International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(2), pp.001-009
DOI: http://dx.doi.org/10.21172/1.82.001
e-ISSN:2278-621X

VOWEL RECOGNITION FOR COCHLEAR IMPLANT DEVICE TESTING

Gautami Pujare[1]

Abstract- Human listeners are better able to identify vowels on the basis of timbre characterization. Timbre is the attribute that distinguishes sounds of equal pitch, loudness and duration. It contributes to our perception and discrimination of different vowels and consonants in speech, instruments in music and environmental sounds. The Mel Frequency Cepstral Coefficient vectors are commonly used in audio analysis. They are described as timbral features as they model the short-time spectral characteristics of the signal onto a psychoacoustic frequency scale.

In this work an audio signal analysis approach based on Mel Frequency Cepstral Coefficients, for vowel recognition is proposed. Mel Frequency Cepstral analysis is utilized as a feature extraction technique. Mel Frequency Cepstral coefficients extracted for vowel phonemes are used as template and matched with audio input signal. To recognize vowels, Adaptive Network based Fuzzy Inference System (ANFIS) is used as a feature matching technique. In simulation results, vowel is recognized correctly, which can be useful in tuning cochlear implant device in medical field. It also can be used for capturing the timbre information of music signals.

Keywords- Pitch, Timbre, Cepstral coefficient

I. INTRODUCTION

Pitch and timbre are two important acoustic characteristics of audio signals. Speech perception involves the perception of many sound attributes including dynamic patterns of pitch, loudness and timbre changes. Speech signals contain a wide variety of acoustic cues from which sound timbre may be derived. At the phonetic level, timbre plays a crucial role in determining the identity of vowels and consonants.

Fundamental Frequency (f_0) or pitch is defined as the frequency at which the vocal cords vibrate during a voiced sound. *Pitch* is the frequency of a sound as perceived by human ear [2]. It is an important attribute of voiced speech. It contains speaker-specific information. Thus estimation of pitch is one of the important issues in speech processing ([1], [10]).

Timbre is the attribute that distinguishes sounds of equal pitch, loudness and duration such as two different musical instruments playing exactly the same note. It contributes to our perception and discrimination of different vowels and consonants in speech, instruments in music and environmental sounds.

The MFCC vectors are commonly used in audio analysis and are described as timbral features because they model the short-time spectral characteristics of the signal onto a psychoacoustic frequency scale.

[1] *Dept. of Medical Electronics, V.P.M.'s Polytechnic, Thane, Maharashtra, India*

The purpose of this work is to propose and evaluate a model of vowel perception which assumes that vowel identity is recognized by a template-matching process. It involves the comparison of Mel Frequency Ceptral Coefficient (MFCC) of word input with MFCCs of
individual phonemes for vowel that are learned through ordinary exposure to speech. For feature matching Adaptive Network based Fuzzy Inference System (ANFIS) is used. ANFIS system uses two, neural network and fuzzy logic approaches. When these two systems are combined, system is capable of reasoning and learning in an uncertain and imprecise environment. It may qualitatively and quantitatively achieve an appropriate result that will include either fuzzy intellect or calculative abilities of neural network.

Vowel recognition has specific applications in many areas. In the medical field one of the applications of vowel recognition is in cochlear implant (CI). CI is a prosthetic device that partially restores hearing function in individual with severe-to-profound sensorineural hearing loss. Because there are many parameters in the CI device that can be optimized for individual patients, it is important to estimate a parameter's effect before patient's evaluation. In CI testing to predict the recognition of speech processed through an acoustic model, vowel perception is investigated in severe noisy speech signal ([4], [7], [12]).

II. RELATED WORK

The most practical speech recognition systems rely on vowel recognition to achieve high performance. The vowels are longer in duration than consonants and are spectrally well defined. Vowels are easily and reliably recognized and therefore have a significant contribution in recognizing speech. Various methods and algorithms are adopted and proposed for speech recognition, vowel recognition, timbre and speech perception in different areas of applications. A speech recognition algorithm for English digits using MFCC vectors to provide an estimate of the vocal tract filter. MFCC parameters are a good perceptual representation for static sounds [13]. Spectral envelope parameters in the form of mel-frequency cepstral coefficients are often used for capturing timbral information of music signals in connection with genre classification applications.

Paul Iverson, Charlotte A. Smith and Bronwen G. Evans demonstrated that both cochlear implant users and normal-hearing individuals use formant movement and duration cues when recognizing English vowels (*Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration-J. Acoust. Soc. Am., Vol. 120, No. 6, December 2006*).

Ying-Yee Kong, Ala Mullangi and Jeremy Marozeau stated that there was a close relationship between timbre perception and vowel recognition with regard to combined (bimodal or bilateral) benefit. That is, individuals who demonstrated a significant combined benefit for vowel recognition in their paper *Timbre and Speech Perception in Bimodal and Bilateral Cochlear-Implant Listeners*

Harald Frostel, Andreas Arzt, Gerhard Widmer in their paper proposed and presented *The Vowel worm: Real-Time Mapping and Visualisation of Sung Vowels in Music*, an approach to predicting vowel quality in vocal music performances, based on common acoustic features (mainly MFCCs). Rather than performing classification, they used linear regression to project spoken or sung vowels into a continuous articulatory space: the IPA Vowel Chart. They introduced a real-time on-line visualization tool, the Vowel Worm, which builds upon the resulting models and displays the evolution of sung vowels over time in an intuitive manner. The concept presented in their work can be used for artistic purposes and music teaching.

Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, in their paper "*Speech Recognition using MFCC," International Conference on Computer Graphics, Simulation and Modeling* (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand), addressed the principle of speech MFCC extraction for performing word recognition when training words with support vector machine(SVM).

III. CEPSTRAL ANALYSIS FOR VOWEL RECOGNITION

According to the source filter theory of speech production, voiced sounds are produced by exciting the time varying system with periodic impulse sequence and unvoiced sounds are produced by exciting the time varying system with a random noise sequence. The resulting speech can be considered as the convolution of respective excitation sequence and vocal tract filter characteristics. If e(n) is the excitation

sequence and h(n) is the vocal tract filter sequence, then the speech sequence s (n) can be expressed as follows: ([2],[11])

$$s(n) = e(n) * h(n) \quad (1)$$

This can be represented in frequency domain as,

$$S(w) = E(w).H(w) \quad (2)$$

Cepstral analysis is used for transforming the multiplied source and system components in the frequency domain to linear combination of the two components in the cepstral domain. From the Eqn. (2) the magnitude spectrum of given speech sequence can be represented as,

$$|S(w)| = |E(w)|.|H(w)| \quad (3)$$

To linearly combine the E(ω) and H(ω) in the frequency domain, logarithmic representation is used. So the logarithmic representation of equation (3) will be,

$$log|S(w)| = log|E(w)| + log|H(w)| \quad (4)$$

Hence in log domain the excitation and the vocal tract shape are superimposed, and can be separated. Cepstrum is computed by taking Inverse Discrete Fourier Transform (IDFT) of logarithm of magnitude of discrete Fourier transform finite length input signal as shown in Fig.1

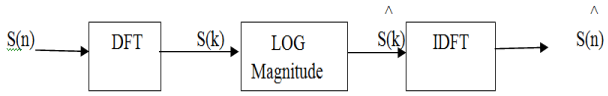

Figure.1 Computation of Cepstrum

$\hat{S}(n)$ is defined as cepstrum.

The samples of S(n) in its first 3ms describe h(n) and can be separated from the excitation. The later is viewed as voiced if S(n) exhibits sharp periodic pulses. Then the interval between these pulses is considered as pitch period. If no such structure is visible in S(n), the speech is considered unvoiced.

IV. MFCC FOR VOWEL RECOGNITION

Mel-Frequency analysis of speech is based on human perception experiments. It is observed that human ear acts as filter. It concentrates on only certain frequency components. These filters are non-uniformly spaced on the frequency axis. More number of filters in the low frequency regions and less number of filters in high frequency regions [7]. MFCC can be used to estimate the acoustic vowel. In speech recognition, it is a widespread method for describing the vocal tract transfer function. MFCC are perceptually motivated features that provide a compact representation of the short-time spectrum envelope. Spectral envelope parameters in the form of MFCCs are also often used for capturing timbre information of music signals. Since timbre similarity and estimating the vocal tract transfer function are closely related, hence MFCCs have also proven successful in the field of music for timbre characterization. Functionally, timbre is a key determinant of sound identity, and plays a pivotal role in speech as it is the principal determinant of phonetic identity [3]. Timbre, often referred to as the color of sound, is believed to play a key role in this recognition process. Measures of spectral shape have thus

been proposed as basic dimensions of timbre (e.g., formant position for voiced sounds in speech, sharpness, and brightness)

Firstly, Database is generated by cropping the individual vowel phoneme sounds from recorded mono sound words. Voice is converted into digital signal form to produce digital data representing each level of signal at every discrete time step. The digitized speech samples are then processed using MFCC to produce voice features. MFCC of these phonemes is calculated and saved as template. After that, the coefficient of voice features can go through neuro fuzzy system to select the pattern that matches the database and input frame in order to recognize the resulting vowel in it. In this work an audio signal analysis approach based on MFCC, for vowel recognition is proposed. Fig. 2 shows the block diagram for 'Vowel recognition system'.

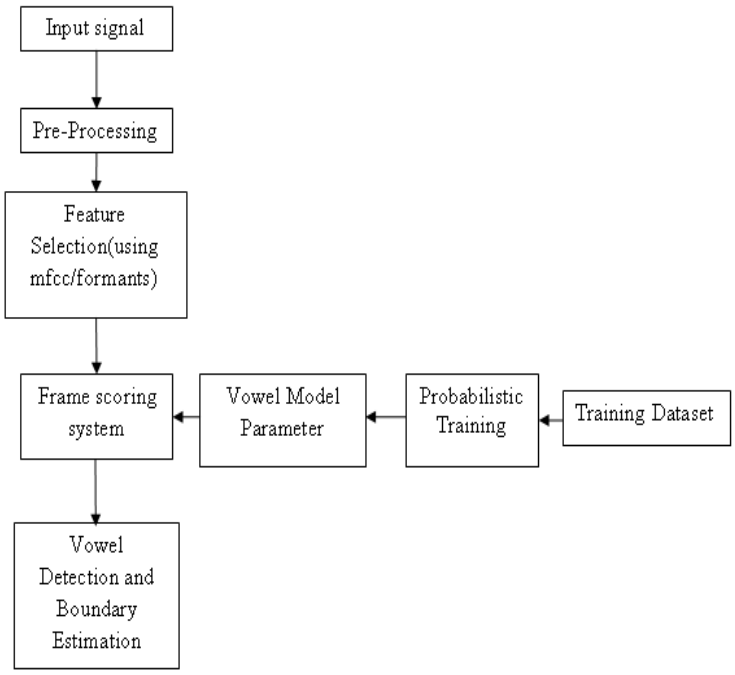

Figure.2 Vowel recognition system

V. MFCC CALCULATIONS

A block diagram for calculating MFCC is as shown in Fig.3

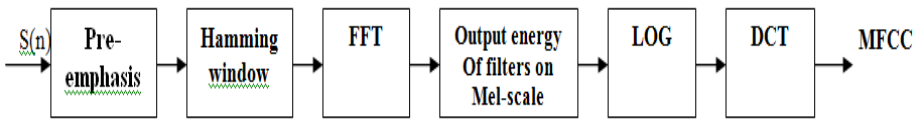

Figure.3 Computation of MFCC

Pre–emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y(n)=X(n)-0.97X(n-1) \quad (5)$$

Let's consider a = 0.97, which make 97% of any one sample is presumed to originate from previous sample.

Framing

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as: If the window is defined as W (n), $0 \leq n \leq N-1$ where

N = number of samples in each frame, Y[n] = Output signal, X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) - W(n)$$
$$W(n) = 0.54 - 0.46 \cos(2n/(n-1)) \quad 0 \leq n \leq N-1 \quad (6)$$

Fast Fourier Transform

To convert each frame of N samples from time domain into frequency domain. 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:

$$Y(w) = FFT[h(t)*x(t)] = H(w)*X(w) \quad (7)$$

If X (w), H (w) and Y (w) are the Fourier Transform of x (t), h (t) and y (t) respectively.

Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale is used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$F(Mel)=[2595*\log10[1+f]700] \quad (8)$$

Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector. ([7], [8], [9])

VI. FEATURE MATCHING

In simulation, the Adaptive network based Fuzzy Inference System (ANFIS) is employed for testing and matching the templates with the given input signals. The Sugeno fuzzy model is used for generating fuzzy rules from a given input-output data set. Using a given input/output data set, the toolbox function *anfis* constructs a fuzzy inference system (FIS) whose membership function parameters are tuned (adjusted) using either a backpropagation algorithm alone or in combination with a least squares type of method. This adjustment allows fuzzy systems to learn from the data they are modeling.[14]
In this case MFCC of cropped vowel phoneme are used for training FIS membership function parameters to emulate a given training data set.
After training, Model validation is carried out. In this process the input vectors from input/output data sets on which the FIS was not trained, are presented to the trained FIS model, it helps to see how well the FIS model predicts the corresponding data set output values.
Testing data input is then presented to the model. It compares the output membership functions of testing data with the input membership function parameters that are tuned (adjusted) in training. Based on matching of these functions the vowel phoneme in specific word input is recognized.

VII. RESULTS

Following results in the form of plots and wave forms as shown in figs 4,5,6,7,8 and 9 for the database that is generated by cropping the individual vowel phoneme sounds from recorded mono sound words. MFCC of these phonemes is calculated and saved as template which is further matched with the given input speech signal to recognize vowels in it. Figure 4, 5 shows the MFCC plots for utterance of vowel phonemes as stem graph.

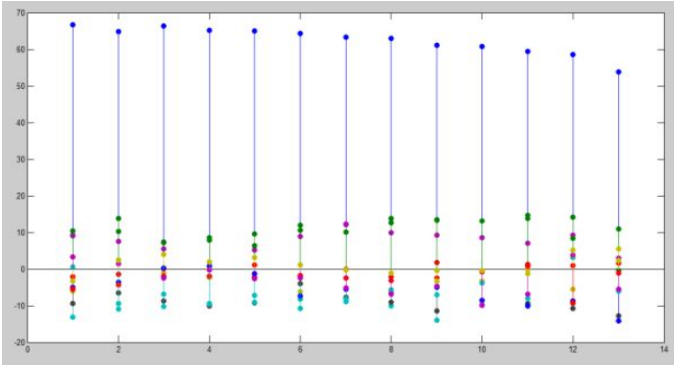

Figure. 4 MFCC plot for short /a/

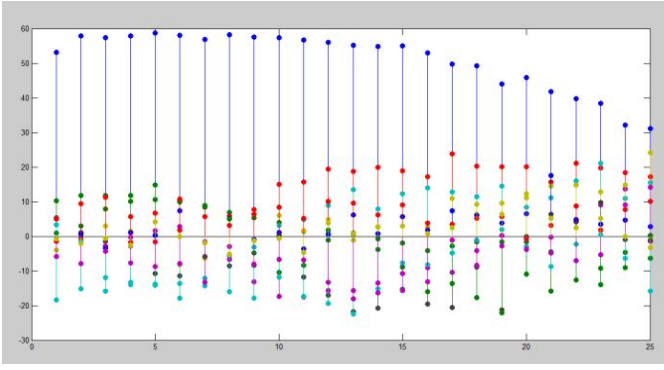

Figure. 5 MFCC plots for /oi/

Figure 6, 7 show the testing wave forms for the individual phoneme for model validation process which are identified correctly.

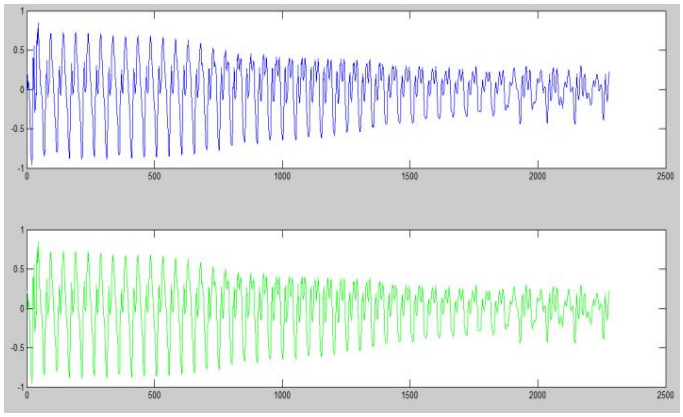

Figure.6 Wave forms of testing of phoneme /a/

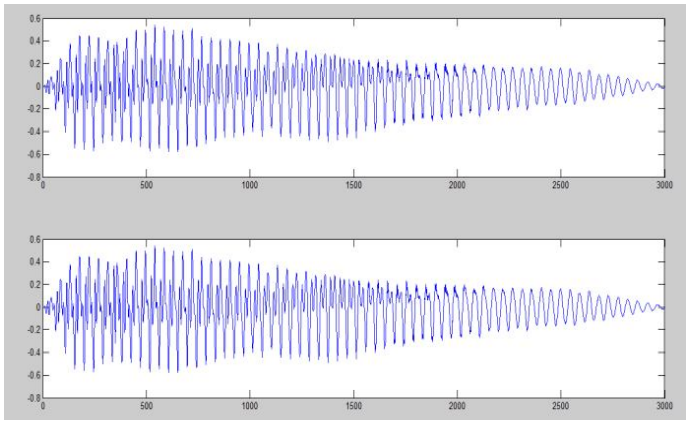

Figure. 7 Wave forms of testing of phoneme /oi/

Figure 8, 9 shows the results of testing of given input signal such as words containing vowel phonemes. In these waveforms the vowel phonemes are located by different colors as they are appearing in the utterance of particular word.

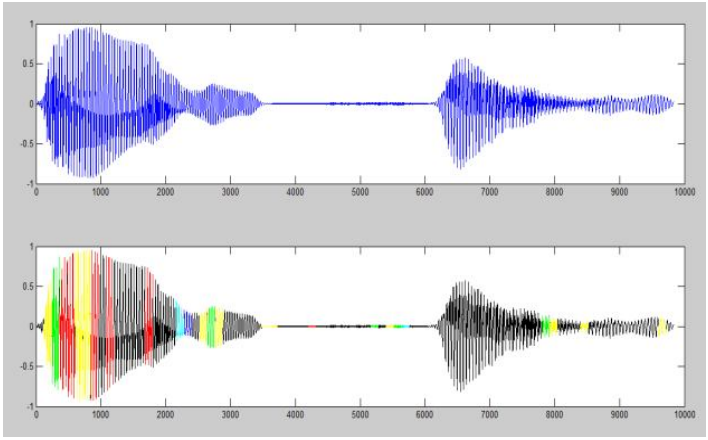

Figure. 8 Wave forms of testing of word ‘also’

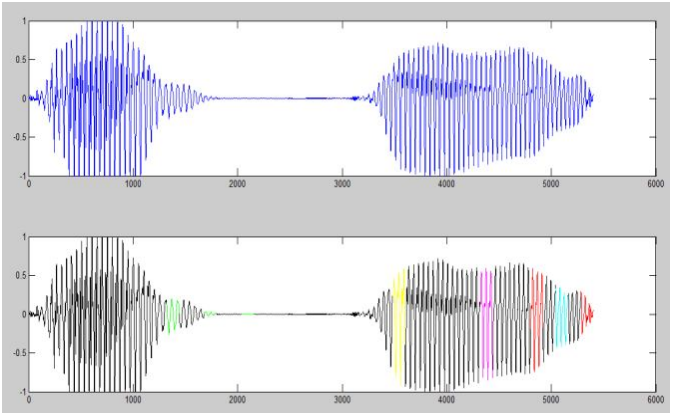

Figure 9 Wave forms of testing of word 'ago'

VIII. APPLICATION

To show the application of this work in CI testing and tuning, cochlear implant device sound is simulated. These simulated words can be presented as input for the vowel recognition system. Proper tuning of the CI device can be claimed if the correct vowel recognition is done by the proposed model. This will be helpful in tuning the CI devices before implantation in the patient's ear.

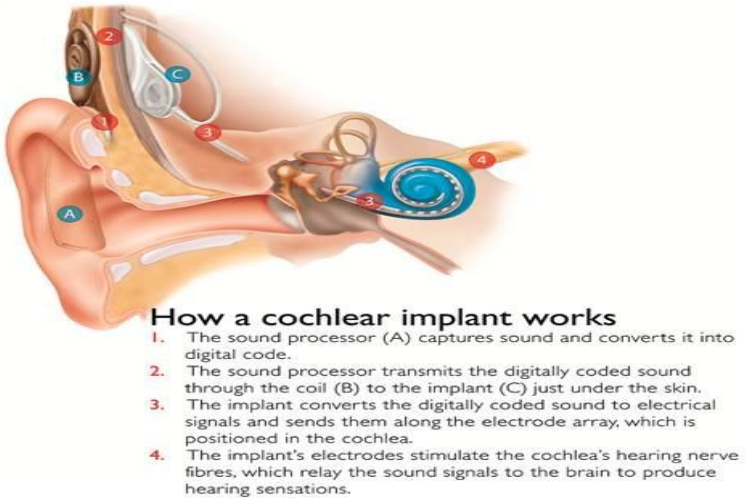

Figure 10 How a cochlear implant works?

IX. CONCLUSION

This paper has discussed vowel recognition algorithm. In future accuracy of the system can be improved by giving more training to the system. This system can be helpful for vowel perception which is to be investigated in severe noisy speech signal for CI testing and tuning. It can also be further explored for timbre analysis (identification) in music signal.

REFERENCES

[1] L.R. Rabiner & R.W. Schafer, "*Digital Processing of Speech Signals*", 10th ed. Pearson Publication 2013
[2] L.R. Rabiner & B. H. Juang, "*Fundamentals of Speech Recognition*", 4th ed. Pearson, 2011
[3] Stephen M. Town and Jennifer K. Bizley, "*Neural and behavioral investigations into timbre perception*" NCBI resources, Published online Nov 13, 2013

[4] 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)
[5] Chuping Liu and Qian-Jie Fu "Estimation of Vowel Recognition with Cochlear Implant Simulations," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, VOL.54, NO. 1, JANUARY 2007
[6] Vibha Tiwari- "MFCC and its applications in speaker recognition," *International Journal on Emerging Technologies* **1**(1): 19-2, 2010
[7] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi- "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," *Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617 https://sites.google.com/Site/Journalofcomputing/138*
[8] Anjali Bala**,** Abhijeet Kumar**,** Nidhika Birla**,** "Voice Command Recognition System Based on MFCC and DTW," *International Journal of Engineering Science and Technology Vol. 2 (12), 2010,7335-7342*
[9] S. R. Suralkar, Amol C.Wani, Prabhakar V. Mhadse,"Speech Recognized Automation System Using Speaker Identification through Wireless Communication," *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE. Volume 6, Issue 1 (May. - Jun. 2013),PP 11-18*
[10] Estimation Of Pitch From Speech Signals [Online] Available: http://iitg.vlab.co.in
[11] Cepstral Analysis of Speech [Online] Available: http://iitg.vlab.co.in
[12] Rikus Swanepoel, "Vowel perception in severe noise", Master Engg, thesis University of Pretoria, Dec 2010
[13] Jesper Hojvang Jensen, Mads Græsboll Christensen, Manohar N. Murthi, and Soren Holdt Jensen, "Evaluation Of MFCC Estimation Techniques for Music Similarity"
[14] anfis and the ANFIS Editor GUI [Online] Available: http://www.mathworks.in/help/fuzzy/anfis-and-the-anfis-editor-gui.html#FP43142
[15] How a cochlear implant works [Online] Available: http://www.nbcnews.com/id/26980383/ns/health-aging/t/older-ears-hear-again-cochlear- implants