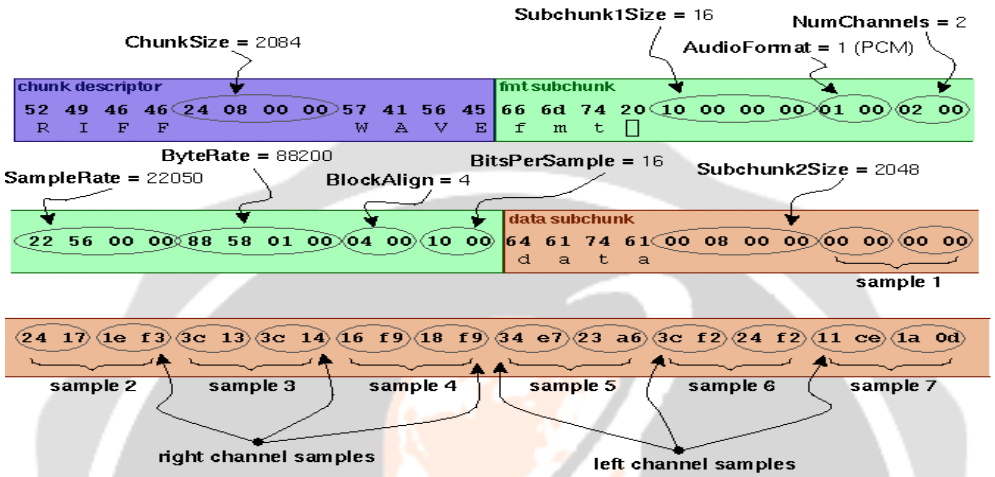


Fig. 6 Bit pattern of audio wave

So as shown in above fig. 5 bits in an audio file are represented and hence our all modules are based on this bit pattern [16].

1. Framing Module

In this as audio wave is not static in nature and hence it has dynamic nature. For this purpose for processing it and accuracy, audio is divided into smaller fractions called as frame and each frame is processed separately. Following is the output generated in this phase:

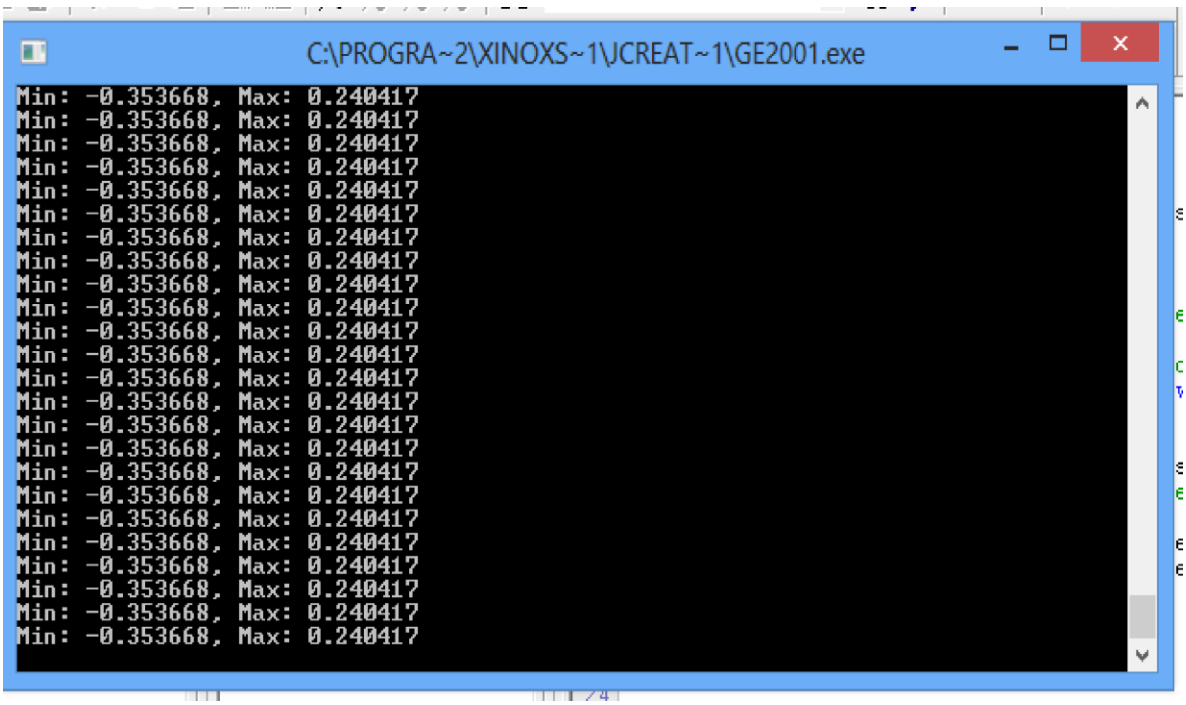


Fig. 7 Reading frames from audio signal

2. Windowing Module

As in framing module we get frames so to remove discontinuities at the start and end of each frame has to be removed. It is done by applying windowing function .In this function for rounding off the signal will be multiplied with zero.

3. Mel Frequency wrapping

As we know our audio wave is in time domain so to convert it to frequency domain FFT is taken. Our voice is not linear in nature, it is logarithmic and hence logarithm is applied on the output obtained in previous step. In this module obtained values will be rounded off by taking discrete cosine transform for smoothing the output. In this final feature vectors are generated.

4. Classification by using Neural Networks (NN)

In this module obtained feature vectors are given as an input to
Neural Networks. Then classification is performed and spoken words are recognized by comparing acquired values and previously stored values.

5. Generating Marathi text file

In this final output is generated on the basis of classification done in the previous phase. Finally output file will be generated, containing Marathi text of recognized speech from speaker.

V. CONCLUSION

This paper shows that the discussed system "Vachantar-Lokbhasha: speech to text conversion for Marathi" will be dealing with accepting speech from speaker and converting it to text in Marathi. In the first phase input signal is divided into multiple smaller blocks & each and every block is analyzed separately. Accuracy obtained in feature extraction phase will be the important factor in recognizing the speech uttered by speaker.

ACKNOWLEDGMENT

The author would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. I am thankful to the authorities of Savitribai Phule PunevUniversity and concerned members of cPGCON2016 conference, organized by, for their constant guidelines and support.I am also thankful to the reviewer for their valuable suggestions. I also thank the college authorities for providing the required infrastructure and support. Finally, I would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] H.B.Chauhan, Prof B.A.Tanawala ,"Comparative study of MFCC and LPC algorithms for Gujrati isolated word recognition", International Journal of Innovative Research in Computer and Communication Engineering, vol-3, Feb 2015.

[2] Archit Kumar,Charu Chhabra,"Intrusion detection system using Expert System(AI) and pattern recognition (MFCC improved VQA)", International Journal of Advanced Research in Computer science Management studies,Vol-2,May2014.

[3] Archit Kumar,Charu Chhabra,"Intrusion detection system using Expert System(AI) and pattern recognition (MFCC improved VQA)", International Journal of Advanced Research in Computer science Management studies,Vol-2,May2014.

[4] Nilu singh,R.a.Khan,Raj Shree, "MFCC Prosodic feature extraction techniques: A Comparitive Study", International Journal of Computer Applications(00975-8883), September 2012

[5] Shushmita Iqbql, Tahira Mehboob, Malik sikandar Hayat khiyat," Voice Recognition using HMM with MFCC for secure ATM", International Journal of Computer Science, vol- November 2011.

[6] Vibha Tiwari," MFCC and its Applications in Speaker Recognition", International Journal on Emerging Technologies 1(1):19-22,2010.

[7] T.B Adam, M.D.Salam, "Spoken English Alphabet Recognition with Mel frequency Ceptral Coefficients and Back propagation Neural Networks", International Journal of Computer Applications(0975-8807) vol-42-no12, March 2012

[8] Chadwan ittichaichareon, Siwat Suksri,Thawersak Yingthawarnsuk, " Speech Recognition using MFCC", International Journal on Computer Graphics, Simulation and Modeling(IJCGM 2012) July 2012

[9] Aldebara klatau "The MFCC", 22nd November 2005.

[10] Rohini Bhujang rao Shinde," Analyzing Childrens speech for Biometric Identification" ,International Journal of Speech and Language Processing, Vol-1, January 2011.

[11] Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh,R.N.V. Sitaram, , "Development of Indian Language Speech Databases for large vocabulary speech Recognition System", International Institute of Information Technology, Hyderabad, Hewlet Packard Labs India, Banglore.

[12] Microsoft Cortana, https://en.m.wikipedia.org/wiki/Cortana_(software).

[13] Google voice search, https://en.m.wikipedia.org/wiki/google_voice_search.

[14] Google voice maps, http://www.idigitaltimes.com/google-maps-tips-and-trics-10-little-known-voice-commands-use-googles-navigation-493297

[15] Vachantar Lokbhasha, http://cdac.in/index.aspx?id=mc_st_vachantar

[16] Wave file format, http://soundfile.sapp.org/doc/WaveFormat/