



An Approach for Objective Assessment of Stuttered Speech Using MFCC Features

K.M Ravikumar, R.Rajagopal, H.C.Nagaraj

Nitti Meenakshi Institute of Technology, Bengaluru, India

kmravikumar@rediffmail.com, Rajagopal.R@lntsec.com, principal@nmit.ac.in

Abstract

Syllable repetition is one of the important parameter in assessing the stuttered speech objectively. The existing method which uses artificial neural network (ANN) and Hidden Markov Model (HMM) requires high levels of agreement as prerequisite before attempting to train and test to separate fluent and nonfluent. We propose automatic detection method for syllable repetition in read speech for objective assessment of stuttered disfluencies which uses a new approach and has four stages comprising of segmentation, feature extraction, score matching and decision logic: Segmentation is assisted manually which is tedious but straightforward. Feature extraction is implemented using well known Mel frequency Cepstra coefficient (MFCC). Score matching is done using Dynamic Time Warping (DTW) between the syllables. The Decision logic is implemented by Support Vector Machine (SVM) and compared with our previous work which uses Perceptron method. The proposed objective approach has an advantage over the manual (subjective), which provide consistent measurement required for assessment. The assessments by human judges on the read speech of 15 adults who stutter are described. 80% of data are used for training and 20% for testing. The average result was found to be 93.45%, which is better than our previous work [80.78%] using HMM.

Keywords: *Assessment, DTW, MFCC, Objective, Perceptron, Stuttering, SVM.*

1. Introduction

Stuttering, also known as stammering in the United Kingdom is a speech disorder. The type of disfluencies that employed are: 1. Interjections (extraneous sounds and words such as “uh” and “well”); 2. Revisions (the change in content or grammatical structure of a phrase or pronunciation of a word as in “ there was a young dog , no , a young rat named Arthur”); 3. Incomplete Phrases (the content not completed); 4. Phrase-repetitions; 5. Word-repetitions; 6. Part-word-repetitions; 7. Prolonged sounds (sounds judged to be unduly prolonged);

8. Broken words (words not completely pronounced) [6].

Stuttering is often associated with “Repetitions”. As described above, part-word or syllabic repetitions are one of the defining elements of stuttering. The dominant features of Normal Nonfluent (NNF) speech reported are: 1. Word Repetitions, but not part-word Repetition is a prevalent feature of early stuttering [25]. 2. In early stuttering, there is a high proportion of Repetition in general, as opposed to other types of disfluency like prolongation [4].

Conventional way of making stuttering assessment are to count the occurrence of these types of disfluencies 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 passage manually (subjective). The main difficulties in making such counts which is subjective are: 1. It is time consuming to make and 2. There are poor agreements when different judges make counts on the same material [7]. Has these counts are subjective they are inconsistent and prone to error. Despite the fact that some researchers have several attempts to use objective methods to evaluate patients progress in speech therapy [1, 10, 11, 12], there is an always a need for improvement. In our previous work we have developed a procedure for recognition of disfluencies [14], using Hidden Markov Model (HMM) and we also tested the perceptron classifier [15]. In our present work we use speech recognition technology with a new approach to automate the disfluency counts, thus providing an objective and consistent measurement. Different Stuttering devices based on Altered Auditory Feedback namely; Delayed Auditory Feedback (DAF), Frequency Shifted Auditory Feedback (FAF), and Masked Auditory Feedback (MAF) and also Digital Speech Aid (DSA) is widely used to treat the stutterer to reduce those counts [13].

The 150 words Standard English passage was selected for preparing the database. All the 15 clients around the age group of 25 on an average were made to read



the passage and these speech were recorded using cool edit version2 at sampling rate of 16000 samples per second with number of bits to represent as 16-bits. The remainder of the paper organized as follows: section (2) focuses on Automatic detection methods and steps involved in it. Section (3) emphasizes on 15 samples collected and accuracy of new approach on those samples. Section (4) concludes by comparing present work with previous.

2. Automatic Detection Method

The detection scheme used for assessment is divided into four steps as shown in Figure 1:

Segmentation: Phonetics gives no exact specification of syllables. The characteristic feature of the syllable is the dynamical transient part consonant-vowel or consonant –vowel –consonant. The feeling of syllable boundaries, although usually very strong, is subjective and often not unique. For Automatic segmentations of syllable many methods are available, which uses signal extremes, first Autoregressive (AR) coefficient, etc [19]. The speech samples collected in the databases are segmented manually, which is tedious but straightforward [5]. The segmented speech syllables are subjected to feature Extraction.

Feature Extraction: A common first step in feature extraction is frequency or spectral analysis. The signal processing techniques aim to extract features that are related to identify the characteristics. The speech signal is analyzed in successive narrow time windows of 10msec width, for its frequency content with 2msec offset [21]. For each and every window we obtain the intensity of several bands on the frequency scale using feature extraction algorithm. Several different feature extraction algorithms exist, namely [5]

1. Linear Predictive Cepstral Coefficients (LPCC)
2. Perceptual Linear Prediction (PLP) Cepstra.
3. Mel Frequency Cepstral Coefficient (MFCC)

Most feature extraction package produce a multidimensional feature vector for every frame of speech. LPCC computes Spectral envelope, before converting it into Cepstral coefficient. The LPCC are LP-derived Cepstral coefficient. PLP integrates critical bands, equal loudness pre emphasis and intensity-to-loudness compression. The PLP is based on the Nonlinear Bark scale. It was originally designed to speech recognition with the removing of speaker dependent characteristics. MFCC is based on signal decomposition with the help of a filter bank, which uses the Mel scale. The MFCC results on Discrete Cosine Transform (DCT) of a real logarithm of the short-term energy expressed on the Mel frequency scale. Our work considers 12MFCC. The Cepstral coefficients are set of features reported to be

robust in some different pattern recognition tasks concerning human voice. They are widely used in speech recognition and also in speaker identification. The human voice is very well adapted to the ear sensitivity, most of the energy developed in speech being in the lower frequency energy spectrum, below 4 kHz. In speech recognition tasks, usually 12 coefficients are retained, which represent the slow variations of the spectrum of the signal, which characterizes the vocal tract shape of the uttered words [16].

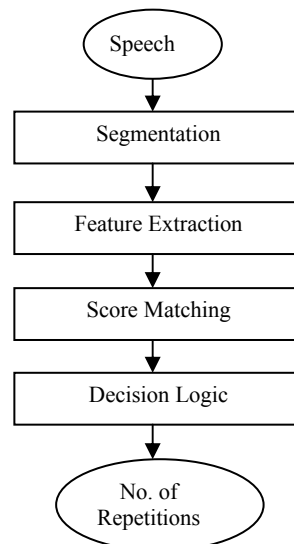


Figure 1: Block diagram of Automatic detection method

The Mel-frequency scaling is done by a bank of triangular band-pass filters, nonuniformly distributed along the frequency axis. The Mel-scale equivalent value for frequency f expressed in Hz is:

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

The MFCCs are computed by redistributing the linearly spaced bins of the log-magnitude Fast Fourier Transform (FFT) into Mel-spaced bins according to equation (1) and applying DCT on the redistributed spectrum. A relatively small number of coefficients (typically 12) provide a smoothed spectral envelope, leading to the isolation of the vocal tract response by the simple retention of the desired amount of information. An additional advantage in using MFCC is that, it has decorrelating effect on the spectral data and maximizes the variance of the coefficients, similar to the effect of Principal Component Analysis. The each dimension is a floating point value. Feature extraction modules are also called front-end or just signal processing modules.

Score Matching: In this paper, the DTW based score matching is done. The DTW procedure combines alignment and distance computation in one dynamic programming procedure. Basic DTW assumes (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 obtained for each syllable are used to compute the angle between them (normalized inner product) which serve as local-distance and represent in the form of matrix. Using Dynamic Programming (DP) the minimum-cost path through matrix is found [8, 22]. These values were given to decision logic to identify whether the syllable were repeated or not.

Decision Logic:

i) Perceptron Method: In our previous work [15] we tested the Decision logic using the Perceptron to take a decision whether a syllable is repeated or not. Perceptron was the first iterative algorithm for learning linear classification. It is a single layer network with threshold activation function:

$$y = \text{sgn}(w^T x + b) \quad (2)$$

The weight vector w is updated each time, a training point is misclassified. The algorithm is guaranteed to converge when data are linearly separable. Two classes of pattern are “linearly separable” if they can be separated by a linear hyperplane.

Suppose that target values (d_t) take either 1 or -1:

$$d_t = \begin{cases} 1 & \text{if } x \in c1 \\ -1 & \text{if } x \in c2 \end{cases} \quad (3)$$

Here we find w such that

$$\begin{aligned} w^T x &> 0 & \text{for } x \in c1 \\ w^T x &< 0 & \text{for } x \in c2 \end{aligned} \quad (4)$$

This implies that

$$w^T x d > 0 \quad \forall x \quad (5)$$

The Perceptron criterion leads to the following objective function

$$\mathcal{E}(w) = - \sum_{x_i \in m} w^T x_i d_i \quad (6)$$

Where m is the set of vectors.

The gradient of $\mathcal{E}(w)$ is:

$$\frac{\partial \mathcal{E}}{\partial w} = - \sum_{x_i \in m} x_i d_i \quad (7)$$

The basic idea behind Perceptron is shown in Figure 2.

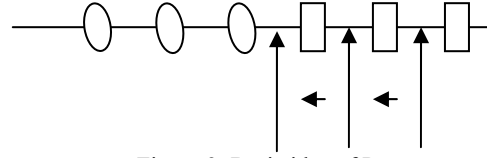


Figure 2: Basic idea of Perceptron

If distinct parameters are separated, do not move. If not, move it to the left. If the pattern is correctly classified, do nothing, else:

$$\Delta w = \eta \sum_{x_i \in m} x_i d_i \quad (8)$$

The Perceptron classifier minimizes the error probability much better than Minimum Mean Square Error (MMSE) classifier.

The Perceptron learning algorithm is given below.

- a) Get a training sample.
- b) Check to see if it is misclassified.
 - i) If classified correctly, do nothing.
 - ii) If classified incorrectly, update w by

$$\Delta w = \eta x_i d_i \quad (9)$$

- c) Repeat steps 1 and 2 until convergence

ii) Support Vector Machine (SVM): In this paper we use the SVM method to classify the fluent with that of Nonfluent. SVM [3, 18, 24] is a powerful machine learning tool which attempts to obtain a good separating hyper-plane between two classes in the higher dimensional space. The equation of the hyper-plane is:

$$w^T x + b = 0 \quad (10)$$

Where w a weight, is vector and b is the bias. Non-linearity is satisfied by mapping the input features x into higher dimensions using a function

$$\phi(x) : R^d \rightarrow R^p, p > d \quad (11)$$

And hence the hyperplane becomes:

$$w^T \phi(x) + b = 0 \quad (12)$$

This leads to the following optimization problem:

$$\min_{\xi, w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (13)$$

Subject to:

$$\begin{aligned} y_i (w^T \phi(x_i) + b) &\geq 1 - \xi_i, \quad i = 1..N, \\ \xi_i &\geq 0 \quad i = 1..N \end{aligned} \quad (14)$$

C is constant determined by a cross validation process. The dual formulation of this problem is:

$$\max_{\lambda} \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j K(x_i, x_j) \quad (15)$$



Subject to:

$$\sum \lambda_i \lambda_i = 0 \tag{16}$$

$$0 \leq \lambda_i \leq C, i = 1..N$$

Here $\lambda_i, i = 1..N$ are the Lagrange multipliers. The function $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is called a kernel function. In SVM literature, there are many forms of the kernel function. If the probability density functions of the feature vectors in both classes are known, there is a possibility of defining natural kernels derived from these distributions [23].

3. Results

Out of fifteen samples of speech collected, twelve samples (80%) were used for training and remaining three samples (20%) for testing. The percentage of accuracy for test data1 and test data2 is computed. The confusion matrix is used for analysis.

It gives the following values: Percentage of Repetition recognized as Repetition, Repetition as Non-repetition, Non-repetition as Repetition and Non-repetition as Non-repetition.

i) Test data1

$$\text{Percentage Confusion Matrix} = \begin{pmatrix} 95.4 & 4.5 \\ 14.6 & 85.3 \end{pmatrix}$$

$$\text{Classification Accuracy Per class} = (95.4 \quad 85.3)$$

$$\text{Overall Classification Accuracy} = (90.35)$$

ii) Test data2

$$\text{Percentage Confusion Matrix} = \begin{pmatrix} 100 & 0 \\ 3.4 & 96.6 \end{pmatrix}$$

$$\text{Classification Accuracy Per class} = (100 \quad 96.6)$$

$$\text{Overall Classification Accuracy} = (98.3)$$

The Figure 3 and 4 shows the result of test data 1 and test data 2. The Percentage of accuracy for two test data is listed in the Table 1.

The syllables per minute (SPM) and percent disfluency (PD) were calculated using following formula:

$$SPM = \frac{\text{Total number of syllables read}}{\text{Total time in seconds}} \times 60$$

$$PD = \frac{\text{Total number of disfluent syllable}}{\text{Total number of syllable}} \times 100$$

Table 1: Percentage of accuracy for test data

Feature Extraction algorithm	Test data1	Test data2	Average accuracy
MFCC	90.35%	98.35%	94.35%

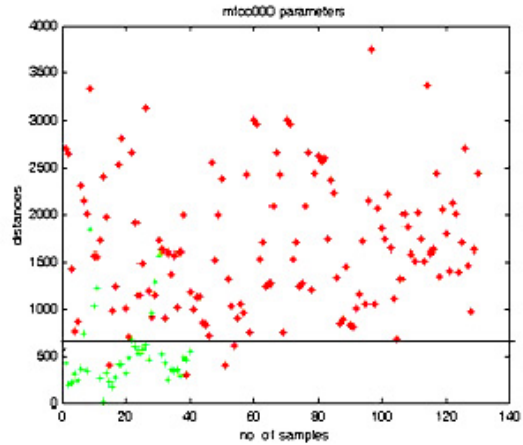


Figure 3: Result of test data 1

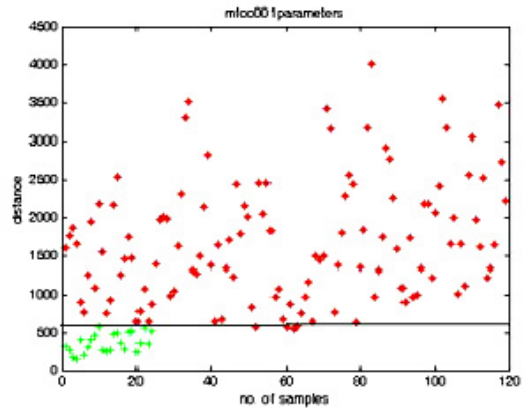


Figure 4: Result of test data 2

The results are tabled in Table 2 for two testing data.

Table 2: Percent Disfluency (PD)

Parameters	Test data1	Test data2
No. of syllable	171	147
Time in Secs	68.4	62.4
Fluent syllable	130	121
Non-fluent syllable	41	26
SPM	150	141
PD (%)	23.97	17.68

The Table 2 helps the speech-language pathologist to assess the client and also improves interjudge agreement about stuttered events [20].

4. Conclusion

In this paper a new approach for automatic detection of syllable Repetition is presented for objective assessment of stuttered disfluencies. We discussed the different steps involved in finding the number of repetitions from the speech samples using MFCC feature extraction algorithm. When compared to the previous work which is implemented using ANN [10, 11], has the result of 78.01% and with HMM [14] 80% and also with Perceptron classifier [15] 83%, our present work which uses SVM performs better with average result of 94.35%. Other Features extraction



Algorithm like fused MFCC and IMFCC may be taken as future work [21]

5. Acknowledgements

We would like to thank *L&T* and *itie Knowledge Solutions* for providing technical support and timely guidance.

6. References

- [1] M. Adams, "Voice onsets and segment duration Of normal speakers and beginning Stutterers," *Journal of Fluency Disorders*, vol. 6, pp. 133- 140, 1987.
- [2] K. R.Aida-Zade, C. Ardil and S. S. Rustamov, "Investigation of combined use of MFCC and LPC Features in Speech Recognition Systems," *International Journal of Signal Processing*, vol. 3, no.2, pp. 105-111.
- [3] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol.2, pp. 121-167, 1998.
- [4] E.G.Conture, "Stuttering" Englewood cliffs, New Jersey: Prentice-Hall, 2nd edition, 1990.
- [5] Dalouglas O' Shaughnessy, "Speech Communication," *Human and Machine*, Universities press, 2nd edition, 2001.
- [6] W.Johnson et al., "The onset of stuttering, minneapolis :University of Minnesata press," 1959.
- [7] D. Kully and E .Boerg, "An investigation of inter_clinic agreement in the identification of fluent and stuttered syllables," *Journal of Fluency Disorders*, vol.13, pp. 309-318, 1988.
- [8] E. Keogh, "Exact indexing of dynamic time warping," In *VLDB*, pp. 406-417, Hong Kong, China, 2002.
- [9] Neeta Awasthy, J.P.Saini and D.S.Chauhan, "Spectral Analysis of Speech: A new Technique," *International Journal of Signal Processing*, vol. 2, no.1, pp. 19-29, 2006.
- [10] Peter Howell, Stevie Sackin and Kazan Glen, "Development of a Two-stage procedure for the Automatic Recognition of Dysfluencies in the speech of children who stutter: I. Psychometric Procedure Appropriate for Selection of Training Material for Lexical Dysfluency Classifiers," *JSLHR*, vol.40, pp. 1073-1084, October 1997.
- [11] Peter Howell, Stevie Sackin, and Kazan Glen, "Development of a Two-stage procedure for the Automatic Recognition of Dysfluencies in the speech of children who stutter: II. ANN Recognition of Repetitions and Prolongations with supplied word segment markers," *JSLHR*, vol.40, pp. 1085-1096, October 1997.
- [12] Peter Howell and Louise Vause, "Acoustic analysis and perception of vowels in stuttered speech", *Journal of Acoustic Society of America*, vol.79, no.5, pp.1571-1579, May 1986.
- [13] K. M. Ravikumar and R. Rajagopal, "Altered Auditory Feedback Systems for Adult Stutter," *Proceedings of the Sonata International Conference on Computer Communication and Control*, pp. 193-196, November 2006.
- [14] K. M. Ravikumar, Sachin Kudva, R. Rajagopal and 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, pp. 514-519, December 2008.
- [15] 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". In *Proceedings of World Academy Science, Engineering and Technology*, vol.36, Bangkok, Thailand, pp. 270-273, October 2008.
- [16] L. Rabiner and B.H. Juang, "Fundamental of speech recognition," PTR Prentice Hall, Englewood Cliffs, New Jersey, 1993.
- [17] Radounae Iqdour and Abdelouhab Zeroual, "The Multi-layered Perceptrons Neural Networks for the Prediction of Daily Solar Radiation," *Internatioal Journal of Signal Processing*, vol 3, no.1, pp. 24-29, 2007.
- [18] A. Reda, El-Khoribi, "Support Vector Machine Training of HMT Models for Land Cover Image Classification," *ICGST-GVIP*, vol.8, issue 4, pp. 7-11, December 2008.
- [19] W. Reichl and G. Ruske, "Syllable segmentation of continuous speech with Artificial Neural Networks," In *Processing of Eurospeech*, Berlin, vol.3, pp. 1771-1774, 1993.
- [20] V.V. Sairam, "Assessment of fluency in Adult". In *Proceedings of the National Workshop on Assessment and Management of Fluency Disorders*, Mysore, India, pp. 11-26, October 2007.
- [21] Sandipan Chakroborthy and Goutam Saha," Improved Text-Independent Speaker Identification Using Fused MFCC & IMFCC Feature Sets Based on Gaussian Filter," *International Journal of Signal Processing*, vol. 5, no.1, pp. 11-19, 2009.
- [22] H. Silverman and D. Morgan, "The application of dynamic programming to connected speech segmentation," *IEEE ASSP Mag*.7, no.3, pp. 7-25, 1990.
- [23] A. Sloin and D. Burshtein, "Support Vector Machine Training for Improved Hidden Markov Modelling," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, January 2008.
- [24] V. N. Vapnik, "Statistical Learning Theory, New York: Wiley," 1998.
- [25] E.Yairi and B.Lewis, "Disfluencies at the onset of stuttering," *Journal of speech & Hearing Research*, vol.27, pp. 154-159, 1984.





Ravikumar.K.M, Assis-tant Professor, Depart-ment Of Electronics and Communication Engg., Ghousia college of Engg., Ramanagara has completed his M.Tech from SJCE, Mysore in the field of Biomedical Instrumentation in

the year 2002 and he is pursuing his PhD program under VTU, Belgaum. He is in the field of teaching from past 12 years and he has published four papers in the International Conference and one in International journal related to his research area. His field of interest includes Digital Signal Processing, Speech Signal Processing and Communication System.



Dr.R.Rajagopal, Str- ategic Electronics C- enter, L&T, Bengaluru, obtained his PhD degree in 1992 from the Bharathidasan Univer- sity, Tiruchirappalli with his research thesis in the area of Array Signal Processing for Passive Sonar. He

had his M.E. degree in Communication Systems from Bharathidasan University in 1985 and his B.E.(Hons) degree from Madras University in 1982.He has worked in the ECE department of Regional Engineering College, Tiruchirappali from 1982 to May 1998. During this period, he carried out many sponsored research projects for N.P.O.L., Cochin,

N.S.T.L., Visakhapatnam and D.O.E., Government of India in the areas of Sonar System Modelling and Simulation, Underwater Propagation Modelling, Beamforming algorithm & software development and Tracking algorithm development. He joined Central Research Laboratory (CRL), Bharat Electronics in June 1998 and served as Head, Radar Signal Processing Group till May 2006. He joined L&T in May 2006 and is Head of Technology Development in Strategic Electronics Center at Bengaluru. His current focus areas are Military Communication, Aviation and UAVs. He has more than 70 publications in the proceedings of International Conferences and International /National Journals. He has served in the Technical Program committees of many International Conferences like IRSI, FUSION 2000, etc.



Dr.H.C.Nagaraj, Princi-pal, Nitte Meenakshi institute of Technology, Bengaluru completed his PhD from IIT madras, Chennai in the year 2000 and M.E in communica- tion system from P.S.G College of Technology, coimbatore in 1984.He is in

the field of teaching from past 26 years and published 24 papers in the National/International conferences and journals. His area of interest are Digital Signal Processing, Image processing, Digital Communication, Mobile Communication and Biomedical Engineering

