Speech Recognition and Correction of a Stuttered Speech

Ankit Dash, Nikhil Subramani, Tejas Manjunath, Vishruti Yaragarala and Shikha Tripathi

Department of Electronics and Communication Engineering
Faculty of Engineering, PES University
Bangalore, India
ankitdsh18@gmail.com, nikhilsubramani47@gmail.com, tejasmanjunath@gmail.com, vishruti.y@gmail.com, shikha@pes.edu

***Abstract*— The aim of this paper is to develop an algorithm to enhance speech recognition of a stuttered speech. Stuttering is a disorder that affects the fluency of speech by involuntary repetition, prolongation of words/syllables, or involuntary silent intervals. Current speech recognition systems fail to recognize stuttered speech. Methods to detect stutter have been reported in literature but efficient techniques for stutter correction have not been reported. This paper addresses this issue and proposes methods to detect and correct stutter within acceptable time limits. To remove prolongation(s) from the sample, amplitude thresholding through neural networks is developed. Repetitions are removed through string repetition removal algorithm using an existing Text-to-Speech (TTS) system. Thus, the output signal, void of all stutters, produces better speech recognition.**

Keywords—speech recognition, stutter, neural networks, TTS system

I. INTRODUCTION

Stuttering is a speech disorder characterized by repetition of sounds, syllables, or words; prolongation of sounds. An individual who faces this disorder knows what he or she intend to say but is unable to produce fluent speech. Millions of people, in today's world, suffer from various speech disorders like stuttering, lisp, and articulation disorder. This often renders them unable to utilize certain things that a normal person takes for granted, like speech recognition systems.

Stuttering disorder is characterized by disruptions in the production of speech sounds, called disfluencies. Disfluencies are not necessarily a problem; however, they can hinder communication when a person produces too many of them. Most people often produce brief disfluencies. For instance, some words are repeated or prolonged while others are preceded by an 'um' or 'uh'. In most cases, stuttering has an impact at least on some daily activities. The specific everyday activities that a person finds challenging to perform, vary across individuals. For example, for some people, communication difficulties happen only during specific activities, like talking on the phone, talking before large groups, utilizing everyday tools that use speech as inputs. An author claims, "Stuttering cannot be permanently cured; it may go into remission for a time, or clients can learn to shape their speech into fluent speech with the appropriate speech pathology treatment" [1].

The current speech recognition systems have good accuracy for fluent speech but fail to recognize speech with repetitions and/or long involuntary pauses. This is mainly because these systems are created to stop the identification process when a pause is encountered. Also, these systems are trained with words without any repetitions and hence, when it encounters a stuttered speech, it is unable to identify the words.

There have been methods to detect stutter from speech samples, but correction of stutter has not been addressed by many researchers. Some of them have used techniques like Artificial Neural Network (ANN), Hidden Markov Model (HMM), Support Vector Machine (SVM) and advanced Digital Signal Processing techniques to remove noise from the samples and then correct them but not with good efficiency.

This paper aims to detect as well as correct these stuttered speech samples in real time mode, and produce the corrected speech samples devoid of any stutter. An algorithm using neural networks along with a few string operations has been proposed to detect and correct the speech. This system can then be integrated with phones and laptops to help people suffering from this speech impairment to control their devices with speech, in the same way as most of the population in today's world does.

Thus the main objective of this work is to help people with speech disability use already accessible tools available to them, without worrying about their speech impairment.

In order to help people suffering from stutter, two algorithms for removing stutter from a speech signal are proposed and implemented.

The first algorithm removes prolongations of syllables or words through amplitude thresholding. This works by measuring the maximum amplitude present in the speech signal and calculating a suitable threshold value based on initial training of the system, whereby signals with an amplitude less than the threshold are removed. This method works satisfactorily since the amplitude of prolongations are usually less than that of the useful words spoken.

The second algorithm is used to remove any repeated words or phrases. Since the repeated words are of similar amplitudes, the first algorithm does not work. To eliminate such a stutter, the speech is converted to text and the designed function removes any repeated word(s) and is then converted back to speech. Implementing these two algorithms in a framework allows the system to remove all types of stutter present in any speech signal.

978-1-5386-5314-2/18/$31.00 ©2018 IEEE

Rest of the paper is organized as follows. Section II discusses the literature review and Section III describes the proposed methodology. Section IV reports the results and the paper concludes in Section V.

II. Literature Review

Several techniques are reported in literature for speech recognition systems using different models, databases, feature extraction methods and classification techniques [2]. A brief description of the existing methods and approaches to stuttered speech recognition is provided in this section.

Chee et al. have provided an overview of stutter recognition systems which highlights the strengths and weaknesses of different techniques used for stutter speech recognition [2].

Besides this, there are other methods proposed for both stuttered speech recognition as well as correction, using a supervised model, stuttering pruning, and automated text-to-speech based recognition [3].

In the supervised model proposed by Surya and Varghese [3] for stuttered speech recognition N audio signals are converted to audio array. The Mel Frequency Cepstral Coefficient (MFCC) features are extracted from the audio samples. During the classification stage, the extracted features are used to train support vector machine [3]. Since automatic speech recognition is a multiclass problem, SVM is extended to perform multiclass classification though SVM is basically a binary non-linear classifier. During testing, SVM classifies the stuttered input to correct word. Several words were manually segmented from the collected dataset. The SVM model is then trained using the segmented stuttered words. An accuracy of 76% was acquired in classifying the words correctly. The accuracy of this method can be improved by including more data in the training dataset. The major limitation of this method is that only the trained words get predicted.

Speech correction method implemented through stuttering pruning is implemented through neural networks in [3]. In this method, trained neural networks take the maximum amplitude of the speech signal as the input, and outputs a threshold for that amplitude value. The threshold output given by the network is used as the reference to remove any silent pauses and prolongations from the stuttered sample. The advantage of this method is that the original speech of the person gets reconstructed devoid of stuttering. But, this method acquired less accuracy i.e. of 62%. Its accuracy can be improved further by using more training data as well as incorporating some other features such as energy, frequency, of the audio input for training. Even though an average of 62% stuttering was removed, few non-stuttered parts also got cleaned from the output speech sample.

This paper aims to improve this method, by improving its accuracy using more training data, and designing the neural network to output an appropriate threshold needed to remove the stutter.

The last method of recognizing the stuttered speech, proposed in the paper [3], is converting the stuttered speech into equivalent text and then back to speech. This method converts words to equivalent texts. Powerful Artificial Neural Networks are used to identify each words in the speech. Then the speech content which is in the form of texts is reverted back to speech. The reversion process uses machine learning techniques to produce the text equivalent audio for each word. This method can remove silent pauses from the stuttered speech. It achieved an accuracy of 80%. Here complete sentences were also recognized instead of single trained words. This method requires powerful neural networks to remove any prolongations from the stuttered speech sample. It was not successful in removing any repetitions from the stuttered speech.

A new approach for correction and recognition of silent paused stuttered speech was also presented in [4]. In this approach, stuttering is eliminated by considering the fact that the voiced speech has more energy than the unvoiced speech. The feature extraction is performed using MFCC algorithm. The Vector Quantization code book is generated by clustering the training features vectors of the dysfluent speech, and then stored in the database. K-means clustering algorithm is used for the clustering purpose. Dynamic Time Warping algorithm is used to match the dysfluent speech with the database. Finally, the silent paused stuttered speech is corrected and the stutter-free speech is recognized. This method works only for isolated silent pause stuttering words, and needs to be further improved for complete sentences and for multi-modal stuttering.

There are papers employing MFCC algorithm followed by a classifier model. Jhawar proposed a method to recognize stutters in a speech sample [5]. The recognition accuracy varied with different classifier models. Another approach for stuttered speech recognition involved the use of Linear Predictive Cepstral Coefficient (LPCC) algorithm with k-nearest neighbors (k-NN) and Linear Discriminant Analysis (LDA) classifiers, which yielded an accuracy of 88.05%. But, these approaches were only developed to recognize and classify the recognized stutters, but not to correct those stutters from the stuttered speech sample.

Limited work has been reported in stuttered speech correction, and most of the papers have used neural networks as their primary technique for this problem by recording a lot of stuttered speech datasets and then feeding the correct speech to the system, trying to make it learn.

A major drawback of all these approaches is the amount of computation power that would be required for these techniques. Using only neural networks for stuttered speech recognition, it is not practically feasible to come up with an efficient solution since there are limited number of sentences than can be put into the system with all its permutations and combinations, to detect a sample correctly every time. Similarly, when using HMM with MFCC, many observations have to be predicted and larger the dataset, more will be the number of observations for HMM to predict from, and it might become less accurate.

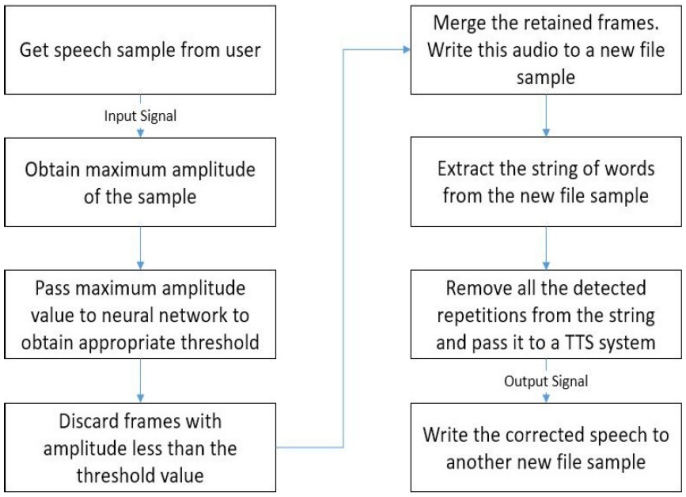
Hence, this paper aims at designing a system to detect as well as correct the detected stutters from the stuttered speech samples using a simple yet effective technique. This is

performed by using amplitude as a factor to remove the silent pauses and prolongations from a speech sample, and then performing a few string operations on the obtained sample to give the corrected speech sample devoid of any stutters.

III. Proposed Design Methodology

The proposed solution to a stutter-free system is based on the algorithm given below:

1. Obtain the speech sample with a duration of six seconds from the user, either through recording or as an audio clip.
2. Through functions from MATLAB, obtain the maximum amplitude of the samples.
3. Pass the maximum amplitude value to the created single neural network to obtain an appropriate threshold. The threshold obtained is always a function of the maximum amplitude.
4. Divide the speech sample into 300 short overlapping frames of equal length.
5. Discard the frames of the signal with amplitude less than the threshold value.
6. Merge the retained frames. Write the corrected audio to a new file sample.
7. After prolongations in the speech sample are removed, convert the new file sample to equivalent text, and write the text to a new text file using Java. This process is done to remove any existing repetitions of words or phrases
8. Perform a check on the file, at word level, for any undesired repetitions of words or phrases or characters.
9. Remove all the detected repetitions from the string.
10. Write the corrected text to a new text file and convert the new text to speech using any of the existing TTS systems. Write the corrected speech to a new file sample.



[Design Methodology of the Stuttered Speech Processing System]

IV. Results and Discussions

The stuttered speech processing system is implemented in MATLAB environment. A user-friendly GUI is created which enables user to record his/her own speech and the speech is then processed to generate the stuttered-free speech as shown below. The GUI offers the user the option to either remove only prolongations or all types of stutter from the user's speech. The types of stutters discussed here are part-word or full-word repetitions, prolongations and interjections.

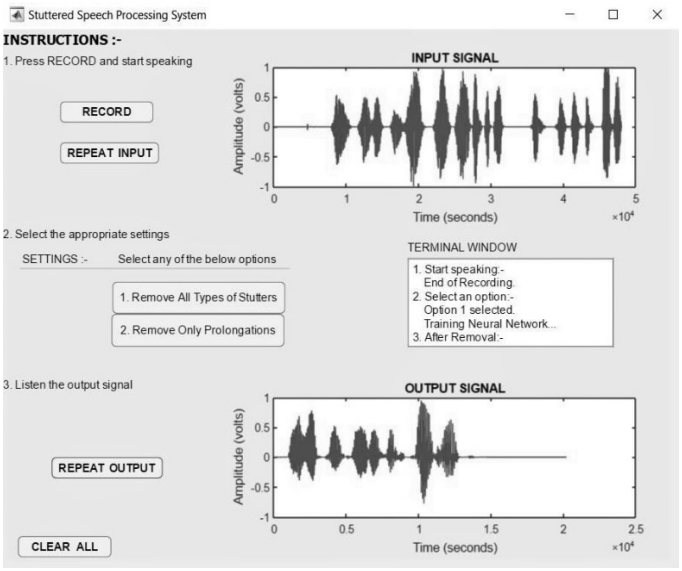


Fig. 1(a). Removal of all types of stutter from the user speech sample.

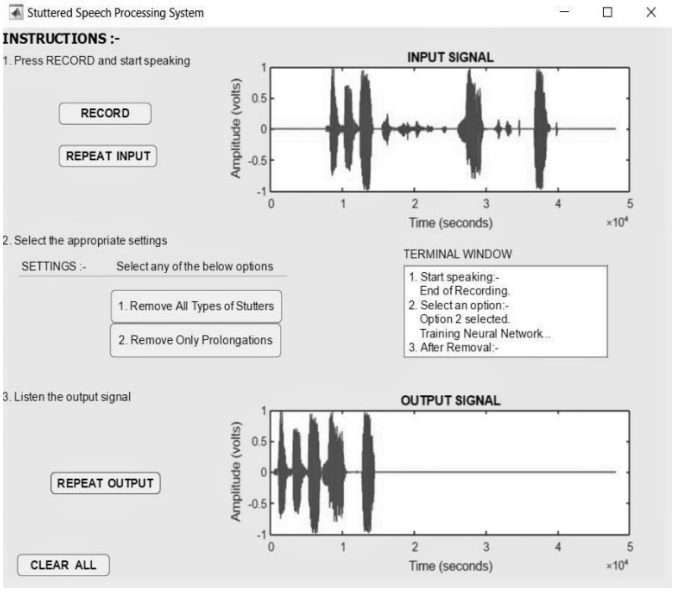


Fig. 1(b). Removal of prolongations from the user speech sample.

From a total of 110 speech samples, the neural network was trained with 60 speech samples, each with a duration of 6 seconds. Back propagation algorithm is used in the neural network for training the speech samples, to generate an appropriate threshold for each trained sample. The threshold value is and can be only used to remove the prolongations and interjections from the sample. Our major contribution to a complete stutter-free system is the elimination of any repetition(s) from the samples. The string repetition removal algorithm created by us effectively removes any repetitions (up to phrase-level) from the resulting sample in no time.

The system was tested with 50 different speech recordings collectively for either of the options selected in the GUI application, with 43 of them being detected and corrected accurately.

A summary of the results obtained is shown in the table below.

TABLE I: Performance of the developed stuttered speech processing system

Number of tested samples	Correct Output obtained	Incorrect output obtained
50	43	7

$$\text{Accuracy} = \frac{\text{Correct Output obtained}}{\text{Number of tested samples}} * 100 = 86\%$$

Hence the accuracy of the system is 86%. The process of obtaining the appropriate threshold consumes the maximum time (nearly 70% of the total time taken to produce the output). Average time taken for detection and correction of a user stuttered speech (of 6 seconds duration) when processed at a sampling rate of 8000Hz with twenty 300ms frames, was around 5-8 sec on a 2.6 GHz dual core i7 machine. The trained sentences were able to be detected and corrected within 5 seconds, while the correction of untrained sentences took around 8-10 seconds depending on the length and complexity of the sentence. Hence it can be inferred that the performance of the system can be improved by including even more samples in the training set.

V. Conclusion

The main objective of this paper is to present an algorithm that efficiently detects and corrects stutter in a speech segment of a person with stuttering speech disability. The proposed algorithm gives an accuracy level of 86% for 50 stutter speech samples. Two algorithms were used for more precise stutter removal system that can be built on any device.

The developed system can be incorporated into any existing speech recognition system. It can also serve as a speech therapy system where a user suffering from stutter can sound like the correct output obtained from the system. Hence the device can be used by people suffering from stutter to use the existing virtual assistant services, or talk to others with confidence using the device. This would enhance the level of communication amongst people with this disorder.

References

[1] Selim S. Awad, Louis Przebienda, Richard Merson, "The Application of Digital Speech Processing to Stuttering Therapy", *IEEE Instrumentation and Measurement Technology Conference*, Ottawa, Canada, pp. 1361–1367, May 1997.

[2] 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.

[3] Arya A Surya, Surekha Mariam Varghese**,** "Automatic Speech Recognition System for Stuttering Disabled Persons"**,** *International Journal of Control Theory and Applications*, Volume 9 Number 43, 2016.

[4] 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*, Vol.4 Issue 3, March 2015.

[5] Gunjan Jhawar, Prajacta Nagraj, P. Mahalakshmi, "Speech Disorder Recognition using MFCC", *International Conference on Communication and Signal Processing* April 6-8, 2016, India.