Vector Quantization and MFCC based Classification of Dysfluencies in Stuttered Speech

P. Mahesha and D.S. Vinod

Abstract--- *Stuttering also known as stammering is a speech disorder that involves disruptions or dysfluencies in speech. The observable signs of dysfluencies include repetitions of syllable or word, prolongations, interjections, silent pauses, broken words, incomplete phrases and revisions. The repetitions, prolongations and interjections are important parameter in assessing the stuttered speech. The objective of the paper is to classify the above mentioned three types of dysfluencies using Mel-Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) framework. For each dysfluency MFCC features are extracted and quantized to a number of centroids using the K-means algorithm. These centroids represent the codebook of dysfluencies. The dysfluencies are classified according to the minimum quantization distance between the centroids of each dysfluency and the MFCC features of testing sample.*

Keywords--- *Stuttering; Vector Quantization; Codebook; Dysfluencies; MFCC*

I. INTRODUCTION

Fluency is a speech pattern, which flows in a rhythmic and smooth manner. The stuttering also known as dysphemia and stammering is a disorder that affects the fluency of speech. Stuttering is one of the serious problems focused in speech pathology. It occurs in about 1% of the population and has found to affect four times as many males as females [1][2][3][4]. This disorder is characterized by disruptions in the production of speech sounds, called dysfluencies. Table I shows types of dysfluencies with examples [5].

Stuttering is the subject of interest to researchers from various domains like speech physiology, pathology, psychology, acoustics and signal analysis [6]. Therefore, this area is basically a multidisciplinary research field of science. There is moderate amount of work noticed on automatic stuttering recognition and classification by methods of acoustic analysis, feature extraction and statistical methods.

In conventional stuttering assessment process, the recorded speech is transcribed and dysfluencies like repetitions, prolongations and injections are identified. Then the frequency of occurrence of each dyfluency is counted. These assessment processes are based on the knowledge and previous experience of speech pathologist. The main drawbacks of making such assessment are time consuming, subjective, inconsistent and also poor agreement when different judges make counts on same material [7].

The objective of our work is to develop a method capable of finding the dysfluency in stuttered speech. This is one of the key parameter in objective assessment of stuttering. This helps Speech Language Pathologists (SLP) to assess stuttering patients, planning appropriate intervention program and monitoring the prognosis during the course of treatment. Also it improves interjudge agreements about stuttered events.

Table 1: Types of Dysfluency with Example

Type of dysfluencies	Example
Repetition	
Whole word	"What-what-what are you doing "
Part word	What t-t-t time is it?
Phrase	I want to-I want to I want to do it
Prolongation	
Sound/ syllable	"I am Booooobbbby James
Interjection (Filled pauses)	
Sound/syllable	"Um – uh -well, I had problem in morning",
Silent pauses	
Silent duration within speech considered normal	"I was going to the [pause] store
Broken words	
A silent pause with in words	" it was won[pause]derful"
Incomplete phrase	
Grammatically in complete utterance	I don't know how to ….let us go, guys"
Revisions	
Changed words, ideas	There was a dog, no rat named Arthur"

II. RELATED WORK

The dysfluent speech processing is one of the areas, where research is still very much in progress. Over recent years number of works have focused on the automatic detection and classification of dysfluencies in stuttered speech by means of acoustic analysis, parametric and non-parametric feature extraction and statistical methods. Which facilitate SLPs for objective assessment of stuttering. In [6], author used Artificial Neural Network (ANN) and rough set to detect stuttering events yielding accuracy of 73.25% for ANN and about 91% for rough set. The authors of [8][9] proposed Hidden Markov Model (HMM) based classification for automatic dysfluency detection using MFCC features and achieved 80% accuracy. In [7], automatic detection of syllable

P. Mahesha, Assistant Professor, Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India. maheshsjce@yahoomail.com

D.S. Vinod, Assistant Professor, Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, Karnataka, India. dsvinod@daad-alumni.de

DOI: 10.9756/BIJMMI.2019

repetition was presented for objective assessment of stuttering dysfluencies based on MFCC and perceptron features. An accuracy of 83% was achieved. Subsequently in [10], same author obtained 94.35% accuracy using MFCC features and SVM classifier. Authors of [11] achieved 90% accuracy with Linear Discriminant Analysis (LDA), k- Nearest Neighbor (k-NN) and MFCC features. In [12] the same author used similar classifiers, LDA and k-NN for the recognition of repetitions and prolongations with Linear Predictive Cepstral Coefficient (LPCC) as feature extraction method and obtained the best accuracy of 89.77%. In our previous work [13], we have developed a procedure for classification of dysfluency using MFCC feature and k-NN classifier for the recognition of three types of dysfluencies and obtained accuracy of 97.78% for k=5.

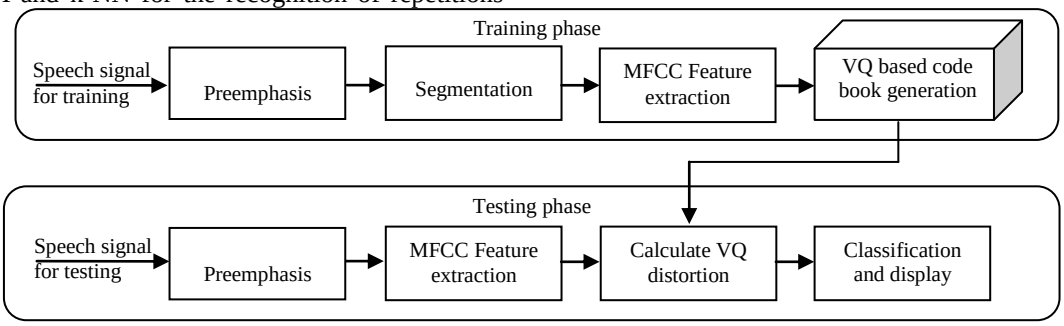


Figure 1: Schematic diagram of classification method

It is evident from literature survey that stuttering is characterized by different dysfluencies. Previous works are based on detecting repetition dysfluency and in few work, prolongation is also considered. However, the stuttered speech contains high occurrences of repetition, prolongation and interjection (filled pause) dysfluencies. Hence, in this paper we are proposing a method to classify repetition, prolongation and interjection dysfluencies.

III. Stuttered Speech Data

The speech samples are obtained from University College London Archive of Stuttered Speech (UCLASS) [14] [15]. The database consists of recording for monologs, readings and conversations. There are 40 different speakers contributing 107 recording in the database. In this work, speech samples are taken from 20 different speakers with age ranging from 10 years 4 months and 20 years 1 month. The samples were chosen to cover a wide range of age and stuttering rate. The dysfluencies such as repetition, prolongation and filled pause are extracted from these 20 different speakers.

IV. Experimentation

The complete process of classification system is depicted in Figure 1. The entire process is divided into training and testing phase. The major phases of classification system are preemphasis, segmentation, feature extraction, VQ codebook generation, VQ distortion calculation and classification.

A. Pre-emphasis

To enhance accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before feature extraction. 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. Generally pre-emphasis is performed by filtering the speech signal with the first order FIR filter, which has the following form:

$$H(z) = 1 - k * z^{-1} \qquad (0.9 < k < 1.0) \qquad (1)$$

In this case, the output of the preemphasis network, $\tilde{s}(n)$, is related to the input to the network, s(n), by the difference equation

$$\tilde{s}(n) = s(n) - k \cdot s(n-1) \qquad (2)$$

where k is the pre-emphasis factor, and the most common value is 0.97 [16]. Consequently the output is formed as follow: The aim of this stage is to boost the amount of energy in the high frequencies.

B. Segmentation

In this paper, dysfluencies such as repetitions, prolongations and interjections were identified by hearing the recorded speech samples and segmented manually. The segmented samples are subjected to feature extraction.

C. Feature Extraction

Feature extraction is to convert an observed speech signal to some type of parametric representation for further investigation and processing. Several feature extraction algorithms are used for this task such as Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) cepstra.

The MFCC feature extraction method is used in this paper. It is one of the best known and most commonly used features for speech recognition. The feature extraction package produces a multi dimensional feature vector for every frame of speech. In this study we have considered 12MFCCs. The human voice is very well adapted to the ear sensitivity and most of the energy comprised in the lower frequency energy spectrum below 4 kHz. Usually 12 coefficients are retained due to slow variation of spectrum of the uttered words [17].

The block diagram for computing MFCC is given in Figure 2. The Mel scale is approximately linear frequency spacing below 1000Hz, and a logarithmic spacing above 1000Hz. Therefore approximate formula to compute the Mel's for a given frequency f in Hz is given by:

$$mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

Where f denotes the real frequency and mel(f) denotes the perceived frequency.

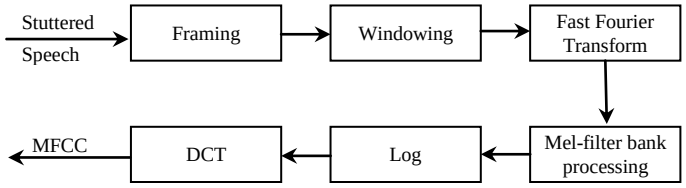


Figure 2: Schematic diagram of MFCC computation

MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore is best known and most commonly used feature in speech recognition. The step-by-step computations of MFCC are discussed briefly in this section.

i. Step1 : Framing

In framing, the pre-emphasized signal is split into several frames and each frame is analyzed in 10-30ms short time duration [18]. In this work, the frame length is set to 25ms with 10ms overlap between two adjacent frames to ensure stationary between frames.

ii. Step2 : Windowing

The effect of the spectral artifacts from framing process is reduced by windowing [18]. Windowing is a point-wise multiplication between the framed signal and the window function. In frequency domain, this combination becomes the convolution between the short-term spectrum and the transfer function of the window. A good window function has a narrow main lobe and low side lobe levels in their transfer function [18]. The Hamming window is applied to minimize the spectral distortion and the signal discontinuities. Hamming window function is shown in equation 4.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \quad (4)$$

If the window is defined as w(n), $0 \le n \le$ N-1. Then the result of windowing signal is

$$Y(n) = X(n) \times W(n) \quad (5)$$

Where: N = number of samples in each frame

Y(n) = Output signal

X (n) = input signal

W (n) = Hamming window

iii. Step 3 : Fast Fourier Transform (FFT)

The purpose of FFT is to convert the signal from time domain to frequency domain. The basis of performing Fourier transform is to convert the convolution of the glottal pulse and the vocal tract impulse response in the time domain into multiplication in the frequency domain [19]. The equation is given by:

$$Y(w) = FFT\ \ h(t) * X(t)\ \ = H(w) * X(w) \quad (6)$$

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

iv. Step 4 : Mel frequency wrapping

The Mel-frequency warping is normally realized by filter banks. Filter banks can be implemented in both time domain and frequency domain. For the purpose of MFCC processing, filter banks are implemented in frequency domain. Mel filter banks are applied on the FFT spectrums. The filter banks have triangular band pass frequency response. The spacing as well as the bandwidth is determined by a constant Mel-frequency interval as shown in Figure 3.

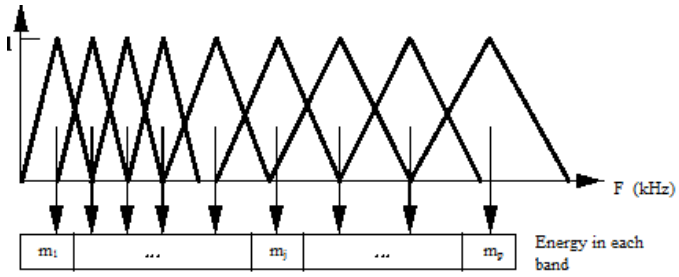


Figure 3: MEL Scale filter bank

The Mel-Frequency analysis of speech is based on human perception. Human ear acts as a filter and concentrates on certain frequency range. In speech, the signal information passed by low frequency component is more important than high frequency component. In order to highlight the low frequency components, Mel scaling is performed. Non-uniform spacing of Mel filter banks on the frequency axis is done through more filters in the low frequency and less filters in the high frequency regions [20]. The elements of each filter are estimated by summing up the convolution result of the power spectrum with a given filter amplitude, according to the formula:

$$S_k = \sum_{j=0}^{J} P_j A_{k,j} \quad (7)$$

where: Sk is power spectrum coefficient, J is subsequent frequency ranges from FFT analysis, Pj is average power of an input signal for j frequency and Ak,j is k-filter coefficient

v. Step 5 : Discrete Cosine Transform (DCT)

In this step, log Mel spectrum is converted back to time domain using DCT as shown in Figure 4. The result of conversion is called Mel Frequency Cepstrum Coefficients (MFCCs). The speech signal represented as a convolution between slowly varying vocal tract impulse response (filter) and quickly varying glottal pulse (source). Similarly speech spectrum consists of the spectral envelope (low frequency) and the spectral details (high frequency).

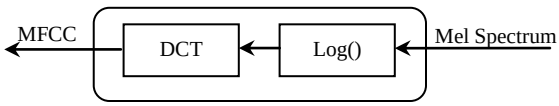


Figure 4: MEL Cepstrum Coefficients

The logarithm has the effect of changing multiplication into addition. The same is applied to separate the spectral envelope and spectral details from the magnitude spectrum. Then, we take the DCT of the logarithm of the magnitude spectrum [21]. With Sk values for each filter given, cepstrum parameter in Mel scale can be estimated by following equation [9].

$$MFCC_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], n = 1, 2, 3, \ldots N \quad (8)$$

where: N is required number of MFCC parameters, Sk is power spectrum coefficient, K is number of filters.

D. Vector Quantization

Vector Quantization (VQ) is a lossy data compression method based on the principle of block coding [22]. It is used to preserve the prominent characteristic of data. VQ is one of the ideal methods to map huge amount of vectors from a space to a predefined number of clusters, each of which is defined by its central vector or centroid [23].

i. Codebook Generation

The first step is to build a dysfluency database. The separate codebook is generated for repetition, prolongation and interjection datasets. Each consisting of N codebooks one for each dysfluency type, as given in equation (9) and (10).

$$REPdatabase = C_{r1}, C_{r2}, \ldots C_{rN} \quad (9)$$

$$PROdatabase = C_{p1}, C_{p2}, \ldots C_{pN} \quad (10)$$

$$INTERdatabase = C_{i1}, C_{i2}, \ldots C_{iN} \quad (11)$$

Where: Cr1 - CrN, Cp1 - CpN and Ci1 - CiN are codebooks for repetitions, prolongations and interjections respectively. This is done by first converting the raw input signal into a sequence of feature vectors X={X1 X2… XT}. These feature vectors are clustered into a set of M codewords X= {C1, C2, ... CM}. The set of codewords is called a codebook. To realize this, K-means clustering algorithm is used.

The K-means algorithm is a straightforward iterative clustering algorithm that partitions a given dataset into a user-specified number of clusters K [24]. Training code vector generation begins with arbitrary initial estimate, continues with iterative nearest neighbor and centriod techniques until a termination criterion is satisfied. The K-means algorithm partitions the T feature vectors into M centroids. The algorithm first randomly chooses M cluster centroids among the X feature vectors [25]. Then each feature vector is assigned to the nearest centroid and new centroids are calculated for new clusters. This procedure is continued until the mean square error between the feature vectors and the cluster-centroids is below a certain threshold or no change in the cluster center assignment. The squared error function given in equation (12), indicates the distance of the n data points from their respective cluster centers.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \quad (12)$$

Where: k is number of clusters, is a distance measure between a data point and the cluster center cj. In brief, the K-means algorithm is composed of the following steps:

$\| x_i^{(j)} - c_j \|^2$

$x_i^{(j)}$

1. Clusters the data into k groups where k is predefined.
2. Selects k points at random as cluster centers.
3. Assigns objects to their closest cluster center according to the Euclidean distance function.
4. Calculates the centroid or mean of all objects in each cluster.
5. Repeats steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

Figure 5 shows the flowchart to generate the codebook using the K-means algorithm.

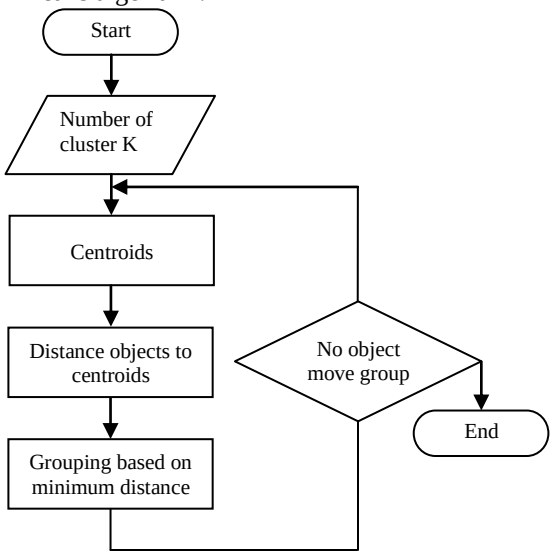


Figure 5: Flow diagram of the K-means algorithm

V. RESULT AND DISCUSSIONS

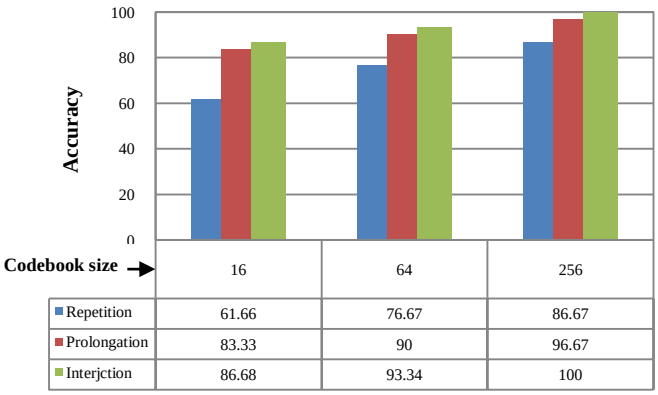
The speech samples are selected from UCLASS database. From the selected speech samples, we have created 150 speech segments of repetition, prolongation and interjection. Using these set of segments, we created training and testing group. The 80% of the segment is used for training and 20% for testing as shown in Table 2. The experiment is conducted three times using different codebook size such as K=16, 64 and 256. Each time different training and testing sets were built randomly. The classification results with different codebook size are presented in Figure 6.

In this work, to get better classification results, VQ codebook is constructed for each dysfluency by clustering the training features. Euclidean distance is employed to measure the proximity between two feature vectors. The results demonstrate substantial influence of VQ on the recognition rate.

Table 2: The Speech Data

	Speech segments	Training	Testing
Repetition	50	40	10
Prolongation	50	40	10
Interjections	50	40	10
Total	150	120	30

In the recognition phase an unknown dysfluency sample, represented by a sequence of feature vectors{x1,x2, …xT} , is compared with the codebooks in the database.



Codebook size	16	64	256
Repetition	61.66	76.67	86.67
Prolongation	83.33	90	96.67
Interjction	86.68	93.34	100

Figure 6: Classification Result

For each codebook a distortion measure is computed using Euclidean distance function. One among repetition, prolongation and interjection is chosen as dysfluency based on lowest distortion. The Euclidean distance is defined by:

$$d_e(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2} \qquad (13)$$

Where xj is the jth component of the input vector and yij is the jth component of the codeword yi.

VI. Conclusion

In this paper a new approach for classification of three types of stuttering dysfluency such as repetition, prolongation and interjection is presented. The feature extraction was performed using MFCC algorithm. The VQ codebook is generated by clustering the training feature vectors of each dysfluency and then stored in the dysfluency database. In this method, the K-means algorithm is used for clustering purpose.

A distortion measure based on minimizing the Euclidean distance was used to match the unknown dysfluency with the dysfluency database. The qualitative analysis of the visual results indicates that as number of centroids increase, identification rate of the system increases. Finally, the proposed VQ and MFCC framework yielded best accuracy of 86.67%, 96.67 and 100% for repetition, prolongation and interjection respectively.

References

[1] Young, M. A., "Predicting Ratings of Severity of Stuttering" [Monograph], Journal of Speech and Hearing Disorders, Pp. 31-54, 1961.

[2] Sherman.D, "Clinical and Experimental use of the Iowa Scale of Severity of Stuttering", Journal of Speech and Hearing Disorders, Pp. 316-320, 1952.

[3] Cullinan. W.L, Prather. E.M, & Williams. D, "Comparison of Procedures for Scaling Severity of Stuttering", Journal of Speech and Hearing Research, Pp. 187-194, 1963.

[4] Oliver Bloodstein, "A Handbook on Stuttering", 5th Edition, Singular Publishing Group, Inc., San-Diego and London, 1995.

[5] M N Hegde, Deborah Davis, Text book on "Clinical Methods and Practicum in Speech Language Pathology", 5th Edition, Cengage learning publisher, 2005.

[6] Andrzej Czyzewski, Andrzej Kaczmarek, Bozena Kostek, "Intelligent Processing of Stuttered Speech", Journal of Intelligent Information Systems archive, Volume 21, Issue 2, Pp. 143-171, September 2003.

[7] K. Ravikumar, B. Reddy, R. Rajagopal, and H. Nagara, "Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies", Proceedings of World Academy Science, Engineering and Technology, Pp. 270-273, 2008.

[8] Wiśniewski. M, Kuniszyk-Jóźkowiak. W, Smołka .E and Suszyński .W, "Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach", In Computer Recognition Systems, Volume 45/2008, Springer Berlin/Heidelberg, Pp. 445–453, October 2007.

[9] Wiśniewski. M, Kuniszyk-Jóźkowiak. W, Smołka. E and Suszyński. W, "Automatic Detection of Prolonged Fricative Phonemes with the Hidden Markov Models Approach", Journal of Medical Informatics & Technologies, Vol. 11, Pp. 293-298, 2007.

[10] Ravikumar. K. M, Rajagopal. R and Nagaraj. H. C, "An Approach for Objective Assessment of Stuttered Speech using MFCC Features", ICGST International Journal on Digital Signal Processing DSP Volume 9, Pp. 19–24, 2009.

[11] Lim Sin Chee & Ooi Chia Ai & M. Hariharan & Sazali Yaacob, "MFCC Based Recognition of Repetition and Prolongation in Stuttered Speech using k-NN and LDA", in proceedings of 2009 IEEE Student Conference on Research and Development (SCOReD 2009), Malaysia, Pp. 146-149, 16-18 Nov 2009.

[12] Lim Sin Chee & Ooi Chia Ai & M. Hariharan & Sazali Yaacob, "Automatic Detection of Prolongations and Repetitions using LPCC", International Conference for Technical Postgraduates, Kuala Lumpur, 14-15 Dec 2009.

[13] P.Mahesha and D S Vinod, "Automatic Classification of Dysfluencies in Stuttered Speech using MFCC", International Conference on Computing Communication & Information Technology (ICCCIT 2012), Chennai, Pp. 181-184, 27-29June 2012.

[14] P.Howell, S.Devis and J.Batrip, "The UCLASS Archive of Stuttered Speech", Journal of Speech, Language and Hearing Research, Volume 52, Pp. 556-559, April 2009.

[15] P.Howell and M. Huckvale, "Facilities to Assist People to Research into Stammered Speech", Stammering Research: An On-line Journal Published by the British Stammering Association, Vol. 1, Pp. 130-242, 2004.

[16] J. Harrington, S. Cassidy, Text book "Techniques in Speech Acoustics", Kluwer Academic Publishers, Dordrecht, 1999.

[17] L. Rabiner and B.H. Juang, "Fundamental of Speech Recognition", PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[18] J. G. Proakis, D. G. Manolakis, "Digital Signal Processing. Principles, Algorithms and Applications", III Edition, Macmillan, New York, 1996.

[19] C. Becchetti and Lucio Prina Ricotti, "Speech Recognition", John Wiley and Sons, England, 1999.

[20] Speech Technology: "A Practical Introduction, Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis", Carnegie Mellon University and International Institute of Information Technology, Hyderabad.

[21] Ibrahim Patel and Y. Srivinasa Rao, "A Frequency Spectral Feature Modeling for Hidden Markov Model Based Automated Speech Recognition", the second International conference on Networks & Communications, Chennai, 2010.

[22] Y. Linde, A. Buzo & R. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Vol. 28, Pp. 84-95, 1980.

[23] K. Sayood, "Introduction to Data Compression", Second Edition, Morgan Kaufmann Publishers, San Francisco, California, 2000.

[24] J. Ghosh, A. Liu, "The K-means Algorithm", In: X. Wu and V. Kumar eds. The Top Ten Algorithms in Data Mining, Boca Raton, FL: Chapman & Hall, 2009.

[25] Rosenberg. A, Rabiner.L and Juang. B, "A Vector Quantization Approach to Speaker Recognition", IEEE International Conference on ICASSP'85, Vol. 10, Pp. 387–390, 1985.

P. Mahesha received his Bachelor's Degree in Electronics and Communications Engineering from University of Mysore, India. Master's Degree in Software engineering from the Visvesvaraya Technological University (VTU), Belgaum, India and currently he is pursuing his PhD under VTU. He has published 4 International Conference papers related to his research area. He is currently working as Assistant Professor at the Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India. He has teaching experience of 7 years. His research interests include Speech Signal Processing, Web Technologies and Software Engineering.

D.S. Vinod received his Bachelor's Degree in Electronics and Communications Engineering and Master's Degree in Computer Engineering from the University of Mysore, India. He did his PhD at Visvesvaraya Technological University (VTU). He did his research work on Multispectral Image Analysis and published 2 International Journals and 10 International Conference papers related to his research area. He is currently working as Assistant Professor at the Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, India. He has teaching experience of 13 years and he was awarded UGC-DAAD Short-term fellowship, Germany in the year 2004 -05. His research interests include Image Processing, Speech Signal Processing, Machine Learning and Algorithms.