Comparison of Multidimensional MFCC Feature Vectors for Objective Assessment of Stuttered Disfluencies

Ravi Kumar K M
Professor & H.O.D, Dept of I.S.E, Ghousia College of Engineering, Ramanagaram, Bangalore
Email: kmravikumar@rediffmail.com
Ganesan S
Lecturer, Dept of I.S.E, Ghousia College of Engineering, Ramanagaram,Bangalore
Email: sganesh41@yahoo.co.in

ABSTRACT

The objective approach has an advantage over the manual, which provides consistence measurement required for assessment of stuttered speech. The number of dimensions (multi dimension) plays a key role in objective assessment of stuttering. The purpose of this paper is to analyze the multidimensional MFCC features and identify which dimensional provides better accuracy. In our work 10 samples in the age group of 25 – 30 years were collected. In which 80% were used for training and remaining 20% for testing. The MFCC features of 12, 13, 26 and 39 dimensional MFCC are compared and it is found that 39 dimensional MFCC are better for assessment of stuttered speech objectively, with 84.58 % accuracy.

Keywords - Assessment; Multidimensional; Objective; Stuttered speech;

Date of Submission: 25 November 2010 Revised: 03 February 2011 Date of Acceptance: 26 February 2011

1. INTRODUCTION

Detection of syllable repetition is one of the important factors in assessing the stuttering speech objectively. This detection method comprising of four stages: segmentation, feature extraction, score matching and decision logic. Segmentation is done by manually, though many methods are available. Feature extraction is implemented using Mel Frequency Ceptra Coefficient (MFCC). Score matching is done by Dynamic Time Warping (DTW) between the syllables. The decision logic is implemented by confusion matrix based on the score given by score matching. Stuttering is a speech disorder in which the normal flow of speech is disrupted by frequent repetition of words or prolongation of speech sounds or syllables. Latest research on the epidemiology of stammering declares that most children stammers between the ages of 2 to 5, the highest peak of stammering is found at the age 4. In cases of injury like stroke or trauma to the brain, acquired stammering may occur. Stuttering exhibits a wide variety of behavioral and psychological symptoms. It is a multi-dimensional problem which involves a particular kind of speech behavior, feelings, beliefs, self-concepts and social interactions. Stuttering is a social-emotional problem as well as a speech problem. People who stutter often display intense fear of speaking and also experiences repeated frustration while communicating and express dissatisfaction. People who stutter are just reacting normally to the stress of their communication disorder.

Types of Stuttering: (i) Developmental stuttering – this is the most common type of stuttering that usually occurs in children. They may not be able to meet verbal demands as their speech and language are underdeveloped. (ii) Neurogenic stuttering – this is usually caused by signal problems that occur between the brain, nerves and muscles.
(iii) Psychogenic stuttering – this is believed to originate in the brain that directs thought and reasoning. This is mostly found in patients with histories of mental illness or mental stress.

Current research suggests that it is caused by a complex interaction between a person's physical makeup and the environment. Stuttering may result when certain factors in the environment combine to produce disfluent speech in a child who is physiological prone to it. The most form of stuttering is thought to be developmental, which is occurring in children who are in the process of developing speech and language. This relaxed type of stuttering is felt to occur when a child's speech and language abilities are unable to meet his or her verbal demands. If stammering is left untreated, the risk of developing an anxiety disorder in later life may be greater. Longitudinal research following children with speech or language disorders from 5 years of age has consistently found that in early adulthood they have increased rates of anxiety disorders compared with other psychiatric illnesses such as schizophrenia or eating disorder. Negative perceptions and concerns about stammering develop by about 10 years of age. conventional way of making stuttering assessment are to count the occurrence of these types of disfuencies and express them either as the number of disfluent words as a proportion of all words in a passage or measure the time the disfluencies take compared with the duration of the entire message. The main difficulties in making such counts are time consuming, poor agreement, etc.

A Confusion matrix is a visualization tool typically used in supervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. When a data set is unbalanced, the error rate of a classifier is not representative of the true performance of the classifier. This can be easily understood by an example. If there are 990 samples from class A and 10 samples from class B, the classifier can easily be biased towards class A. If the classifier classifies all the samples as class A, the accuracy will be 99 %. This is not a good indication of the classifier’s true performance. The classifier has a 100 % recognition rate for class A but a 0 % recognition rate for class B.

TABLE 1 TABLE OF CONFUSION

		Actual value		
		P	N	Total
Prediction outcome	P’	True Positive	False Positive	P’
	N’	False Negative	True Negative	N’
	Total	P	N	

In predictive analytics, a table of Confusion, also known as a confusion matrix, is a table with two rows and two columns that reports the number of True Negatives, False Positive, False Negative and True Positive.

For example, consider a model which predicts for 10,000 insurance claims whether each case is fraudulent. This model correctly predicts 9,700 non-fraudulent cases, and 100 fraudulent cases. The model also incorrectly predicts 150 cases which are not fraudulent to be fraudulent and 50 cases which are fraudulent to be non-fraudulent. The resulting table of confusion is shown below.

TABLE 2 EXAMPLE OF TABLE OF CONFUSION

		Actual value		
		P	N	Total
Prediction outcome	P’	100	150	P’
	N’	50	9700	N’
	Total	P	N	

2. LITERATURE SURVEY

To enable objective assessment of stuttering, in a more reliable way, the features extracted will play the major role. This chapter described the different feature vectors used and their reliability. The present work is the extension of the work done by K.M.Ravikumar et al (2008), where they have used 12 dimensional MFCC feature vector to recognize the disfluency. Altered Auditory Feedback systems for Adult Stuttered are available which are widely used for treatment[13]. The samples were obtained by making the client / patient read a Standard English passage. It is also described in the paper that, different feature extraction methods may be tried to increase accuracy.

3. METHODOLOGY

3.1. SUBJECTS AND MATERIALS

A group of people between the age group of 25 to 30 were monitored. A Standard English passage of 150 words was selected for preparing the data base. The clients were made to read the passage and these speech samples were recorded using cool edit version 2 at sampling rate of 16000 samples per second with 16 bits representation [12]. Ten samples were collected, out of which eight samples were used for training and two samples were used for testing.

3.2 APPROACH

Automatic Detection Method: The process of counting stuttering events could be carried out more objectively through the automatic detection of stop-gaps, syllable repetitions and vowel prolongations. The alternative could be based on the subjective evaluations of speech fluency and may be dependent on a subjective evaluation method. This method requires vectors of parameters, which characterize the distinctive features in a subject's speech patterns. In addition, an appropriate selection of the parameters and feature vectors while learning may augment the performance of an automatic detection system. The detection procedure is divided into four.

(i) Segmentation: The characteristic feature of the syllable is the dynamical transient part consonant-vowel or consonant-vowel-consonant. For automatic segmentation of syllable many methods are available, which uses signal extremes, first autoregressive (AR) coefficient, etc[17]. The speech samples collected in the data bases are segmented manually to obtain the syllables. The segmented speech syllables are subjected to feature extraction.

(ii) Feature extraction: Speech recognition at its most elementary level comprises a collection of algorithms including statistical pattern recognition, communication theory, signal processing and linguistics. The signal processing converts the speech waveform to some of type of parametric representation. This parametric representation is then used for further analysis and processing. The speech signal is analyzed in successive narrow time windows of ten milliseconds width, for its frequency content with two millisecond offset [17]. For each and every window we obtain the intensity of several bands on the frequency scale using feature extraction algorithm.

There are different types of feature extraction LPC (Linear Prediction Coefficient Cepstra), MFCC (Mel Frequency Cepstra Coefficient), PLP (Perceptual Linear Prediction Cepstra).

Linear Predictive Cepstral coding computes a LPC spectral envelope, before converting it into cepstral coefficient. Linear predictive analysis of speech has become the predominant technique for estimating the basic parameters of speech. This analysis provides both an accurate estimate of the speech parameters and also an efficient computational model of speech. The basic idea behind this analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences between the actual speech samples and linear predicted values a unique set of parameters or predictor coefficient can be determined. These coefficients form the basic for linear predictive analysis of speech. The LPCC is the most used coding method in speech recognition.

Mel frequency cepstra coefficient is based on signal decomposition with the help of a filter bank, which uses the Mel scale expressed on the Mel-frequency scale. The MFCC results of a discrete cosine transform of the real logarithm of the short term energy. Mel scale cepstral analysis is very similar to perceptual linear predictive analysis of speech, where the short term spectrum is modified based on psychophysically based spectral transformations. In this method, the spectrum is warped according to the MEL scale, where as in PLP the spectrum is warped according to the Bark scale. The main difference between Mel scale cepstral analysis and perceptual linear prediction is related to the output cepstral coefficients. The output cepstral coefficients are then computed based on this model. In contrast Mel scale cepstral analysis uses cepstral smoothing to smooth the modified power spectrum. This is done by direct transformation of the log power spectrum to the cepstral domain using an inverse Discrete Fourier Transform (DFT). The MFCC has good performances in speech recognition.

Perceptual linear prediction is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short-term spectrum of the speech by several psychophysically based transformations. The PLP cepstral coefficients are computed using the PLP functions defined in the analysis library. Most feature extraction package produce a multi dimensional feature vector for every frame of speech. This study considers 12, 13, 26, 39 MFCC.

The ceptral coefficients are a set of features reported to be robust in some different pattern recognition tasks concerning human voice. The human voice is very well adapted to the ear sensitivity. The energy developed in speech being comprised in the lower frequency energy spectrum, below 4 KHz. In speech recognition tasks, usually the 12 coefficients are retained, that they represent the slow variations of the spectrum the signal characterizing the vocal tract shape, the spectrum of shuttered words [7].

The Mel-scale equivalent value for frequency f expressed in HZ is

$$\text{Mel}\,(f) = 2595 \log_{10}\left(1+ \frac{f}{700}\right) \tag{1}$$

The MFCC's are computed by redistributing the linearly spaced bins of the log-magnitude Fast Fourier Transformer (FFT) into Mel-spaced bins according to the above equation and applying discrete cosine Transform (DCT) on the redistributed spectrum. The 13, 26, and 39 dimension MFCC's are calculated using below formulas:

$$c(t) = \text{mel}(f) + \text{DCcomponent} \tag{2}$$

$$\Delta c(t) = c(t + \tau) - c(t - \tau) \tag{3}$$

$$\Delta\Delta c(t) = \Delta c(t + \tau) - \Delta c(t - \tau) \tag{4}$$

(iii) Score matching: we have done the score matching using DTW. The DTW procedure combines alignment and distance computation in one dynamic programming procedure. DTW assumes that (a) global variation in speaking rate for a person uttering the same word at different times can be handled by linear time normalization (b) local rate variations within each utterance are small and can be handled using distances penalties (c) each frame of test utterance contributes equally to recognition (d) single distance measure applied uniformly across all frames is adequate. These give intuitive distance measurements between time series by ignoring both global and local shifts in the time dimension. The 12 dimensional MFCC obtain for each syllable are used to compute the angle between them which serve as local distance and represent in the form of matrix. Using Dynamic Programming (DP) the min-cost path through matrix is found [4, 6]. These values were given to decision logic to identify whether the syllable were repeated or not.

(iv) Decision logic: The distance between the syllable obtained in the form of scores is used to decide whether repetition has occurred or not. Out of ten samples collected, the scores of eight samples are used to fix up the threshold to separate repetition (R) and non repetition (NR). Using the threshold obtained, the scores of remaining two samples are checked and the confusion matrix is drawn.

4. RESULTS

The separation of two classes of data, for test data1 and test data2 is shown in figure 1 to 8. The confusion matrix for two test data with respect to 12, 13, 26, and 39 is shown and the overall accuracy is computed. Initially the confusion matrix for score of 12 MFCC was compared with 13 MFCC, which indicated no improvement. The process was continued for 26 MFCC and 39 MFCC, which clearly indicated the improvement over 12 MFCC. The confusion matrix shows the efficiency of multidimensional MFCCs. Fig. 9 and Fig. 10 shows the bar chart comparing the multidimensional MFCC with respect to test data 1 and test data 2.

1. Confusion matrix for 12 dimension test data 1 = $\begin{bmatrix} 95.68\% & 4.32\% \\ 40.82\% & 59.18\% \end{bmatrix}$

Percentage of accuracy = (95.68 % 59.18 %)
Overall accuracy = (77.43 %)

2. Confusion matrix for 12 dimension test data 2 = $\begin{bmatrix} 96.11\% & 3.89\% \\ 28.21\% & 71.79\% \end{bmatrix}$

Percentage of accuracy = (96.11 % 71.79 %)
Overall accuracy = (83.95 %)

3. Confusion matrix for 13 dimension test data 1 = $\begin{bmatrix} 97.41\% & 2.59\% \\ 69.39\% & 30.61\% \end{bmatrix}$

Percentage of accuracy = (97.41 % 30.61 %)
Overall accuracy = (64.01 %)

4. Confusion matrix for 13 dimension test data 2 = $\begin{bmatrix} 96.90\% & 3.10\% \\ 51.3\% & 48.71\% \end{bmatrix}$

Percentage of accuracy = (96.90 % 48.71 %)
Overall accuracy = (72.81 %)

5. Confusion matrix for 26 dimension test data 1 = $\begin{bmatrix} 98.27\% & 1.73\% \\ 32.66\% & 67.34\% \end{bmatrix}$

Percentage of accuracy = (98.27 % 67.34 %)
Overall accuracy = (82.81 %)

6. Confusion matrix for 26 dimension test data 2 = $\begin{bmatrix} 96.11\% & 3.89\% \\ 25.65\% & 74.35\% \end{bmatrix}$

Percentage of accuracy = (96.11 % 74.35 %)
Overall accuracy = (85.23 %)

7. Confusion matrix for 39 dimension test data 1 = $\begin{bmatrix} 96.55\% & 3.45\% \\ 30.62\% & 69.38\% \end{bmatrix}$

Percentage of accuracy = (96.55 % 69.38 %)
Overall accuracy = (82.97 %)

8. Confusion matrix for 39 dimension test data 2 = $\begin{bmatrix} 98.03\% & 1.97\% \\ 25.65\% & 74.35\% \end{bmatrix}$

Percentage of accuracy = (98.03 % 74.35 %)
Overall accuracy = (86.19 %)

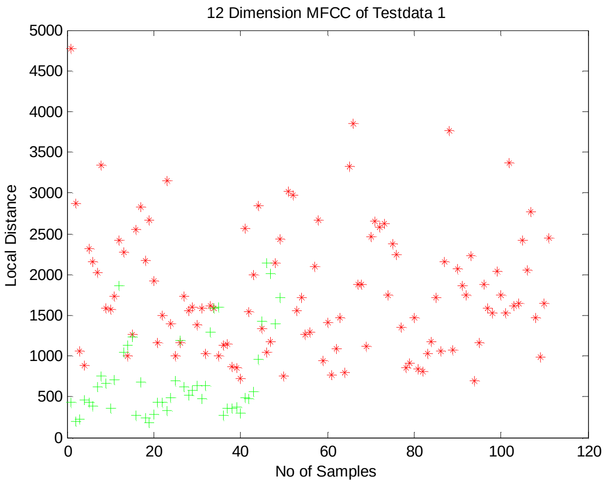

Figure 1: 12 dimensional Test data 1 for two classes of data

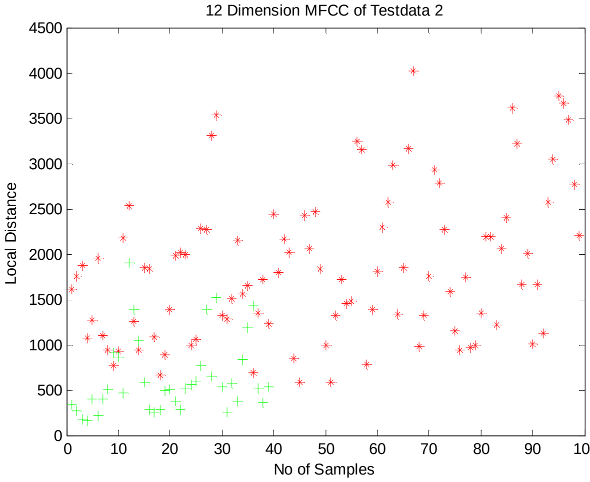

Figure 2: 12 dimensional Test data 2 for two classes of data

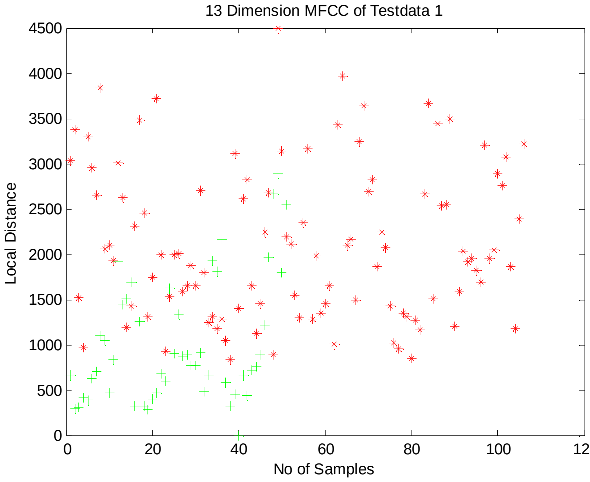

Figure 3: 13 dimensional Test data 1 for two classes of data

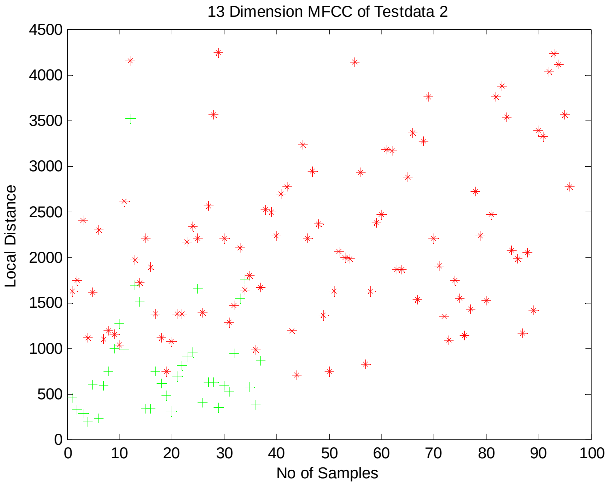

Figure 4: 13 dimensional Test data 2 for two classes of data

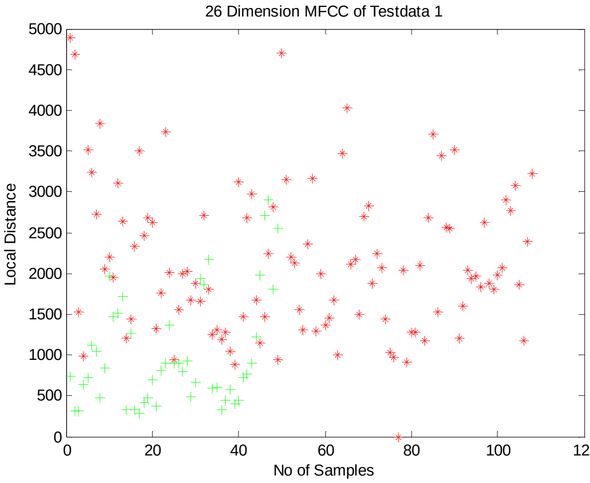

Figure 5: 26 dimensional Test data 1 for two classes of data

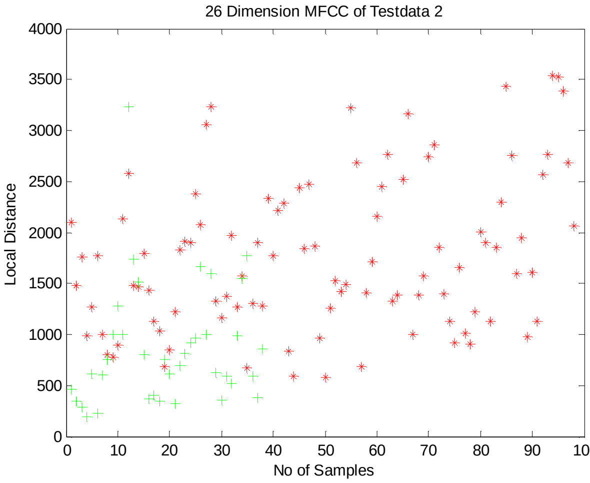

Figure 6: 26 dimensional Test data 2 for two classes of data

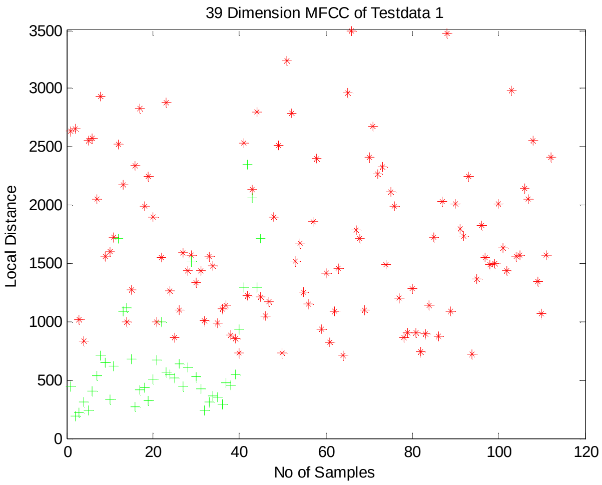

Figure 7: 39 dimensional Test data 1 for two classes of data

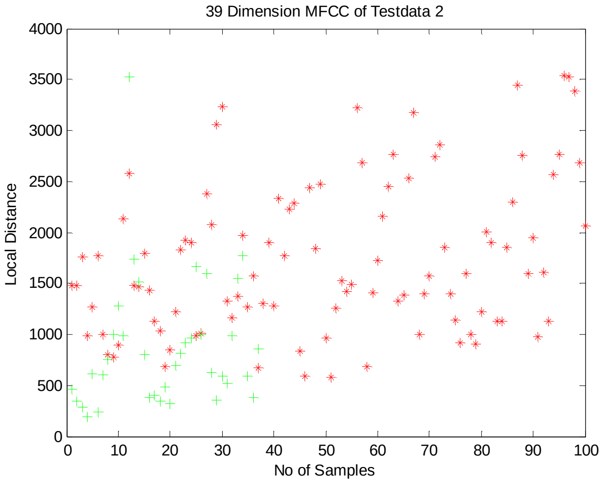

Figure 8: 39 dimensional Test data 2 for two classes of data

TABLE 3. COMPARISON OF MULTIDIMENSIONAL MFCC

	Test data 1	Test data 2
12 Dimension MFCC	77.43 %	83.95 %
13 Dimension MFCC	64.01 %	72.80 %
26 Dimension MFCC	82.80 %	85.23 %
39 Dimension MFCC	82.96 %	86.19 %

5. CONCLUSION

From table 3 it is clear that, the 39 MFCC separate the two classes of data more precisely than the other multidimensional MFCC. Therefore for objective assessment of stuttered disfluencies, the 39 dimension MFCC feature vector obtained for each syllable performs better than other multidimensional feature vectors. Compared to earlier methods [8,9] which uses Artificial Neural Network (accuracy 78%) and Hidden Markov Model (accuracy 81%), the present work using 39 dimensional MFCC provides better results with 84.58%. Due to this improvement the work done to obtain result in [15,16] may be improved further. As a future work to check for improvements other feature extraction methods like IMFCC (Inverse Mel Frequency Cepstral Coefficient) may be tried.

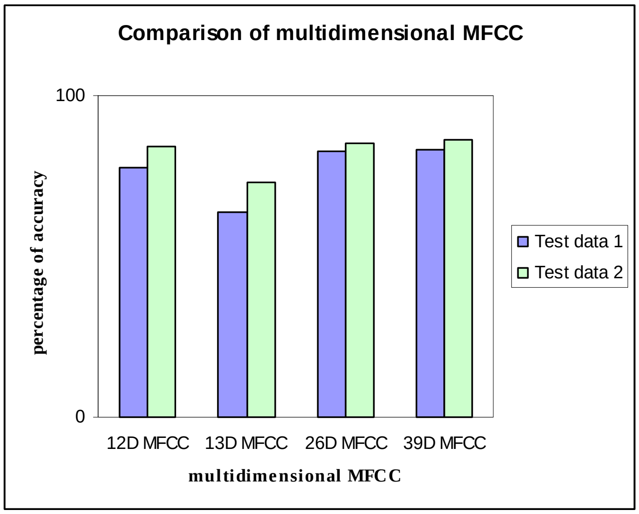

Fig. 9 : Comparison of multidimensional MFCC for testdata1 & testdata2

ACKNOWLEDGMENT

The authors would like to thank Dr. Chandrashekar Institute of Speech and Hearing, Lingarajapura, Bangalore for enabling us to collect data for our work in their institution.

REFERENCES

[1] D. Kally and E.Boerg, "*An investigation of inter-clinic agreement in the identification of fluent and stuttered syllabus", Journal of fluency disorders, vol . 13*, pp.309-318 , 1988.

[2] Dalouglas O Shaughnessy, "*Speech Communications ", Human and Machine, University Press,* second edition, 2001.

[3] E.G.Conture, "*Englewood cliffs*", New Jersey: Prentice Hall, second edition, 1990.

[4] E.Keogh, "*Exact indexing of time warping" , in VLDB pp.*406-417. Hong Kong, Chine, 2002.

[5] E.Yairi & B.Lewis, "*Disfluencies at the onset of stuttering", Journal of speech & hearing Research, vol.27*, pp.154-159, 1984.

[6] H. Silverman & D. Morgan, "*The application of dynamic programming to connected speech segmentation", IEEE ASSP Mag.7, no.3*, 7-25, 1990.

[7] L. Rabiner and B.H.Juang. "*Fundamental of speech recognition", PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.*

[8] Peter Howell, Stevie Sacking and Kazan Glen, *"Development of a Two-stage procedure for the*

Automatic Recognition of Disfluencies in the speech of children who stutter. I. Psychometric procedure appropriate for selection of training material for Lexical Disfluency Classifiers", JSLHR, vol. 40 , pp.1073-1084, October 1997.

[9] Peter Howell, Stevie Sacking and Kazan Glen, "*Development of a Two-stage procedure for the Automatic Recognition of Disfluencies in the speech of children who stutter. II ANN Recognition of Repetitions and prolongation with supplied word segment markers", JSLHR, vol. 40,* pp.1085-1096, October 1997.

[10] Tack Mu Kuson, Michael E. Zervakis, "*Gaussian Perceptron : Learing Algorithms", IEEE International Conference on Systems, Man and Cybernetics, vol. 1* pp. 105-110, October 1992.

[11] W. Johnson et al., "*The onset of stuttering, minneapolies university of minnesata press*", 1959

[12] K.M. Ravikumar, Balakrishna Reddy, R. Rajagopal and H.C. Nagaraj, "*Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies"*, 2008.

[13] K.M.Ravikumar, and Dr. R.Rajagopal "*Altered Auditory Feedback Systems for Adult Stutter", Proceedings of the Sonata international Conference on Computer Communication and Control,* November 2006, pp. 193-196.

[14] K.M.Ravikumar , Sachin Kudva , Dr. R.Rajagopal and Dr.H.C.Nagaraj, "*Development of a Procedure for the Automatic Recognition of Disfluencies in the Speech of People who Stutter", International Conference on Advanced Computing Technologies,* Hyderabad, India, December 2008, pp. 514-519.

[15] K.M.Ravikumar , S.Mahadev , Dr. R.Rajagopal and Dr.H.C.Nagaraj, "*An Algorithm to Compute the number of Repetitions and its Iterations from Stuttered Speech for Objective Assessment of Stuttering, ICOICT, International Conference on Optoelectronics", Information and Communication Technologies,* Trivandrum, Kerala, India, February 2009, pp. 147-150.

[16] K.M.Ravikumar , T. Satish , Dr. R.Rajagopal and Dr.H.C.Nagaraj, "*Bayseian Classifier for Classification of Normal Nonfluency and Fluent Speech from Stuttered Disfluencies", ICOICT,* International Conference on Optoelectronics, Information and Communication Technologies, Trivandrum, Kerala, India, February 2009, pp. 202-205.

[17] K.M.Ravikumar , Dr. R.Rajagopal & Dr.H.C.Nagaraj, "*An Approach for Objective Assessment of Stuttered Speech Using MFCC Features", ICGST, International Journal on Digital Signal Processing, Vol.9,* June 2009, Issue.1, pp. 19-24.

Authors Biography

RAVIKUMAR K.M, Professor and Head of Information Science and Engineering, Ghousia collage of Engineering, Ramanagaram. He has completed his M.Tech in the field of Biomedical Instrumentation in the year 2002 and submitted his PhD thesis in 2009 (VTU, Belgaum). He is in the field of teaching from past 13 years and he has published ten papers in the International Conference and three in the International Journal related to his research areas. He is member of various professional societies which include ISTE, MIE, SIS, and BMESI. His field of interest includes Digital Signal Processing and Communication Systems.

GANESAN S, M.Tech (IT) working as a Lecturer in the Department of Information Science and Engg., Ghousia College of Engineering, Ramanagaram, Bangalore. He has nine years of teaching experience. He is life member of ISTE. His field of interest includes Data Structure, DBMS and Software Engineering.