International Journal of Electronic and Electrical Engineering.
ISSN 0974-2174, Volume 7, Number 3 (2014), pp. 269-280

http://www.irphouse.com

Intoxicated Speech Detection using MFCC Feature Extraction and Vector Quantization

Risha Mal[1], R.K. Sharma[2] and Naveen Kumar[3]

[1]M.Tech. ECE [VLSI], Department of Electronics & Communication Engineering, National Institute of Technology, Kurukshetra, Haryana, India.
[2]Department of Electronics & Communication Engineering, National Institute of Technology, Kurukshetra, Haryana, India.
[3]M.Tech. CSE, The Institution of Electronics & Telecommunication Engineers, New Delhi, India.

Abstract

This study has been done on a technique which is suitable for tapping the telephonic conversation from a remote location to identify intoxication and consequent impaired brain activity that may cause criminal events e.g. DUI (driving under influence). This technique is time efficient, easy to use, non–invasive for the peoples and affordable for law enforcement personnel, bartenders/servers, court of law, co-workers/supervisors, clinicians, teachers and individuals who need to identify the presence and level of intoxication state in other peoples. The peaks in log Mel Filter Bank are main cues for identifying the sounds of speech. If a person is found drunk and his/her voice shows a great deal of variation, then this study describes an effective unsupervised method for query-by-audio sample speaker retrieval firstly by extracting MFCC features and then VQ (vector quantization) algorithms on the alcoholic audios. This method is also supported by verifying some speech parameters (fundamental frequency, jitter, shimmer). A set of twelve mel-frequency cepstrum coefficients computed every 10ms and which resulted the best performance i.e. 95% recognition with each of 8 speakers. The superior performance of the mel-frequency cepstrum coefficients may be an attributed to the fact that they better represent the perceptually relevant aspects of the short-terms speech spectrum.

Key Words: Alcoholic or Intoxicated speech detection; Mel-frequency cepstral coefficients (MFCC), vector quantization; Euclidean distance.

1. Introduction

The speech pattern of an individual changes with the consumption of alcohol, due to which slurred speech is produced. Speaking in any language exhibit disordered speaking patterns when they are under the influence of alcohol. The presence of alcohol in the speaker represents host of challenges to them in all aspects of language production. Any or all of the lexical, syntactic, morphological, and phonological processes may become degraded. The identification of inebriation is a problem faced not only by clinicians and teachers, but also by any individual who needs to identify the presence and level of this state in others. If a person is lightly intoxicated his/her voice cannot be identified being under influence, hence, it is a challenge to identify people's speech even in any amount alcohol ingestion. Most of the people depend on breath analyzers or blood alcohol detection which is a passive way of detection. Researchers have shown that the most important symptoms of alcohol ingestion are: (a) speech degradation, (b) the severity and/or type of impairment varies from person to person, (c) various types of drinking behaviors - and/or gender - have an influence on the process, (d) severity of speech impairment correlate with increasing intoxication, (e) speech patterns different for increasing and decreasing involvement, (f) the presence of inebriation, and level of severity, always be detected in speech. [1],[2]. Speech have many features like fundamental frequency (F0), jitter, shimmer, format frequency, SNR, harmonics etc. All these parameters were first tested and then analyzed for speech features that mostly affected the intoxicated person. [3].

Speech features extraction has done using Mel-frequency cepstral coefficients (MFCC). These speech features are unique numerical values which are representing the different values for characteristics of an individual person. These coefficients are very huge in number therefore specific algorithm is required for classification of these values. Vector quantization using LBG algorithm is used where the speech feature vectors are classified using k-means and identified by distance measurement by Euclidean distance.

2. Speech Feature Analysis

(i) Test Setup: This step includes an introduction of software 'Praat' which is used to analyse the speech sample. A voice recording in .wav format is read by software and saved with the frequency of 44100 Hz. The spectrogram of the speech signal is viewed. In this window there are options to view spectrum, pitch, intensity, formant, pulses etc. We will take into consideration of fundamental frequency of pitch from voice report of pulses.

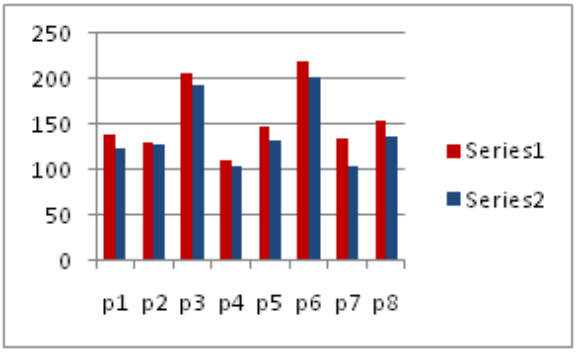
(ii) Analysis of speech sample - Fundamental Frequency: The previous studies indicate that the fundamental frequency increases in most of the cases when a person is lightly intoxicated [4], [5]. The challenge is to prove if it is right with the help of speech samples. The vibration that has the slowest rate is called the fundamental frequency. The fundamental frequency, often referred to simply as the fundamental and abbreviated F0, is defined as the lowest frequency of a periodic waveform. In terms of a superposition of sinusoids (e.g. Fourier series), the fundamental frequency is

the lowest sinusoidal frequency in the sum. It is referred that fundamental frequency of males generally varies from 85-200 Hz and for females it is 165-300 Hz. The analysis of F0_Hz of the voice samples is done where the values of F0_Hz is plotted between intoxicated speech against sober speech. Three speech samples of the same person are taken during intoxication and sober condition both. Now, the mean value of F0_Hz is calculated at different time instants of intoxication.

Table 1: Standard deviation and median value of F0_Hz

Persons	MeanF0_Hz (intox)	MeanF0_Hz (sober)	Std. Deviation
P1	138.1484	123.8374	10.11940515
P2	128.8405	127.2004	1.159725832
P3	206.9155	193.1159	9.757790738
P4	109.9314	103.0747	4.848419067
P5	146.565	130.514	11.34977094
P6	220.0789	201.119	13.40667386
P7	134.3082	102.4677	22.51463347
P8	153.6494	135.5852	12.77331832

Persons from P1-P5 are the voices saying “I am fine’’, from which the vowel ‘i’ is extracted for analysis and persons ‘P6-P8’ are the voices saying “kurukshetra” from where the vowel ‘u’ is extracted for analysis. Hence, both gives a comparative study between two vowels ‘a’ and ‘u’ of their F0_Hz of different persons. The study shows a increasing trend when lightly intoxicated about 110mg. F0_Hz of lightly intoxicated person is likely to rise may be due to hyperactivity of the motor system of a person. But it is not necessary the F0_Hz will always rise during any level of intoxication [4]. It is presumed that during heavy intoxication the F0_Hz is likely to decrease due to sleepiness and lack of consciousness. The graphical representation of F0_Hz in both the states is shown as below: -



Graph 1: Mean fundamental Frequency of Talker - Red bars- alcoholic, blue bars- non-alcoholic

3. Speech Information Processing

Feature extraction and Feature classification - Mel-Frequency Cepstral Coefficients: The block diagram of speaker recognition is shown in fig. 1. This represents the speaker recognition model, where MFCC features are extracted and classified using VQ and finally Euclidean Distance measurements is done.

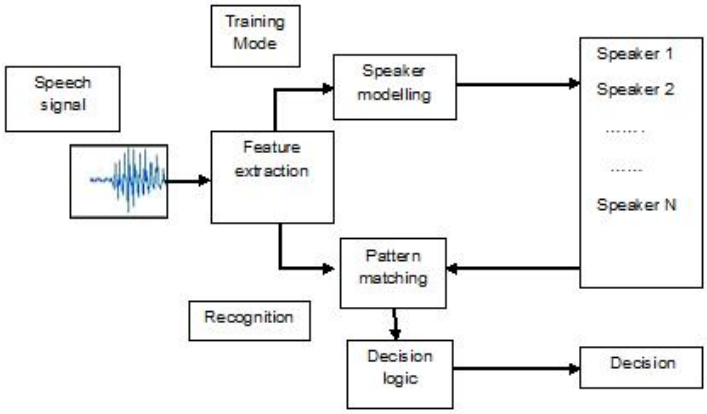


Fig 1: Speaker Recognition Model.

The frequency content of sounds is defined in nonlinear scale called the Mel scale. MFCC is the most robust and accurate algorithm that uses acoustic features for speech/speaker recognition [8]. MFCC is the acoustic approach that takes human perception sensitivity with respect to frequencies into consideration, and therefore is best for speaker recognition. The performance of speaker recognition improved significantly when complementary information is Ex-Ored with MFCC in feature vectors either by concatenation or by combining models scores. The main complementary information is like residual phase [7]. The audio file converts the speech waveform to some type of parametric representation for analysis and processing. This is referred to as the signal-processing front end. The speech signal is a slow time varying signal (quasi-stationary). The speech signal is analyzed over a sufficiently short period of time (between 5 and 100 msec), its characteristics are almost stationary. Therefore, short time spectral analysis is the most common way to characterize the speech signal. Such parametric representation best uses MFCC for feature extraction. The speech input is typically recorded at a sampling rate above 44100 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which covers most of the energy sounds that are generated by humans. As discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to variations. Figure 2 shows the block diagram of the MFCC processor. The process is divided into 5 portions, namely Frame Blocking, Windowing, FFT, Mel-Frequency Wrapping and Cepstrum.

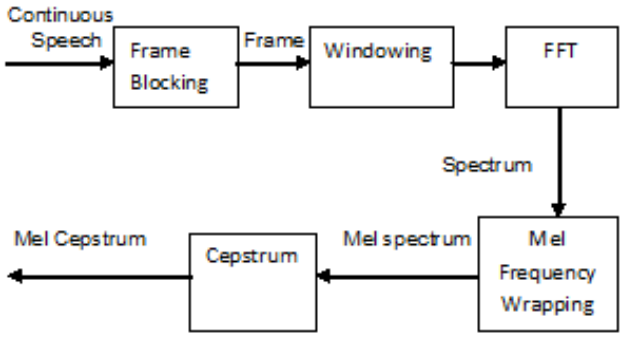


Fig. 2: Block diagram of the MFCC Processor.

A. Frame Blocking: The continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The start index and stop index is specified for what section of speech sample to be taken. The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 512 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and M = 100.

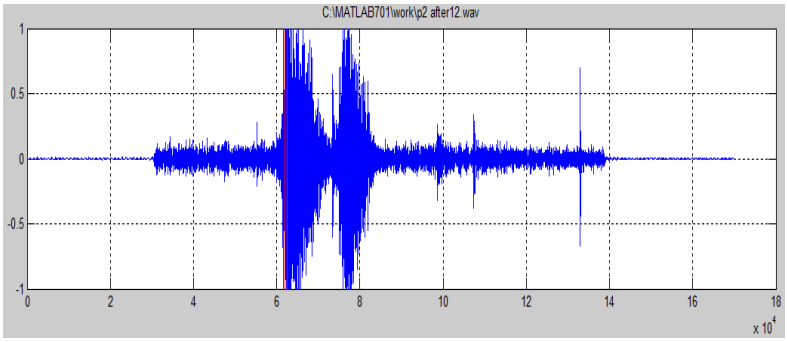


Fig 3: Frame blocking of continuous speech sample "I am fine" using index no.+frame size-1(62000Hz+512 byte-1)

B. Windowing: The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as w (n), $0 \leq n \leq N-1$, where N is the number of samples in each frame, then the result of windowing is the signal as in Equation 1:

$$y(n) = x(n)w(n), \quad 0 \leq n \leq N-1 \qquad \ldots\ldots \qquad \ldots (1)$$

The Hamming window is used for the windowing process, which has the form as in Equation 2:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1 \ \ldots\ldots \quad (2)$$

C. Fast-Fourier Transform (FFT): The next processing step is the Fast Fourier Transform, which 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 {xn}:

$$Xn = \sum_{k=0}^{N-1} xnke^{-2\pi jkn/N} \quad (3)$$

j here is to denote the imaginary unit, i.e. j = $\sqrt{-1}$. In general Xn's are complex numbers. The result after this step is often referred to as spectrum or periodogram.

D. Mel-Frequency Wrapping: Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency 'f' measured in Hz, is a subjective pitch measured on a scale called the 'mel' scale. The mel frequency scale is a linear frequency spacing below 1000 Hz and 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. Therefore we can use the formula in Equation 4 to compute the mels for a given frequency 'f' in Hz.

$$\text{mel}(f) = 2595 * \log10(1 + f / 700) \quad \ldots \quad \ldots\ldots\ldots (4)$$

One approach to simulate the subjective spectrum is to use a filter bank, uniformly spaced on the mel scale. The filter bank has a triangular bandpass frequency response. The spacing as well as the bandwidth is determined by a constant mel frequency intervals. The modified spectrum of S (w) thus consists of the output power of these filters when s (w) is the input. The number of mel spectrum coefficients K is typically chosen as 12. This filter bank is applied in the frequency domain, therefore it is simply amounts to take only those are triangle-shape windows on the spectrum.

E. Cepstrum: In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). 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, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore, if we denotes those mel power spectrum coefficients that are the result of the last step are $\hat{S}k$, k = 1, 2, ..., K, we can calculate the MFCC's, as in Equation 5:

$$\hat{C}n = \sum_{k-1}^{K} (\log \hat{S}k) \cos\left[n(k - 0.5)\frac{\pi}{K}\right],\ n = 1,2,\ldots,K \quad (5)$$

We exclude the first component, $\hat{C}0$ from the DCT since it represents the mean value of the input signal which carried little speaker specific information.

F. First Modification – Feature Matching: The goal of pattern recognition is to classify objects of interest into one or more of a number of categories or classes. The objects of interest are generally called patterns and in our case are sequences of acoustic vectors that are extracted from an input speech using MFCC techniques. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can also be referred to as feature matching. Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. This is exactly our case since during the training session, we label each input speech with the ID of the speaker. These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm. These patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm. In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is 'vector-quantized' using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

G. Second Modification - Clustering the Training Vectors: The acoustic vectors extracted from input speech of a speaker provide a set of training vectors. Next important step is to build a speaker-specific VQ codebook for these speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980] [9], for clustering a set of L training vectors into a set of M codebook vectors. The algorithm is formally implemented by the following recursive procedure:-

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook y (n) according to the rule.
3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of M is designed.

The LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialise the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained. Figure 4 shows, in a flow diagram, the detailed steps of the LBG algorithm. 'Cluster Vectors' is the nearest-neighbour search procedure which assigns each training vector to a cluster associated with the closest codeword. 'Find centroids' is the centroid update procedure. 'Compute D (distortion)' sums the distances (Euclidean distance) of all training vectors in the nearest neighbour search so as to determine whether the procedure has converged.

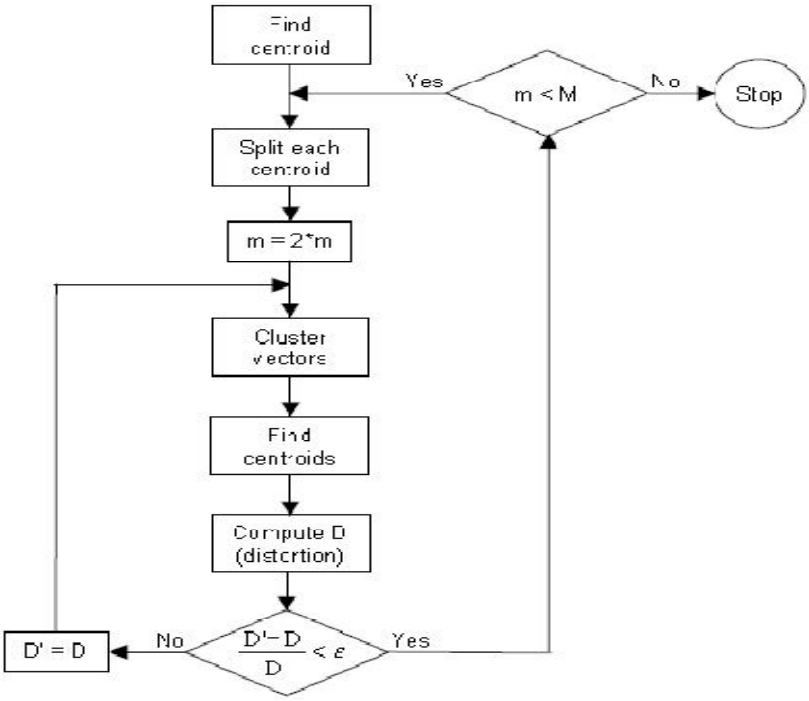


Fig. 4: Flow diagram of the LBG algorithm (Adapted from Rabiner and Juang, 1993)

4. Results

The speech samples are analyzed in various ways. After feature extraction using MFCC the vectors were classified by Vector Quantization and tested using Euclidean Distance for identification. Analyzing of Euclidean Distance matrix, rank and error rate could be found. It is necessary that if a person is to be identified the distance between the training sample of that particular speech sample and the testing sample must be minimum as compared to other training samples.

Speech samples of same speaker are taken and trained with sober speech samples (3 samples) and tested with corresponding intoxicated speech samples (3 samples) taken in about 2 min. duration. As the number of coefficients of MFCC are increases the identification rate and accuracy also increases. The accuracy attains saturation after 4 Cepstral coefficients due to less number of speakers.

Table 2: Number of Cepstral Coefficients vs. accuracy

No of Cepstral co-eff	Accuracy
4	50%
8	62.5%
12	62.5%
20	62.5%
28	62.5%

Speech sample of same speaker is taken and trained with sober speech sample (3 samples) and tested with corresponding intoxicated speech samples (3 samples) taken in about 2 min. duration. Same sampling frequency and same number of coefficients is considered. There are 8 speakers: 5 with same word and 3 with another word. The distance was calculated between same speakers with sober speech samples as training vectors and intoxicated speech samples as testing vectors.

Table 3: Same persons' training and testing vectors, Success Rate

Person	Identification accuracy
P1	95%
P2	100%
P3	98%
P4	95%
P5	95%
P6	97.2%
P7	93%
P8	92.3%

Table 4: Different persons' training and testing vectors, identification accuracy and Success Rate./

Person	Identification Accuracy	Success Rate
P1	100%	92.5%
P2	100%	93%
P3	100%	97%
P4	100%	89%
P5	100%	97%

P6	100%	96.173%
P7	100%	92.26%
P8	100%	97.3%

c) Made a database of training of all the 8 speakers, 3 samples of sober voices and training with 3 samples of intoxicated voice samples and found whether the system indentifies the right person and with how much error rate the distance deviates from same person's sober sample?

5. Conclusion

The distance between the same persons's speech samples taken within 2 min. duration gave the highest identification accuracy (average distance = 0.357). The identification accuracy between same person's sober speech and intoxicated speech sample gave 100% identification but higher error rate (average distance = 1.7). The identification of speech samples of 8 speakers (3 speech samples) of sober speech gave 95% identification against 1 person's intoxicated speech accuracy with high error rate (average distance >3). Furthermore the result shows that if Euclidean Distance of same person's intoxicated speech against sober speech is approximately greater than 3.

References

[1] Hollien H, DeJong G, Martin CA Schwartz, R. and Liljegren, K. 2001. Effects of ethanol intoxication on speech suprasegmentals [Dec]. Journal of the Acoustical Society of America, 110.3198- 206.

[2] Harry Hollien, James D. Harnsberger, Camilo A. Martin, Rebecca Hill, and G. Allan Alderman, Gainesville. Perceiving the Effects of Ethanol Intoxication on Voice, Florida

[3] Hollien H, DeJong G, Martin CA. Production of intoxication states by actors: perception by lay listeners. J Forensic Sci 1998;43(6):1153–1162

[4] K. Johnson, D.B. Pisoni, R.H. Bernacki. 1990. Do voice Recordings Reveal whether a Person is Intoxicated? A Case Study. Phonetica, vol. 41, pp. 215-237

[5] Zheng F., Zhang, G. and Song, Z. 2001 Comparison of different implementations of MFCC, J. Computer Science & Technology 16(6), pp. 582-589.

[6] Ganchev, T., Fakotakis, N., and Kokkinakis, G. 2005 Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proc. of SPECOM Patras, Greece, pp. 1191-194.

[7] Davis, S. and Merlmestein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on ASSP, Aug, 1980. pp. 357-366.

[8] Clarence Goh Kok Leon,Politeknik Seberang Perai, Jalan Permatang Pauh. Robust Computer Voice Recognition Using Improved MFCC Algorithm, 13500 Seberang Perai, Penang, Malaysia.

[9] Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[10] Klingholz, R. Penning, E. Liebhardt. 1988. Recognition of low-level—alcohol intoxication from speech signal. Journal of the Acoustical Society of America, vol. 84, pp. 929-935.

[11] L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, N.J., 1993

[12] Jyh-Shing Roger Jang MIR Lab, CS Dept, Tsing Hua Univ. Hsinchu, Taiwan http://www.cs.nthu.edu.tw/~jang.

[13] Vibha Tiwari, ―MFCC and its applications in speaker recognition‖, International Journal on Emerging Technologies 1(1): 19-22(2010), ISSN : 0975-8364.

[14] VOICEBOX: Speech Processing Toolbox for MATLAB. (n.d.), Retrieved April 17, 2012 from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[15] D. Liu and F. Kubala. Online speaker clustering, in Proc. 2004 IEEE International Conference Acoustics, Speech, and Signal Processing, vol. 1, pp. ;333-336, 2006.

[16] K.R. Aida–Zade, C. Ardil and S.S. Rustamov,‖ Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems‖,World Science and engineering and Technology 19 2006.

[17] R.K. Sharma and Tripti Kapoor. Parkinson's disease Diagnosis using Mel-frequency Cepstral Coefficients and Vector Quantization, International Journal of Computer Applications (0975 – 8887) Volume 14– No.3, January 2011.

[18] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", 3rd Proceedings of International Conference on Electrical & Computer Engineering , ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh,pp 565-568.

[19] Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka. Speaker Identification and Verification by Combining MFCC and Phase Information, IEEE Transactions on audio, speech, and language processing, VOL. 20, NO. 4, May 2012.