Determination Of Disfluencies Associated In Stuttered Speech Using MFCC Feature Extraction

[1]Raghavendra M, [2]P Rajeswari

[1]M.Tech-Biomedical Signal Processing & Instrumentation, Dept. of Instrumentation Technology, SJCE, Mysuru, India
[2]Assistant Professor, Dept. of Instrumentation Technology, SJCE, Mysuru, India

***Abstract* – Stuttering also known as stammering is fluency disorder in which it affects the flow of speech, an involuntary repetitions, prolongation of sounds, syllables, phrase or words, and involuntary silent pause or blocks in communication. This involuntary speech disorder involves frequent and significant problems with the normal fluency and flow of speech. The number of disfluencies present in a speech is an important factor in determining in stuttering. The objective of this paper is to determine the disfluencies in speech which helps speech language pathologist to monitor the stuttering condition and its event. The existing methods uses Artificial neural network, Hidden Markov Model etc. In this method it involves novel approach to detect the disfluencies. The proposed method comprises of segmentation, feature extraction, SVM classifier. Feature extraction is implemented using well known Mel frequency Cepstrum coefficient.**

Keywords : Stutter, Disfluency, Syllable, SVM, MFCC.

I. Introduction

Stuttering is a communication disorder involving disruptions, or disfluencies in a person's speech. The word stuttering can be used to refer either to the specific speech disfluencies that are commonly produced by people who stutter or to the overall communication difficulty that people who stutter may experience. The term stuttering is most commonly associated with involuntary sound repetition, but it also encompasses the abnormal hesitation or pausing before speech, referred to by people who stutter as blocks, and the prolongation of certain sounds, usually vowels or semivowels.

Stuttered speech often includes repetitions of words or parts of words, as well as prolongations of speech sounds. These disfluencies occur more often in persons who stutter than they do in the general population. Some people who stutter appear very tense or "out of breath" when talking. Speech may become completely stopped or blocked. Blocked is when the mouth is positioned to say a sound, sometimes for several seconds, with little or no sound forthcoming. After some effort, the person may complete the word. Interjections such as "um" or "like" can occur, as well, particularly when they contain repeated ("u- um- um") or prolonged ("uuuum") speech sounds or when they are used intentionally to delay the initiation of a word the speaker expects to "get stuck on."

Some examples of stuttering include:

- Repetition: a single syllable, word, sound or phrase is repeated (for example: on—on—on a chair) or a part of a word which is still a full syllable such as "un—un—under the..." and "o—o—open".
- Prolongation : Holding onto a sound for an extended period of time.("lllllike this").
- Interjection : Speech sounds are when they are used intentionally to delay the initiation of a word the speaker expects to "get stuck on." An example of this is: "I'll meet you – *um um you know like* – around six o'clock." The person expects to have difficulty smoothly joining the word "you" with the word "around." In response to the anticipated difficulty, he produces several interjections until he is able to say the word "around" smoothly

The Standard English passage of stuttered speech is obtained from the UCLASS archive [9]. It consists of 3 types of recordings: monologs, readings, and conversation. There are 43 different speakers contributing 107 reading recordings. The samples are chosen to cover a broad range of both age and stuttering rate. Most of the reading samples do not have a text script, and hence only the reading samples with text scripts are chosen for our investigation. In this study, disfluencies like prolongation and repetition etc. are investigated. These types of disfluencies can be detected easily in monosyllabic words [5]. Each of the samples consists of more than 300 words and syllable. Disfluencies are identified and segmented automatically.

II. METHODOLOGY

In order to carry out the process, initially, stuttered speech data are collected. The input speech is subjected to pre-emphasis where low frequency components are boosted to higher frequency for easy analysis then features are extracted using Mel frequency cepstral coefficient[5]. Extracted features are used to predict the presence of fluency and non fluency depending upon certain statistical features. After obtaining these features, they are fed to the classifiers which can automatically label the speech in to fluent and non-fluent syllable. By this help it is easy to evaluate the disfluencies present in the stuttered speech.

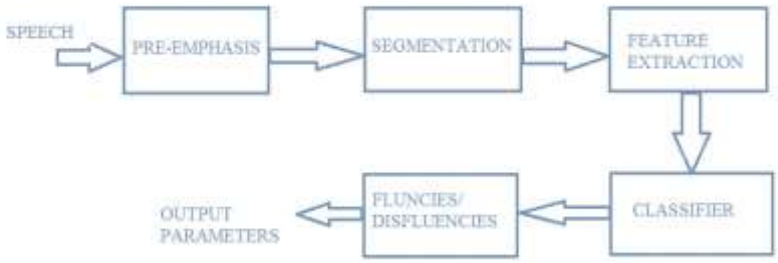

Fig.1 Block diagram of speech disfluencies determination

i. Segmentation: Syllable segmentation refers to the ability to identify the components of a word, phrase, or sentence, identifying how many syllables are in a word or phrase. Speech segmentation is the process of dividing the continuous speech into basic units having finest boundaries. It is an important step in speech recognition [8]. It also plays an important role in certain applications. It is a new method to automatically segment a continuous speech signal into syllable-like segments [6]. The technique for segmentation is based on processing the short-term energy function of the continuous speech signal. The short-term energy function is a positive function and can therefore be processed in a manner similar to that of the magnitude spectrum. In this paper, we employ a method, based on group delay processing of the magnitude spectrum to determine segment boundaries in the speech signal.

ii. Feature extraction: The time domain waveform of a speech signal carries all of the auditory information. From the phonological point of view, very little can be said on the basis of the waveform itself. However, past research in mathematics, acoustics, and speech technology have provided many methods for converting data which can be considered as information if interpreted correctly. In order to find some statistically relevant information from incoming data, it is important to have mechanisms for reducing the information of each segment in the audio signal into a relatively small number of parameters, or features. These features should describe each segment in such a characteristic way that other similar segments can be grouped together by comparing their features [16]

The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Campestral Coefficients (MFCC)[5]. MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale. MFCCs being considered as frequency domain features are much accurate than time domain features [9],

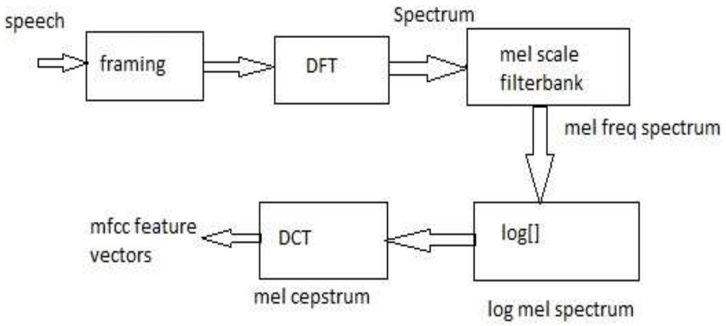

Fig.2 Blocks involved in mfcc feature extraction

Speech signal is initially passed to pre-emphasis which is carried out before segmentation. Pre-emphasis is a very simple signal processing method which increases the amplitude of high frequency bands and decreases the amplitudes of lower bands. In simple form it can be implemented as

$$y(n) = x(n) - a \cdot x(n-1)$$

Where *x(n)* is value of input signal at discrete time step *n*, *y(n)* is value of output signal at discrete time step *n*. a is constant (a=0.95).

The input speech signal is framed into 15~20ms with overlap of 50% of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, zero padding is done to the nearest length of power of two. Overlapping is used to produce continuity within frames.
The extraction of the signal takes place by multiplying the value of the signal at time *n*, *s[n]*, with the value of window at time *n*, *w[n]*.

$$Y[n]=w[n]s[n]$$

A commonly used algorithm for computing the DFT is Fast Fourier Transform or FFT. This implementation of the DFT is very efficient, but only works for values of N which are powers of two. The FFT of each frame and obtain its magnitude. The FFT is a computationally efficient algorithm

of the Discrete Fourier Transform (DFT). If the length of the FFT, is a power of two ($K = 2^n$), a faster algorithm can be used, so a zero-padding to the nearest power of two within speech frame length is performed.
The Mel-scale equivalent value for frequency f expressed in Hz is:

$$Mel(f) = 2595\log 10\left[1 + \frac{f}{700}\right]$$

Compute the Mel-spaced filter bank. This is a set of triangular filters. During MFCC computation, this intuition is implemented by creating a bank of filers which collect energy from each frequency band, with 10 filters spaced linearly below 1000 Hz, and the remaining filters spread logarithmically above 1000 Hz.

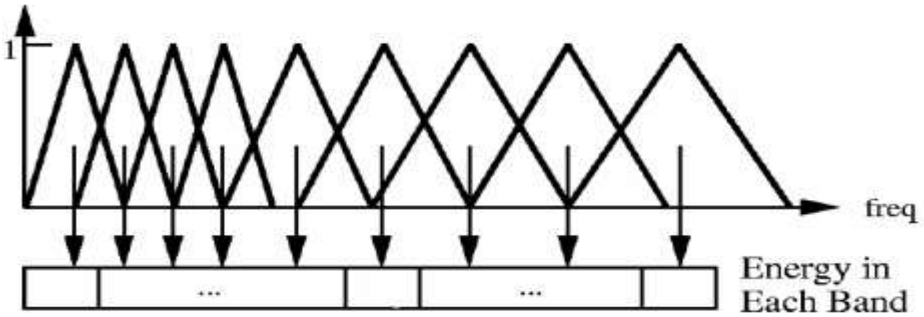

Fig.3 Mel filter bank

Once we have the filter bank energies, we take the logarithm of them. The logarithm allows to use cepstral mean subtraction, which is a channel normalization technique. The final step is to compute the DCT of the log filter bank energies. There are 2 main reasons this is performed. Because our filter banks are all overlapping, the filter bank energies are quite correlated with each other. The DCT decorrelates the energies which mean diagonal covariance matrices can be used to model the features. This feature extracted vectors are subjected to support vector machine.

iii. Support vector machine**:** SVM are classifiers which discriminate data points of two categories.Each data point is represented by a vector of n-dimension. Every data points belong to one of the classes. Hyperplane is plotted between the two classes. SVM has two steps, training and testing. Initially more number of stuttered speech data are trained. To achieve maximum separation between the two classes, SVM picks the hyperplane which has the largest margin. The margin is the summation of the shortest distance from the separating hyperplane to the nearest data point of both categories. Such a hyperplane is likely to generalize better to classify testing speech data.

If we take y_i as the class of data point x_i then the classifier function is given by

$$f(x_i) = \text{sign}(w^T . x_i + b)$$

Functional margin of $\mathbf{x}_i$ is given by

$$f(x_i) = y_i (w^T x_i + b)$$

Given training data (x_i, y_i) for i = 1. . . N, with $x_i \in R^d$ and $y_i \in \{-1, 1\}$, learn a classifier f(x) such that

$$f(x_i) \quad \begin{array}{l} \geq 0 \ y_i = +1 \\ < 0 \ y_i = -1 \end{array}$$

i.e. $y_i f(x_i) > 0$ for a correct classification. Based on the sign fluent and non-fluent syllables are separated [17].
The margin of separation is given by

$$Width = 2/||w||$$

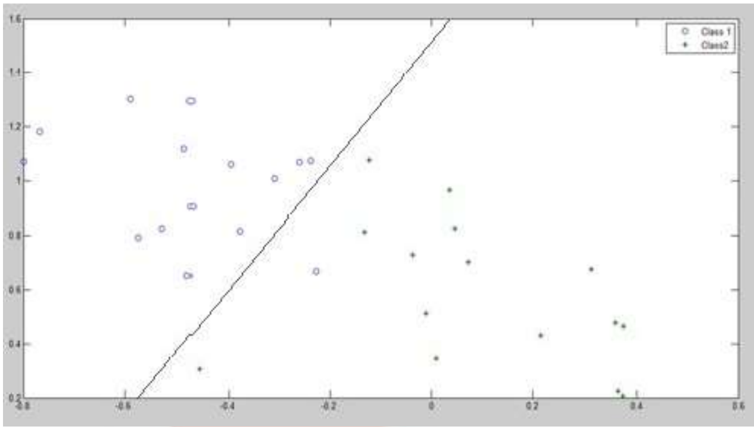

Fig.4 Two classes separated by hyperplane.

In this work, to get better classification results, disfluency obtained by support vector machine of training features. Euclidean distance is employed to measure the proximity between two feature vectors.
In the recognition phase an unknown disfluency sample, represented by a sequence of feature vectors{x1,x2,…xT} , is compared with the speech database. Compute parameters using Euclidean distance function [11]. Repetition, prolongation and interjection are chosen as disfluency based on lowest distortion. The Euclidean distance is defined by:

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

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

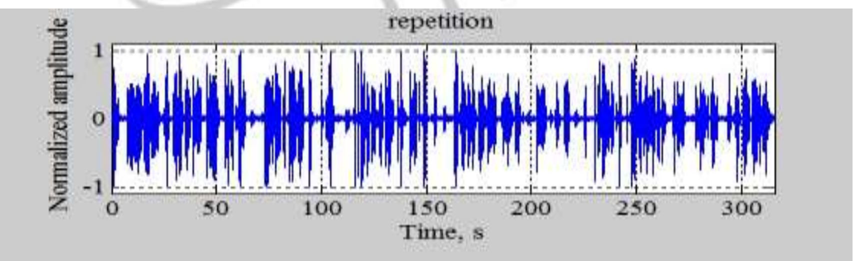

Fig.5 Repetition

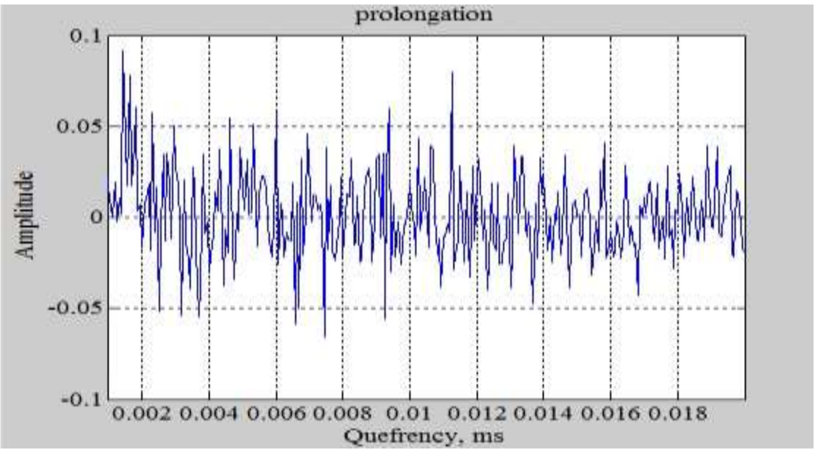

Fig.6 Prolongation

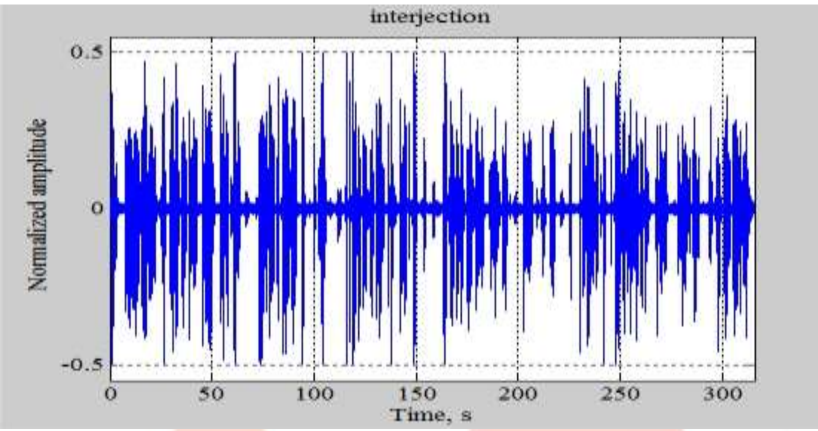

Fig.7 Interjection

III. RESULTS

Speech data of 20 samples from UCLASS were trained and 5 test data are used for testing. Based on the analysis disfluencies are obtained such as repetition, prolongation and interjection. This methodology was implemented using MATLAB software. The results obtained in this method are tabulated in Table-I as follows,

Syllable per minute and percentage disfluencies are crucial parameter in stuttered speech to decide the severity of stutter. According to Language Severity Rating Scale if the fluency is more 11% for a 100 word then it is said to be severe below that is mild and moderate condition. SPM and PD can be calculated by using formula [2].

$$Syllable\ per\ minute(SPM) = \frac{Total\ number\ of\ syllable}{Time\ in\ seconds} X\ 60$$

$$Percentage\ disfluency(PD) = \frac{Total\ number\ of\ disfluent\ syllable}{Total\ number\ of\ syllable} X\ 100$$

Table I

Test data	Number of Syllable	Total time in seconds	SPM	PD (%)	Condition
1	194	130	89.5	42%	Severe
2	163	120	81.5	38%	Severe
3	112	68	98.8	21%	Mild

Table I shows the stuttered events of patient. This event helps Speech language pathologist to assess and also improve interjudge agreements about stuttered events.

IV. CONCLUSION

In this paper a new approach to determine the stuttered disfluencies such as repetition, prolongation and interjection are presented. An automatic segmentation of syllable is implemented to obtain syllable like units to replace manual technique. The feature extraction was performed using MFCC algorithm. Stuttered database were trained and tested using classifier called SVM.

A lot more other techniques such as LDA and Perceptron method can also be used instead of SVM but accuracy level varies. However SVM classification has high accuracy. The number of Training data can be increased and checked with testing data to improve the accuracy. To find out repetition, prolongation, interjection. DTW, Correlation technique are alternative methods to detect and hence in future, enhancements can be made on this [1] [15].

REFERENCES

[1]. K. Ravi Kumar, B. Reddy, R. Rajagopal, and H. Nagaraj, "Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies," in Proceedings of World Academy Science, Engineering and Technology, 2008, pp. 270-273.

[2]. R.Rajagopal, K N Ravikumar and H.C.Nagaraj, "An Approach for Objective Assessment of Stuttered Speech Using MFCC Features," ICGST International Journal on Digital Signal Processing, DSP, vol. 9, pp. 19-24, 2009.

[3]. Lim Sin Chee,Ooi Chia Ai, Sazali Yaacob "Overview of Automatic Stuttering Recognition System" Proceedings of the International Conference on Man-Machine Systems(ICoMMS),11-13 October 2009, Batu Ferringhi, Penang, MALAYSIA

[4] P. Howell, S. Sackin, K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysuencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dys
uency classiers", Journal of Speech, Language, and Hearing Research, Vol. 40, pp. 1073-1084, 1997.

[5] P. Howell, S. Sackin, K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfuencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers", Journal of Speech, Language, and Hearing Research, Vol. 40, pp. 1085-1096, 1997.

[6]. Aimee Sidavi & Renee Fabus "A Review of Stuttering Intervention Approaches for Preschool-Age and Elementary School-Age Children" CICSD Brooklyn College, Brooklyn, NY

[7]. Chong Yen Fook, Hariharan, Lim Sin Chee "Comparision of speech parameterization techniques for the classification of speech disfluencies". Turk J of Elec & Comp Sci(2013) TUBITAK.

[8]. Okko Rasanen, Gabriel Doyle, Michael C "Unsupervised word discovery from speech using automatic segmentation into syllable like units" Aalto university, Stanford university, Language & Cognition Lab, California.

[9]. Paul Mermelstein "Automatic segmentation of speech into syllabic units" Haskins laboratories, New Haven, Connecticut 06510. 14 April 1975.

[10]. T.Jayasankar, Dr R.Thangarajan, Dr.J.Arputha Vijaya "Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis". International Journal of Computer Applications (0975 – 8887) Volume 25– No.1, July 2011

[11] P. Howell, S. Davis, J. Bartrip, "The UCLASS archive of stuttered speech", Journal of Speech, Language, and Hearing Research, Vol. 52, pp. 556-569, 2009.

[12]. Erfan Loweimi, Seyed Mohammad Ahadi, Thomas Drugman, and Samira Loveymi "On the Importance of Pre-emphasis and Window Shape in Phase-based Speech Recognition" TCTS Lab, University of Mons, 31, Boulevard Dolez, B7000 Mons, Belgium,Computer Engineering Department, Buali Sina University, Hamedan, Iran

[13]. T.Nagarajan, Hema A. Murthy and Rajesh M. Hegde "Segmentation of speech into syllable-like units" EUROSPEECH 2003

[14]. P. Mahesha and D.S. Vinod, " Vector Quantization and MFCC based Classification of Dysfluencies in Stuttered Speech" Bonfring International Journal of Man Machine Interface, Vol. 2, No. 3, September 2012

[15]. Andrzej Czyzewski, Andrzej Kaczmarek1 and Bozena Kostek "Intelligent processing of stuttered speech" Gdansk University of Technology, Sound & Vision Engineering Dept., Gdansk, Poland and Institute of Physiology & Pathology of Hearing, Warsaw, Poland.

[16] K.M. Ravikumar, S. Ganesan, "Comparison of multidimensional MFCC feature vectors for objective assessment of stuttered disfluencies", International Journal of Advanced Networking and Applications, Vol. 2, pp. 854{860, 2011.

[17]. P.Mahesha and D.S.Vinod "An Approach For Classification Of Dysfluent And Fluent Speech Using K-Nn And SVM" International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.6, December 2012.