Speaker Independent Speech Recognition using MFCC with Cubic-Log Compression and VQ Analysis

Neeraj Kaberpanthi
Department of Electronics and Communication Engg.
Samrat Ashok Technological Institute
Vidisha, M.P., India

Ashutosh Datar
Department of Bio-Medical Engg.
Samrat Ashok Technological Institute
Vidisha, M.P., India

ABSTRACT
Speech processing is developed as one of the paramount requisition region of digital signal processing. Different fields for research in speech processing are speech recognition, speaker identification, speech bland, speech coding etc. The objective of Speaker Independent Speech Recognition is to concentrate, describe and distinguish information about speech signal and methodology towards creating the speaker free speech recognition system. Extracted information will be valuable for the directing and working different electronic contraptions and hardware through the human voice proficiently. Feature extraction is the first venture for speech recognition. Numerous algorithms are recommended / created by the scientists for feature extraction. In this work, the cubic-log compression in Mel-Frequency Cepstrum Coefficient (MFCC) feature extraction system is utilized to concentrate the characteristics from speech sign for outlining a speaker independent speaker recognition system. Extracted features are used to train and test this system with the help of Vector Quantization approach.

General Terms
Speech Recognition, Mel Frequency Cepstrum Coefficient, Vector Quantization, Cubic-Log Compression.

Keywords
Speech Recognition, Speaker Independent Speech Recognition, MFCC, Mel Frequency Cepstrum Coefficient, Vector Quantization, VQ Approach, Cubic-Log Compression.

1. INTRODUCTION
Speech Recognition or Automatic Speech Recognition is the process by which a machine can be commanded or controlled by human voice command. In Speech Recognition System (SRS) a person has to speak in a microphone. The speech signal is digitized by an analog-to-digital converter and is put in to the memory. To focus this, the computer attempts to match the input with a digitized voice sample, or template that has a known meaning to the computer [1].

Throughout the speech recognition prepare a stream of voice samples are entered as data to the system. When that, the speaker database must be prepared and tantamount feature must be processed from the input voice signal. The feature vectors are then matched to the substance of the database. The system holds information formats, and endeavors to match these samples with the real inputs voice signal provide to the system [2]. The best discourse recognizers were giving a 4-6 % of lapse rate because of nature.

Speech recognition system can be classified on the basis of three parameters, on the bases of speech patterns, on the bases of speaker recognition capability and on the bases of vocabulary size of SRS database [3]-[5]. Per speech patterns SRS can be classified as isolated word SRS and continues word SRS. Isolated word SRS works for isolated words not for sentences but continues word SRS works for sentences. As per speaker recognition capability SRS can be classified as speaker dependent, speaker independent and speaker adaptive SRS. Speaker dependent SRS can be operated by a particular speaker only, but speaker independent SRS can be operated by any user or these types of SRS are not bounded for particular user. Speaker adaptive SRS have the adaptive capability. They can update their parameter performance as per user voice characteristics. As per vocabulary size of SRS database it can further classified as small, medium and large vocabulary SRS. Small vocabularies have only 2 to 200 words. Medium vocabulary can have 200 to 2000 words. Large vocabulary can have 2000 to 200K words in their database.

Speech recognition can get affected by some variability within the speaker, this are speaking style and voice quality. Pitch, intensity and speaking rate variation affects the speaking style of a speaker, which affects speech recognition efficiency [6]. Accents, dialects and nation verses non-nation are the variables which create differences between the speakers, and affect the speech recognition efficiency. Background noise and microphone noise affects the speaking environment and affects speech recognition efficiency.

There are several methods has been proposed for feature extraction from speech signals like Perceptually Based Linear Predictive Analysis (PLP) [7], Linear Discriminant Analysis (LDA) [8], Linear Predictive Coding (LPC) Analysis [9] and Mel-Frequency Cepstrum Coefficients (MFCC) [10]-[18].

In speech processing, the mel-frequency cepstrum (MFC) is a illustration of the short-term power spectrum of a speech signal, based on a linear cosine transform of a log power

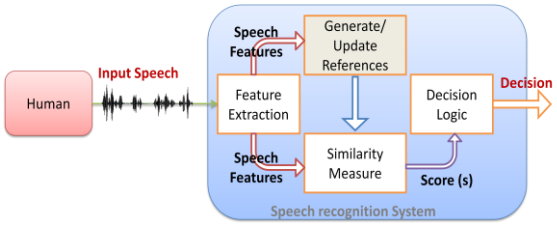

Fig 1: Speech Recognition System Model

spectrum on a nonlinear Mel-scale of frequency [10] – [12]. Mel-frequency cepstram coefficients (MFCCs) are coefficients that collectively make up an MFC. Mel-frequency

scale is a perceptually motivated scale [13] which is linear below 1kHz, and logarithm above, with equal numbers of samples below and above 1kHz. It represents the pitch (perceived frequency) of a tone as a function of its acoustics frequency. One mel is defined as one thousandth of the pitch of a 1kHz tone. Mel-scale frequency can be approximate by Eq. (1):

$$m=2595\ \log_{10}\left(1+\frac{f}{700}\right) \quad (1)$$

The Cepstrum is closely related to the auto correlation function. The Cepstrum separates the glottal frequency from the vocal tract resonances. The Cepstrum is obtained in two steps. A logarithmic power spectrum is calculated and declared to be the new analysis window. On that an inverse FFT is performed [19]. The result is a signal with a time axis. The word Cepstrum is a play on spectrum, and it denotes mathematically as in Eq. (2). Where s(n) is the sampled speech signal, and c(n) is the signal in the Cepstral domain.

$$c(n) = \text{ifft}\,(\log|\text{fft}(s(n))) \quad (2)$$

MFCC (Mel Frequency Cepstral Coefficients) is the most common technique for feature extraction. MFCC tries to mimic the way our ears work by analyzing the speech waves linearly at low frequencies and logarithmically at high frequencies. MFCC coefficients are a set of DCT decorrelated parameters, which are computed through a transformation of the logarithmically compressed filter-output energies, derived through a perceptually spaced triangular filter bank that processes the Discrete Fourier Transformed (DFT) speech signal. In cubic-log compression MFCC process instead of logtithmic energy in standard MFCC process, Cubic-log energy at each of the mel frequencies is taken [18].

The featured vectors of speech signal after processing by MFCC with cubic log compression is expressed as Eq. (3):

$$c[n] = \sum_{k=1}^{M} \log^3(s[k]) * \cos\left[n * (k+0.5) * \frac{\pi}{M}\right] \quad (3)$$

Where s[k] is energy in each mel window, $1 \le k \le M$; M is Mel windows number, which ranges from 20 to 24 generally, $1 \le n \le L$; L is desired order of M.

Extracted feature vectors through MFCC with cubic log compression are used to train SRS and recognize the correct speech signal with the help of vector quantization analysis.

Vector quantization (VQ) analysis is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. It was originally used for data compression [20]. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means, and some other clustering algorithms [21]. The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error [22] – [23]. This is why VQ is suitable for lossy data compression.

The paper is structured as follows: Section 2 describes the problems focused by the previous researches, Section 3 describes the proposed algorithm for the problem formulated in section 2, Section 4 shows the Matlab simulations and obtained results for the proposed algorithm and in Section 5 paper will conclude the research.

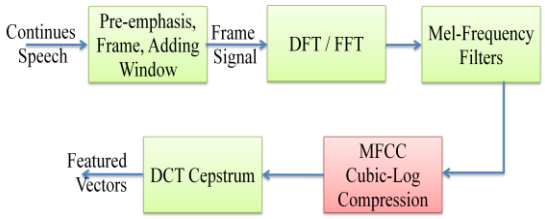

Fig 2: MFCC Process with Cubic-Log Compression

2. PROBLEM FORMULATION

By the literature survey related with speech recognition system, it is found that most of the researches are concerned with speaker identification systems and speaker dependent speech recognition systems. In both cases the efficiency of the SRS is concerned with training the system by the particular user's speech samples to be recognized by themselves. Because of the uncertain appearance of the noises with the input speech signal due to surrounding, throttle friction or some other causes this kind of SRS shows many time problem to recognize correct speech signal. This kind of SRS's can be operated by only users who had trained the system previously, even after training the success of the correct recognition is impacted due to the noisy environment. To overcome these problems speaker independent speech recognition system is a solution, which can break the limitations of the users after the one time proper training the SRS. This will also increase the success rate of speaker identification system and speaker dependent speech recognition systems.

3. PROPOSED METHODOLOGY

To develop the speaker independent speech recognition system we are proposing a method which uses MFCC with cubic log compression for the features extraction from the speech signal and these extracted features will be passed through the vector quantization analysis for the system training and recognition purpose.

To train the system a sufficient number of speech signals of a particular word by the different persons are used as speech sample data base. To extract there feature, MFCC with cubic log compression is used which produce speech signal feature vectors according to the number of mel-frequency filter bank used in MFCC process, which is represented by M.

It is necessary to normalize the extracted features vectors to make all the vectors from different persons in a same range for the vector quantization process. Each set of vectors for a particular person and word is separately normalized in the rage of 0 to 1. If P is the number of persons and n is the numbers of words to be trained the system, normalized vector set matrix for each person of dimension n x M is obtained.

Normalize vectors of vector set matrix of each person is fragmented according to the n and p. which will yield c[n x P] matrix of vectors which shows the local centroid of each words by different persons.

Matrix c[c x P] is now used for finding the global centroids for each words. Which will yield c[n x 1]. These set of centroids will be added with the c[c x p] matrix, by which it provides a centroids matrix of dimension c[n x p+1]. These centroids will be used for the recognizing the test speech signal.

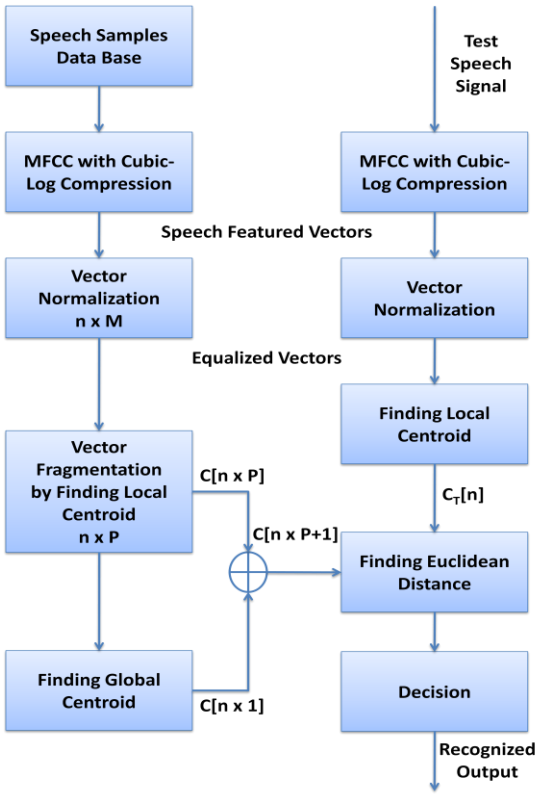

Fig 3: Proposed Methodology for Speaker Independent Speech Recognition System

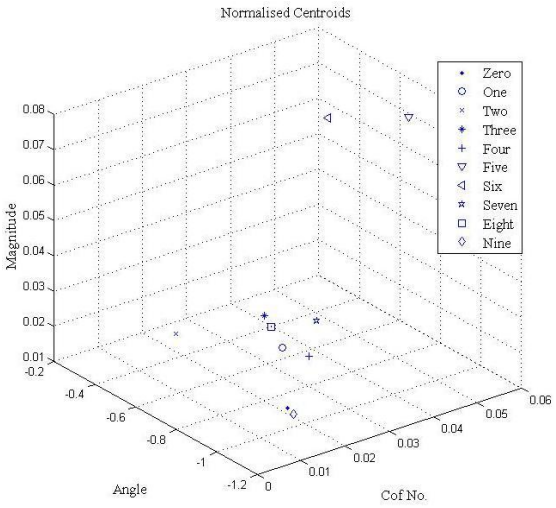

Fig 4: Global Centroids Plot for 'Zero' to 'Nine'

During the recognition, first the features of the test speech signal are extracted through the MFCC with cubic log compression of order M mel frequency bank. As similar to the training, module in testing also extract vectors will be normalized. Through the normalized vectors local centroid is found which is represented by CT[N].

To recognize the test speech signal, minimum Euclidean distance is calculated between c[n x P+1] and CT[n].

$$Ed[i,j] = \sqrt{(c[i,j] - C_T)^2} \qquad (4)$$

Where $1 \leq I \leq n$; $1 \leq j \leq P+1$ and Ed[i , j] is the euclidean distance matrix of order n x P+1.

According to the minimum Euclidean distance matrix decision can be taken for the recognized speech signal.

4. MATLAB SIMULATION

In this research work we have considered 20 persons speech samples of numerical words (Zero, One, Two, Three, Four, Five, Six, Seven, Eight and Nine) in less noise environment with sampling frequency 44.1 kHz, as the speech data base for the training the system. Speech signal sampling size is 10 milliseconds. Band stop Kaiser Window filter of 1-10 kHz is used for windowing the sampled speech signal. Mel Frequency filter bank size is 24. Another speech sample of same words is taken from the 20 persons for the testing purpose. Due to this M = 24, n = 10, P = 20.

Training speech samples are processed with the proposed method of speaker independent SRS, and constructed the clusters of centroids of each wards separately. We have used 3D vector quantization plot for the plotting of the centroids clusters. Calculated vectors and centroids are in complex form, due to this we used their Angle, Magnitude and coefficient number as the parameters for the 3D plot.

Figures 4 show the comparative positions of all global centroids of 'Zero' to 'Nine'. With this figure it is clear that all words centroids obtained an separate position with each other in an 3D space of censored, which is mandatory for the recognition of the correct speech signal.

We have tested this algorithm for numerical words samples 'Zero' to 'None', but showing the plots of sum of them. Figure 5 shows the plot of the clusters of speech samples of 'Zero' for 20 persons in blue color and their respective global centroids in red color which shows the identical location for each word with respect to each other. Similarly Figure 6 shows the plot of the clusters of speech samples of 'Two', Figure 7 shows the plot of the clusters of speech samples of 'Three', Figure 8 shows the plot of the clusters of speech samples of 'Five', Figure 9 shows the plot of the clusters of speech samples of 'Seven' and Figure 10 shows the plot of the clusters of speech samples of 'Nine'.

The experimental results are arranged in the confusion matrix form for the analysis between actual class and the predicted class by our SRS, as shown in Table 1. These results show the efficient prediction probability of the system trained and test for the 20 persons. Its probability can be easily improved with the help of training this system for the large number of samples.

Table 1. Recognition Result

		Prediction									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Actual Class	**0**	**6**	1	3	1	1	2	0	3	2	1
	1	3	**6**	1	2	2	2	0	2	1	1
	2	4	1	**6**	0	1	0	1	4	0	3
	3	4	2	1	**5**	3	2	0	0	1	2
	4	5	1	3	1	**6**	0	1	2	1	0
	5	2	2	1	1	1	**6**	5	1	0	1
	6	0	1	1	3	3	1	**6**	0	2	3
	7	0	3	2	0	1	3	3	**6**	1	1
	8	2	3	0	2	2	1	0	2	**5**	3
	9	0	1	0	2	0	3	3	3	3	**5**

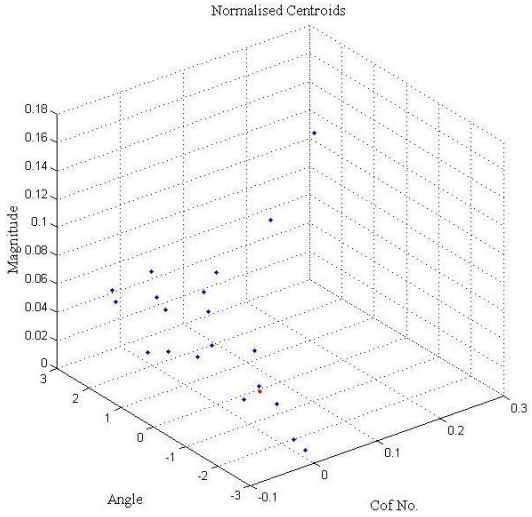

Fig 5: Local Centroids Plot For 'Zero'

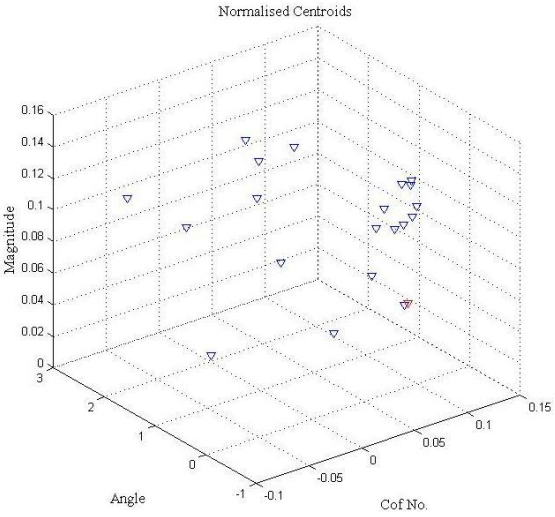

Fig 8: Local Centroids Plot For 'Five'

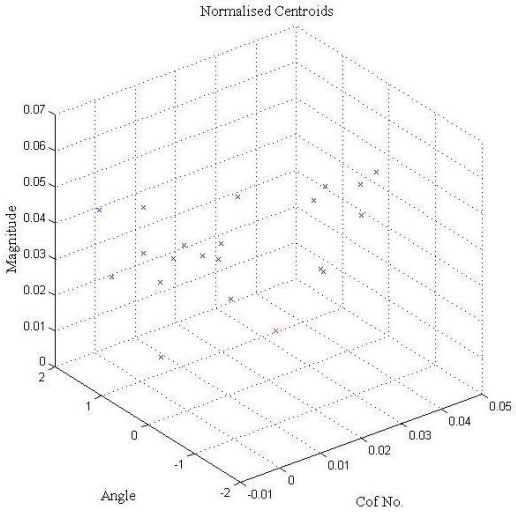

Fig 6: Local Centroids Plot For 'Two'

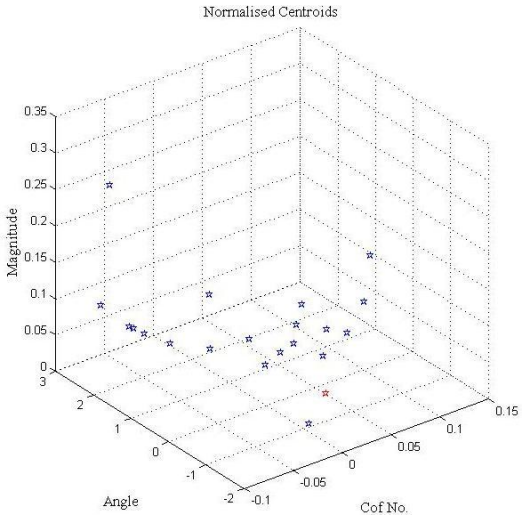

Fig 9: Local Centroids Plot For 'Seven'

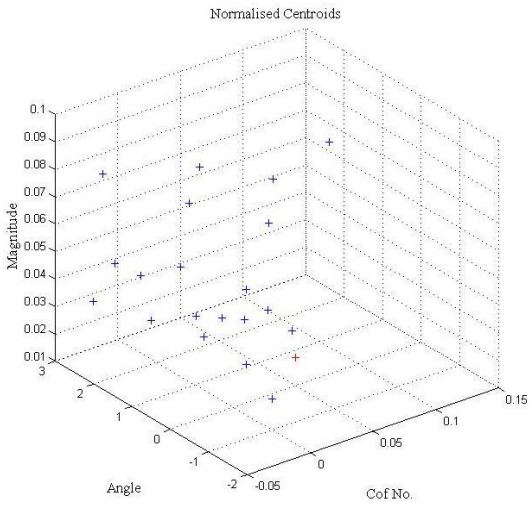

Fig 7: Local Centroids Plot For 'Three'

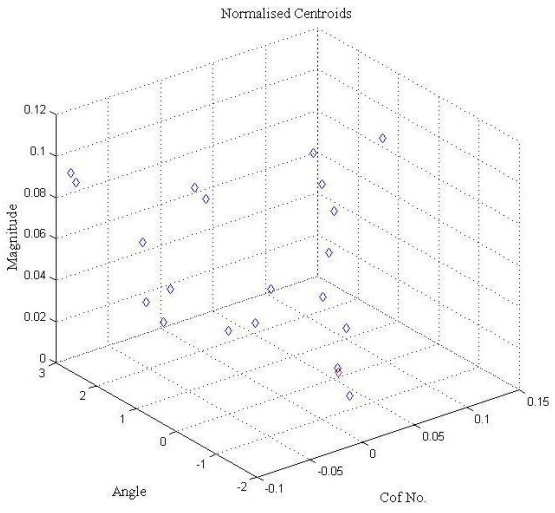

Fig 10: Local Centroids Plot For 'Nine'

5. CONCLUSIONS

The proposed speaker independent speech recognition system using MFCC cubic-log compression and vector quantization analysis is trained and test in less noise environment for the 20 persons. The result shows individual centroid positions in 3D vector space, which is important for the efficient recognition of the samples. This system is trained for 20 persons and achieved good probability of correct prediction. As this system uses MFCC with cubic-log compression for the feature extraction, this system has less number of vector coefficients, fast execution speed and due to use of VQ analysis it shows good accuracy.

6. ACKNOWLEDGMENTS

The author would like to express their sincere thanks to Dr. S. N. Sharma Head of the Department of Electronics and Communication, all authors of the reference for this research, all faculty members of the department and of all the persons who have supported in any way for this research.

7. REFERENCES

[1] Rabiner, L., and Juang, B. H. 2003. Fundamentals of Speech Recognition. Pearson Education (Singapore).

[2] Pathak, P. 2010. Speech Recognition Technology: Applications & Future. International journal on Advance Research on computer science. Vol. 1.

[3] Anusuya, M. A., and Katti, S. K. 2009. Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security. Vol. 6. No. 3. 181-205.

[4] Gaikwad, S. K., Gawali, B. W., and Yannawar, P. 2010. A Review on Speech Recognition Technique. International Journal of Computer Applications. Vol. 10. No. 3. 16-24.

[5] Prabhakar, O. P., and Sahu, N. K. 2013. A Survey On: Voice Command Recognition Technique. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3. 576-585.

[6] Karam, Zahi N., and Campbell W. M. A new Kernel for SVM MIIR based Speaker recognition. MIT Lincoln Laboratory, Lexington, MA, USA.

[7] Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. US WEST Advanced Technologies, Science and Technology, Englewood, Colorado. 1738-1752.

[8] Umbach, R. H., and Ney, H. 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. Acoustics, Speech, and Signal Processing, 1992. ICASSP-92. 1992 IEEE International Conference on, San Francisco, CA. Vol. 1. 13-16.

[9] Jiang, H., and Joo, M. 2003. Improved linear predictive coding method for speech recognition. Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on IEEE. Vol. 3. 15-18.

[10] Hossan, M. A., Memon, S., and Gregory, M. A. 2010. A Novel Approach for MFCC Feature Extraction. Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference, Gold Coast, QLD. 1-5.

[11] Junqin, W., and Junjun, Y. 2011. An Improved Arithmetic of MFCC in Speech Recognition System. Electronics, Communications and Control (ICECC), 2011 International Conference on, Zhejiang. 719 - 722.

[12] Ittichaichareon, C., Suksri, S. and Yingthawornsuk, T. 2012. Speech Recognition Using MFCC. International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Thailand. 135 - 138.

[13] Stevens, S. S., Volkman, J., and Newman, E. B. 1937. A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America 8 (3). 185 - 190.

[14] Zhu, W., and O'Shaughnessy, D. 2004. Incorporating Frequency Masking Filtering in a Standard MFCC Feature Extraction Algorithm. INRS-EMT, Quebec Univ., Montreal, Que., Canada. Vol. 1. 617 - 620.

[15] Firoz S. A., Vimal Krishnan, V. R., Sukumar, R., Jayakumar, A., and Anto, B. P. 2009. Speaker Independent Automatic Emotion Recognition from Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms. Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09. International Conference on, Kottayam, Kerala. 528 – 531.

[16] Homberg, M., and Gelbart, D. 2006. Automatic speech recognition with an adaptation model motivated by auditory processing. IEEE Transactions on Audio, Speech, and Language Processing. Vol. 14. 43 - 49.

[17] Wang, H., Xu, Y., and Li, M. 2011. Study on the MFCC Similarity-based Voice Activity Detection Algorithm. Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on, Deng Leng. 4391 - 4394.

[18] Devi, M. R., and Ravichandran, T. 2013. A Novel Approach for Speech Feature Extraction by Cubic-Log Compression in MFCC. Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on, Salem. 182 – 186.

[19] Childers, D. G., Skinner, D. P., and Kemerait, R. C. 1977. The Cepstrum: A Guide to Processing. The IEEE. Vol. 65. No. 10.

[20] Kekre, H. B., and Tanuja K. S. 2008. Speech Data Compression using Vector Quantization. World Academy of Science, Engineering and Technology. Vol. 2. No. 3. 568 – 571.

[21] Kekre H. B., and Sarode, T. K. 2013. New Clustering Algorithm for Vector Quantization using Rotation of Error Vector. International Journal of Computer Science and Information Security. Vol. 7. No. 3. 159 – 165.

[22] Singh, S., and Rajan, E. G. 2011. Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC. International Journal of Computer Applications. Vol. 17. No. 1. 1-7.

[23] Gupta, D., Mounima, R. C., Manjunath, N., and Manoj, P. B. 2012. Isolated Word Speech Recognition Using Vector Quantization (VQ). International Journal of Advanced Research in Computer Science and Software Engineering, Bangalore, India. Vol. 2. 164 - 168.