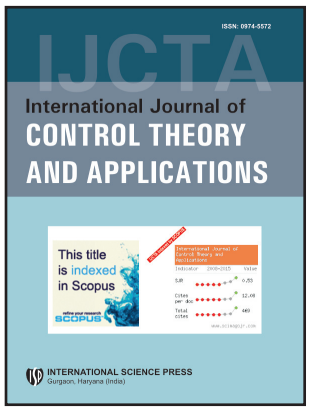

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Automatic Speech Recognition System for Stuttering Disabled Persons

Arya A Surya[a] and Surekha Mariam Varghese[b]

[a-b]Department of Computer Science and Engineering, M.A College of Engineering, Kothamangalam, Kerala, India. Email: [a]aryasurya92@gmail.com; [b]surekh.var@gmail.com

Abstract: About 1% of population suffers from stuttering. Stuttering or stammering is a speech disorder. It affects the fluency of speech. Stuttered speech contains the disfluencies, characterized by prolonged sounds, repetitions, incomplete phrases so on. In present world Automatic Speech Recognition (ASR) find its relevance in many applications. But Automatic Speech Recognition systems developed are not efficient in recognizing stammered speech. This paper proposes three methods i.e., using trained model, by removing prolongations/repetitions and by converting to text for recognizing stuttered speech.

Keywords: Stuttering; Disfluencies; Automatic Speech Recognition; Feature Extraction; Classification; MFCCs; SVM; Neural Networks.

1. INTRODUCTION

Speech is accustomed and universally used form of communication form of communication. It is very effective form and is researchers [10] are interested in this field. Automatic speech recognition(ASR) [18] can be viewed as a technology of machine driven reproduction of uttered language into a decipherable text in real time i.e., it allows computers to understand the words that is said by a person. In this era of machines, having a machine to identify spoken speech has steered speech researchers and analysts for more than 5 decades. The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950s and ever since it find its many appliance [19] in our day to day life. ASR gain its attention due to its convenience for the deaf and hard of hearing, Cost cut down through automation and Searchable text capability.

But not all the humans are blessed with proper speaking capability. Around one percentage of total population suffers from speech disfluencies like dysarthria, apraxia, stuttering, cluttering, lisping, whispering, mumbling and many more. Stuttering otherwise known as stammering is one of the serious disorder found in speech pathology. Stammering is most commonly found to affect males than females.

Stuttering is a speech disorder in which normal flow of speech is cluttered by occurrences of dysfluencies. The most prevalent types of stuttering are:

(i) Interjections

(ii) Revisions

(iii) Incomplete phrases

(iv) Repetition

(v) Prolonged sounds

(vi) Broken words

The causes of stuttering can be language problem like anxiety, genetical, heredity, stress and nervousness, brain disorders etc.

2. RELATED WORK

Most In last few decades researchers have been focusing on developing systems for assessing stuttered speech but not many has been working on recognizing the stuttered speech. And prevalent current speech recognition systems[12], [13], [6] use hidden Markov models (HMMs), Gaussian mixture models (GMMs), artificial neural networks (ANNs) [9] etc.

A research [20] presented by Geetha, Pratibha, Ashok, and Ravindra in year 2000, on classification of childhood dysfluencies used ANNs. They achieved an accuracy of 92% in predicying normal, non-fluency and stuttering. Similarly Czyzewski, aczmarek, and Kostek (2003) and Prakash (2003) presented papers on classifying dysfluent speech using ANNs. In 2007, Wis´niewski, Kuniszyk-Joz´kowiak, Smoka, & Suszyn´ proposed an automatic detection system mainly focussed on recognition of prolongations of phonemes with HMM as classification method and they parameterized samples using MFCCs. They achieved approximately 80% accuracy. in 2009,Ravikumar et al. [3] implemented technique for the automatic detection method using SVM to classify between fluent and dysfluent speech. The system yielded 94.35% accuracy.

The speech recognition systems involves mainly two stages of speech processing, feature extraction and feature classification. Feature extraction [7] can be defined as a process of Converting the original speech signal into parametric representation for further investigation and processing by retaining useful information from speech signal while abandoning unwanted signal like noise. To recognize these extracted feature vectors different classifiers [6] can be used. There are different speech feature extraction techniques such as

(a) Linear Predictive coding

(b) Mel-frequency cepstrum (MFFCs)

(c) Fast Fourier Transform (FFT)

(d) Perceptual Linear Prediction Coefficients (PLPC)

(e) Discrete wavelets (DWs)

The commonly used classification methods [6,7] for speech include

(a) Artificial Neural Networks

(b) Hidden Markov Model

(c) Support vector machines

(d) Gaussian mixture model (GMM)

(e) K-Nearest Neighbor (K-NN)

(f) Linear Discriminant Analysis (LDA)

3. PROPOSED WORK

Three methods are proposed for the recognition of the stuttered speech.

(a) Supervised model for stuttered speech recognition

(b) Stuttered speech recognition by stuttering pruning

(c) Automated text-to-speech based stuttered speech recognition

A. Supervised Model for Stuttered Speech Recognition

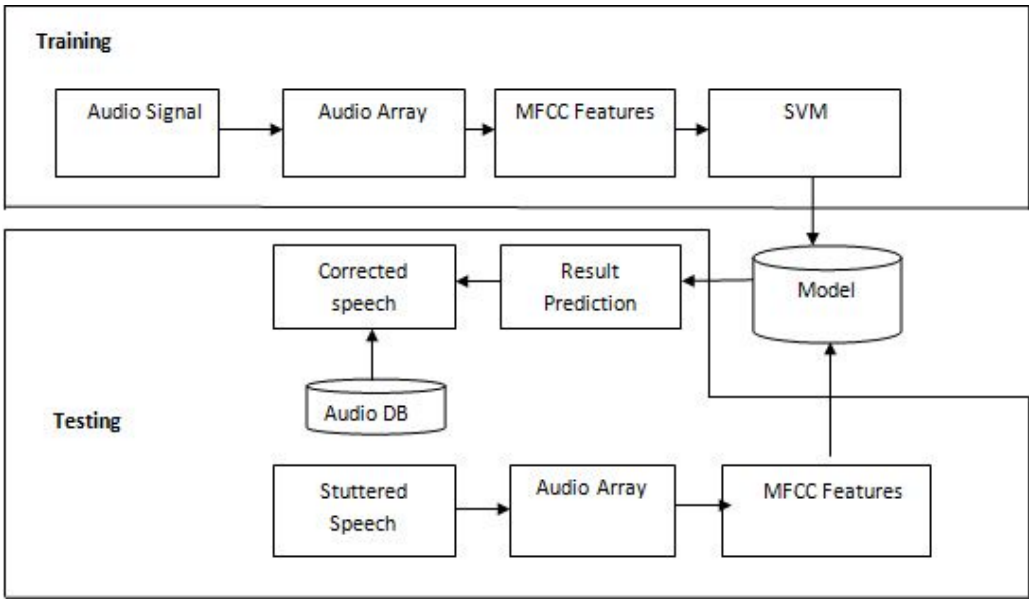

Figure 1: Block Diagram of supervised model for stuttered speech recognition

The first method is implemented using a model. It involves two phases: Training and Testing. N audio signals are converted to audio array. The MFCC [14] features are extracted from audio. The unique features of human voice can be extracted by using Mel Frequency Cepstral Coefficient (MFCC) [1], [2], [4] and this MFCC also represents the short term power spectrum of human voice. To calculate the coefficients which represent the frequency Cepstral, MFCC [3] is used and these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. The frequency bands are equally spaced in Mel scale and because of this it approximates the human voice more accurate. The steps involved in calculating MFCCs is as follows:

1. *Frame blocking:* The pre-emphasized signal is blocked into frames of P samples with adjacent frames are separated by Q (Q < P). First frame consists of first P samples. Second frame begins Q samples after first frame and overlaps by P-Q samples and so on.

2. *Windowing:* Windowing is the process of multiplying of speech waveform with a rectangular pulse whose width equal to the frame length resulting in each frame. Thus, it will reduce significant high frequency noise caused due to abrupt changes from zero to signal and from signal to zero present the

beginning and ending of the frame. In order to reduce the edge effect, each frame is multiplies by N sample points hamming window. The concept here is to minimize the spectral distortion and the signal discontinuities. The mathematical expression of the hamming window is:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\Pi n}{N-1}\right) \tag{1}$$

If window is defined as $w(n)$, $0 \le n \le N-1$, where N is the number of samples in each frame, and then the result of windowed signal is

$$y(n) = x(n)w(n) \tag{2}$$

3. *FFT:* Fast Fourier transform is applied to windowed signal to convert each frame of N samples from the time domain into the frequency domain. After the FFT block, the spectrum of each frame is filtered by a set of filters, and the power of each band is calculated.

4. *Mel-Scale Filter:* A filter bank which spaced uniformly on the Mel-scale is used to simulate the subjective spectrum. Mel-Scale is defined as logarithmic scale of frequency based on human pitch perception. Mel-scale is linear frequency spacing below 1 khz and logarithmic spacing above 1 khz. The mapping from linear frequency to Mel-frequency is

$$\text{mel}(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

Finally, the log Mel spectrum is converted into time. The output is called as Mel Frequency Cepstrum Coefficients (MFCC). The Mel Frequency Cepstrum Coefficients are real numbers and can be converted into time domain using Discrete Cosine Transform (DCT).

During the classification stage the extracted features used to train support vector machine (SVM) [15,16,17]. SVM has two stages training and testing. SVM is a classifier which performs classification methods by constructing hyper planes in a multidimensional space. Since automatic speech recognition is a multiclass problem, SVM can be extended to multiclass classification, though SVM is basically a binary non linear classifier. During testing SVM classifies the stuttered input to correct word.

Algorithm 1: Pseudo code for Stuttered speech recognition

INPUT: Stuttered speech data, D

OUTPUT: Classified word, *w*

Step 1: Sn ← Each speech signal Є D

Step 2: C ← MFCC(S_n, *a*)

Step 3: *w* ← svm Method (C)

Step 4: return *w*

Procedure MFCC(S_n, a)

Step 1: $x(n) \leftarrow S_n - aS_{n-1}$

Step 2: Divide $x(n)$ into N frames with overlap of M(M < N)

Step 3: $y(n) \leftarrow w(n)x(n)$, $0 \le n \le N-1$

Step 4: $x(n) \leftarrow \text{DFT}(y(n))$, $n = 0, 1, ..., N-1$

Step 5: FBE(g) = $\log\left(\sum_0^{k/2} x(k)\text{Hmel}(k, g)\right)$, $g = 0, 1, ..., G - 1$

Step 6: $C_n \leftarrow \text{DCT}(\text{FBE}_n)$

Procedure svm Method (C)

Step 1: for each C_i

Step 2: Compute svm solution w, b for data set with imputed labels

Step 3: Compute outputs $f_i =< w, x_i >+ b$

Step 4: Set $y_i = \text{sgn}(f_i)$

Step 5: Return label based on y_i

The above algorithm gives the summary of the proposed method. In the above algorithm, G is logarithmic filter bank energy coefficient, Hmel is the mel filter and C gives cepstral coefficients. Discrete Fourier Transform (DFT) is calculated efficiently using FFT algorithm.

B. Stuttered Speech Correction System

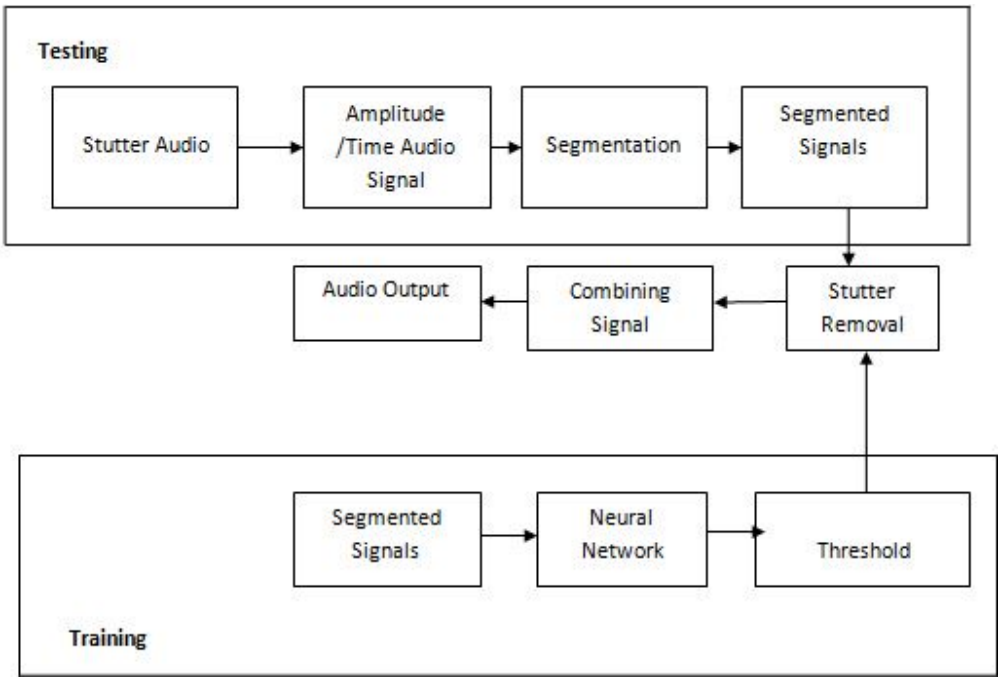

Figure 2: Block Diagram of stuttered speech recognition by stuttering pruning

The steps for the correction and recognition of stuttered speech are as follows:

1. Take a speech sample converting into amplitude/time audio signal.
2. Obtain the maximum amplitude of the speech.
3. Pass the maximum amplitude to the neural network to compute a threshold value.
4. Segment the audio samples into small frames of equal length.

5. Analyse each frame and if the max value of the frame is greater than the threshold value, copy the frame
6. Once all frames have been analysed, pass the signal to a speech recognition module.

Algorithm 2: Pseudo code for removing stuttering

INPUT: Stuttered speech sample

OUTPUT: Cleaned. wav audio file

Step 1: x ← sample s_i for each s_i € x

Step 2: AmpMax ← max(Amplitude of x)

Step 3: threshold← neuralNetwork(AmpMax)

Step 4: Divide x into small frames

Step 5: For each frame, if max(frame amplitude) > threshold

Step 6: copy frame to z

Step 7: Convert z into. wav format

Step 8: return

procedure neuralNetwork(AmpMax)

Step 1: Build neural network specifying number of input layers, hidden and output layers

Step 2: Randomly initialize the weights

Step 3: Add training set to network with AmpMax as input and threshold as output

Step 4: Update weights until network converges

Neural network[5,11] is trained using backpropogation. Trained Neural Networks will take the maximum amplitude of the speech signal as an input. It outputs the threshold for that amplitude as the output. The threshold output given by the network is used to clean the stuttered sample that could be easily recognized.

C. Automated Speech-to-text Based Stuttered Speech Recognition

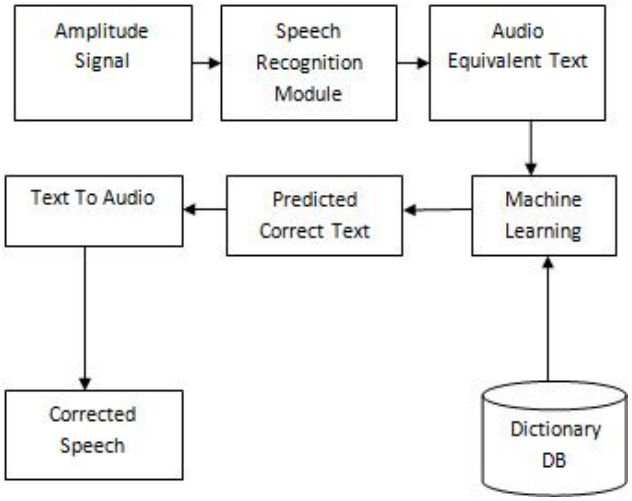

Figure 3: Block Diagram of Automated speech- to-text based stuttered speech recognition

This method converts words to equivalent texts[19]. Powerful Artificial Neural Networks are used to identify each letters in the speech. ANN is trained with intelligent guess to predict the vowels and consonants terms in the speech. This ANN analyses the inputted speech with its training experience and produces equivalent texts. The outputted text is then inputted to a dictionary and picks the most matched word. In this way all stuttered speeches are eliminated. Then the entire speech which is in the form of texts are reverted back to speech. The reversion process again uses machine learning techniques to produces the text equivalent audio for each speeches. Thus we can obtain corrected speech devoid of stammering.

Algorithm 3: Correction of stuttered speech converting to equivalent text

INPUT : Audio file

OUTPUT: corrected audio signal

Step 1: Input the audio file, F

Step 2: Input F to the RecogAnn

Step 2.1: Returns the equivalent text

Step 3: Split each words in text

Step 4:Input each words in text to DictFunc

Step 4.1: Returns corrected version of each word

Step 5: Combine the words

Step 5.1: Returns original sentence,S

Step 6: Input S into TextToSpeech

Step 7: Returns the corrected audio signal

4. RESULT AND DISCUSSION

Data samples were obtained from the University College London Archive of Stuttered Speech (UCLASS) and National Institute of Speech and Hearing (NISH). The database consists of recordings for monologs, readings and conversations of different speakers. In first and third methods i.e. recognition of speech using a trained model and recognition by converting to equivalent respectively, the Speech samples of prolongations and repetitions were segmented from the recordings manually. While in the method of removing prolongations and repetition reading of different people were used to collect the their highest speech threshold and corresponding stuttering threshold.

The first method implemented was stuttered speech recognition using a classifier model. The classifier model used was support vector machine. Several words were manually segmented from the collected dataset. The SVM model was trained using the segmented stuttered words. We acquired an accuracy 76% in classifying the words correctly. The accuracy of this method can be improved by using more training data. Limitation of this method is that only the trained words gets predicted.

Speech correction method implemented using neural network acquired less accuracy i.e. of 62% which can be improved further by using more training data as well as incorporating some other features such as energy, frequency etc of the audio input for training. Even though an average of 62% stuttering was removed, few non stuttered part got cleaned which would be considered in future work. Comparing to the previous method advantage

of this method is that the original speech of the person gets reconstructed devoid of stuttering. The Figure 4 and Figure 5 below shows the original stuttered speech and cleaned speech.

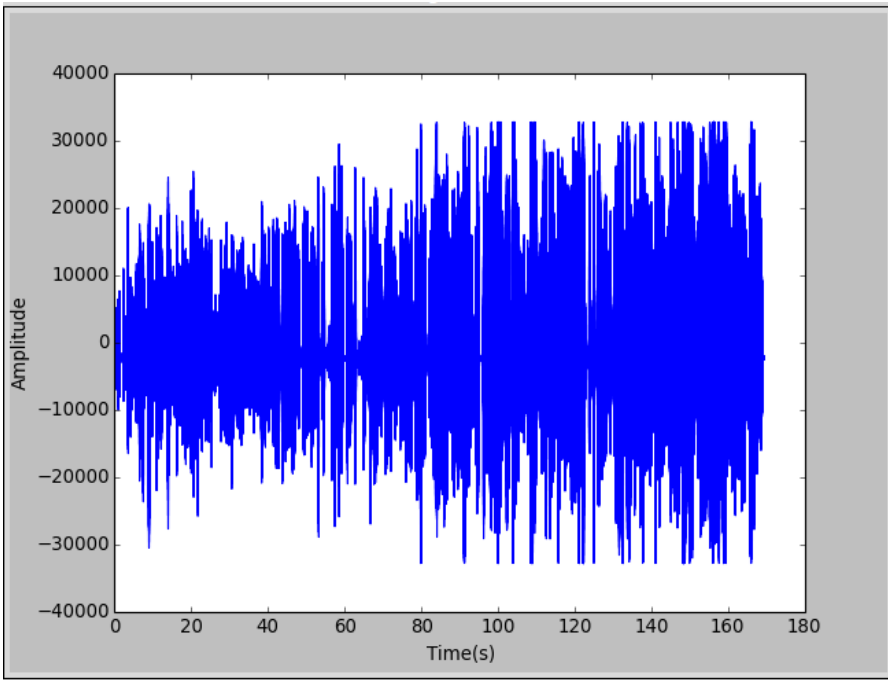

Figure 4: Stuttered speech sample

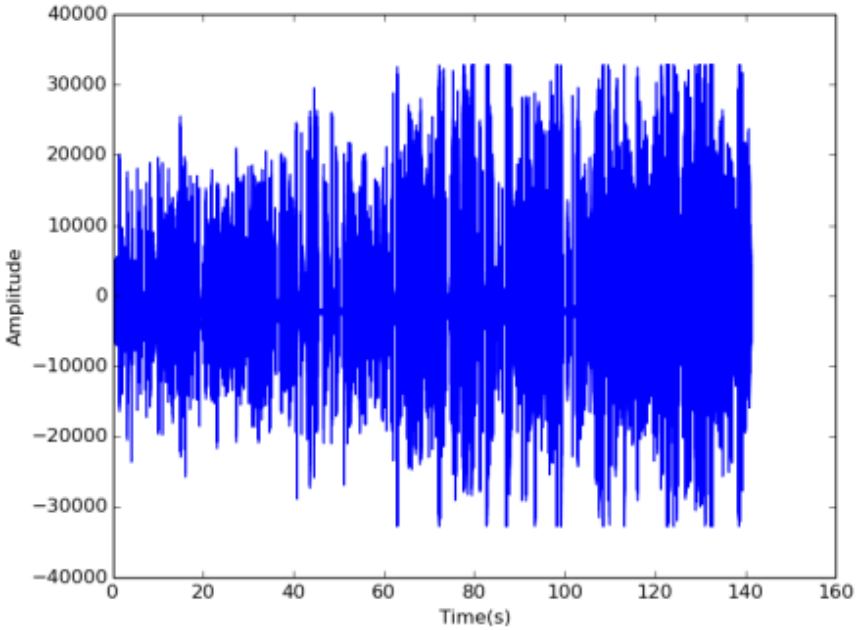

Figure 5: Cleaned speech sample

Finally the last method which tried to recognize the stuttered speech by converting it into equivalent text achieved an accuracy of 80%. Here complete sentences were able to recognize rather single trained words. Graph shown below plots the accuracy of different proposed methods. As the number of training data increases we can see a is improvement in the accuracy of the methods.

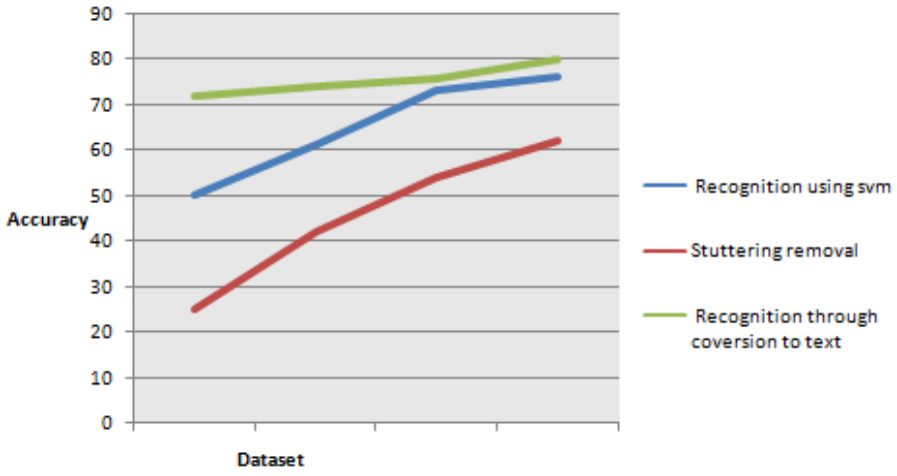

Figure 6: Accuracy of different methods

5. CONCLUSION

The increasing usage of speech recognition systems by people has led to the ease of access in their day to day lives. People use personal assistants like Apple's Siri, Microsoft Cortana or Google Now and make their lives easier, however people with speech impairments like stuttering cannot benefit from these services because these companies have catered their speech recognition algorithms to the majority of people, that is, the people without any speech disorders even when about 70 million people in the world suffer from stuttering alone. These voice recognition systems are unable to detect when people afflicted by stuttering use it because when the person starts stuttering the service thinks that the person has completed speaking and doesn't process what comes subsequently. In order to make these above mentioned applications more universal we propose 3 methods that would actually help in solving a real world problem. The methods were implemented and its accuracy were measured. The accuracy can be improved using more training data and considering other features.

6. SCOPE FOR FUTURE WORK

Our solution can be integrated with already existing speech recognition services across all platforms like PC, mobile, etc. Which would enable the affected persons to use speech recognition tools and services even with their stutter. We can further increase the accuracy and effectiveness of our technique by acquiring more data samples from affected individuals, which would result in a larger training set, thus making our methods more robust. We could also use another parameters, classification methods and features to better detect and correct the stuttered speech. Stuttering is only one of the common speech disorders, we could also implement the same with other speech impediments like lisp, etc

Acknowledgment

The Authors would like to thank Speech Language Pathologist at "National Institute of Speech and Hearing" (NISH) for providing support in collecting the data.

7. REFERENCES

[1] Lim Sin Chee, Ooi Chia Ai, M. Hariharan and Sazali Yaacob, "MFCC based Recognition of Repetitions and Prolongations in Stuttered Speech using K-NN and LDA" *Proceedings of 2009 IEEE student conference on Research and Development*

[2] M. Hariharan, Lim Sin Chee, Ooi Chia Ai and Sazali Yaacob, "Clssification of speech dyfluencies with MFCC and LPCC features"

[3] Ravikumar, K. M., Rajagopal, R., & Nagaraj, H. C. (2009). An approach for objective assessment of stuttered speech using MFCC features. *ICGST International Journal on Digital Signal Processing*, DSP, 9(1), 19–24.

[4] Lim Sin Chee, Ooi Chia Ai and Sazali Yaacob, "Overview of Automatic stuttering recognition system" International conference on Man-Machine Systems(ICoMMS) October 2009, Batu Ferringhi, penang, Malaysia

[5] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Networks used for Speech Recognition*", Journal Of Automatic Control, University Of Belgrade*, Vol. 20:1-7, 2010

[6] G. Manjula, M. Shiva Kumar, "Overview Of Analysis And Classification Of Stuttered Speech", *Proceedings of 11th IRF International Conference*, 8th May 2016, Hyderabad

[7] S. Poornima, J. Satheesh Kumar, "Feature Extraction and Signal Classification methods for Stuttering Speech Analysis ", *International Journal of Modern Computer Science and Applications* Volume No.-1, Issue No.-5, November, 2013 http://www.resindia.org

[8] Mahesha and D.S. Vinod, "Automatic Classification of Dysfluencies in Stuttered Speech using MFCC", *International Conference on Computing, Communication and Information Technology* (ICCCIT 2012), 27 - 29 June, 2012

[9] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "DNN for acoustic modelling in speech recognition", *IEEE Signal Processing Magazine* November 2012

[10] Rekha Hibare and Anup Vibhute," Feature Extraction Techniques in Speech Processing: A Survey", *International Journal of Computer Applications* (0975 – 8887)Volume 107 – No 5, December 2014

[11] V.Naveen Kumar, Y Padma Sai, C Om Prakash, "Design and Implementation of Silent Pause Stuttered Speech Recognition System", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering(IJAREEIE)*, Vol. 4, Issue 3,March 2015.

[12] I.Szczurowska, W.Kuniszyk-Jozkowiak, and E.Smolka, "The application of Kohonen and Multilayer Perceptron Networks in the speech nonfluency analysis," *Archives of Acoustics*, Vol. 31, p. 205, 2006

[13] M. Kesarkar, "Feature Extraction For Speech Recognition, " Indian Institute of Technology, Bombay 2003

[14] C.Burges,, " A tutorial on support vector machines for pattern recognition," Data Mining Knowl. Discov., Vol. 2, pp. 121-1998.

[15] Sonia Sunny, David Peter S, K Poulose Jacob, "Design of a Novel Hybrid Algorithm for Improved Speech Recognition with Support vector Machines Classifier*", International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, pp. 249-254, June 2013.

[16] P. Bhuvaneswari, J.Satheesh kumar, Support Vector Machine Technique for EEG Signals, *IJCA*, 63(13), pp. 1-5, ISSN: 0975-8887,2013

[17] M.A.Anusuya, S.K.Katti, "Speech Recognition by Machine: A Review", *International Journal of Computer Science and Information Security (IJCSIS*), Vol. 6, No. 3, 2009.

[18] Dr. Shaila D. Apte, "*Speech Processing Applications*", in Speech and Audio Processing, Chapter 3, Pages 105- 118, Wiley IndiaEdition.

[19] Prachi Khilari, Prof. Bhope V. P, "A Review On Speech To Text Conversion Methods", *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)* Vol. 4,Issue7,July2015

[20] Y.V. Geetha, K. Pratibha, R.Ashok, and S.K.Ravindra" Classification of childhood disfluencies using neural networks" *Journal of Fluency disorders*, Vol. 25, pp. 99-117,2000.