Stuttered Speech Recognition using Convolutional Neural Networks

1st Phani Bhushan S
Dept. Information Science and Engg JSS Science and Technological University
Mysuru, India

2nd Vani H Y
Dept. Information Science and Engg JSS Science and Technological University
Mysuru, India

3rd D K Shivkumar
Dept. Information Science and Engg JSS Science and Technological University
Mysuru, India

4th Sreeraksha M R
Dept. Information Science and Engg JSS Science and Technological University
Mysuru, India

***Abstract*—Stuttering or Stammering is a speech defect within which sounds, syllables, or words are rehashed or delayed, disrupting the traditional flow of speech. Stuttering can make it hard to speak with other individuals, which regularly influence an individual's quality of life. There are more than 70 million people worldwide who suffer from stuttering problems. Many people who suffer from this issue lose their confidence to speak in public and they find it difficult to fit in society. The proposed method is called SSR, which is based on the Weighted Mel Frequency Cepstral Coefficient feature extraction algorithm and Convolutional Neural Network for classification of stuttered events. In this work the focus is on detecting and removing the prolongation, silent pauses, and repetition to generate proper text sequence for the given stuttered speech signal. The work has utilized the UCLASS stuttering dataset which provides the data for stuttered speech in .wav format for analysis and customized dataset for validating the built system. The speech samples are parameterized to Weighted MFCC feature vectors. Then extracted features are inputted to the CNN for training and testing of the model. The test results show that the proposed method reaches the best accuracy of 89 Percent.**

***Index Terms*—Stuttered Speech Recognition (SSR), Convolution Neural Network (CNN), Mel Frequency Co-efficient (MFCC).**

I. INTRODUCTION

In today's world there are millions of people suffering from various speech disorders like stuttering, lisp, etc., this often renders them unable to utilize certain things that we take for granted, like speech recognition systems. Stuttering is one such speech disorder affecting the fluency of speech. It begins during childhood and, in some cases, lasts throughout life. The disorder is characterized by disruptions in the production of speech sounds, also called disfluencies. Most people produce brief disfluencies from time to time. For instance, some words are repeated, and others are preceded by um or uh. Disfluencies are not necessarily a problem; however, they can impede communication when a person produces too many of them.

In most cases, stuttering has an impact on at least some daily activities. The specific activities that a person finds challenging to perform vary across individuals. For some people, communication difficulties only happen during specific activities, for example, talking on the phone or talking before large groups, utilizing everyday tools that use speech as inputs. Currently, the speech recognition systems have a great accuracy for fluent speech but are unable to recognize speech with repetitions or long involuntary pauses, i.e., stuttering. This is mainly because the systems are created to stop the identification process when a pause is encountered. Also, these systems are trained with proper words without any repetitions and so, when it encounters a stuttered speech, it is unable to identify the words, since it has not been trained to do so. This paper aims to detect as well as correct these stuttered speech samples and then give the corrected speech sample devoid of stuttering. We will be using MFCC and CNN for stuttered speech recognition.

II. LITERATURE SURVEY

Marek Wisniewski, et.al. [1] In this paper, the authors applied speech recognition techniques to find disfluent events. A speech recognition system based on the Hidden Markov Model Toolkit was built and tested. Authors were not concentrated on specific disfluency type but tried to find any extraneous sounds in a speech signal. Patients read prepared sentences, the system recognized them and then results were compared to manual transcriptions. The aim of the studies on the automatic diagnosis of stuttering people was to develop a complete, non- supervised and fully objective tool that can support speech diagnosis and therapy process. Obtained results were good - sensitivity 88 percent and predictability 79 percent.

Ankit Dash, et.al. [2] The aim of this paper was to develop an algorithm to enhance speech recognition of a stuttered speech. 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.

Vikhyath Narayan, et.al. [3] The objective of this paper was to develop a method that could detect the dysfluency in stuttered speech. This helps Speech Language Pathologists (SLP) to assess stuttering patients, planning appropriate intervention program, and monitoring the prognosis during treatment. In the proposed system Mel-frequency cepstral coefficients (MFCC) were used for feature extraction of a signal. Decision logic was used for analysis of speech stuttering. Support vector machine was a classifier used for classification of stuttered speechsignal. The SVM classifier yielded an accuracy of 90 percent and 96.67 percent for dysfluent and fluent speech, respectively. In this work we have considered combination of three types of dysfluencies which are important in classification of dysfluentspeech.

Khalid A. Darabkh, et.al. [04] In this paper, the authors have proposed a speech recognition algorithm that is based on double threshold voice activity detection and Mel-frequency cepstral coefficients. Interestingly, a pre-emphasis is made for noise reduction and normalization when a word is recorded. The proposed speech recognition algorithm, in this article, uses the double threshold VAD technique which has proven to have a critical effect on the system performance. Including delta and acceleration coefficients has helped in enhancing the overall accuracy of the algorithm. The MFCC of the wavelet channels are computed to get the characteristics of the speech signals. Results showed that this approach gives better recognition ratethan MFCC features.

III. PROPOSED METHODOLOGY

A. *Datasets*

Stuttering speech recording is released by University College London's Archive in three versions UCLASS Releases One, Two and UCLASS-FSF in 2004 and 2008 respectively. The recording is collected between the age groups 18 to 45 with equally divided among male and female. The UCLASS one and two recordings were made in normal speaking conditions and the UCLASS-FSF was made when the sound of the speaker's voice was altered as he or she spoke. The data and software are freely available to anyone for research and teaching purposes. The dataset was available in both MP3 as well as WAV format along with the Orthographic and Phonetic transcription. The dataset consists of 490 audio recordings including monologues, readings, and conversations of children with known stutter disfluency issues. The resulting applicable data consisted of 25 unique conversations between an examiner and a child between the ages of 8 and 18, totalling to just over one hour of audio. Also 50 customized datasets were used to validate the accuracy of the built classifier.

B. *Stuttered Speech Recognition*

The figure 1 depicts the architecture of Stuttered Speech Recognition system using MFCC as feature extraction method and CNN as classifier. The Stuttered Speech Recognition system is carried out in four main steps namely., Pre-processing the data which involves Silence removal and Noise reduction, feature extraction, training, testing and classification. The architecture is illustrated in figure 1.

The first step is to extract the features from speech signal uttered by the speaker. The features will become the basic unit for classifying. The signals from the same users are tested and verified with CNN for the required output. At the end, the speech signal which is given as input is converted into text without any stutters.

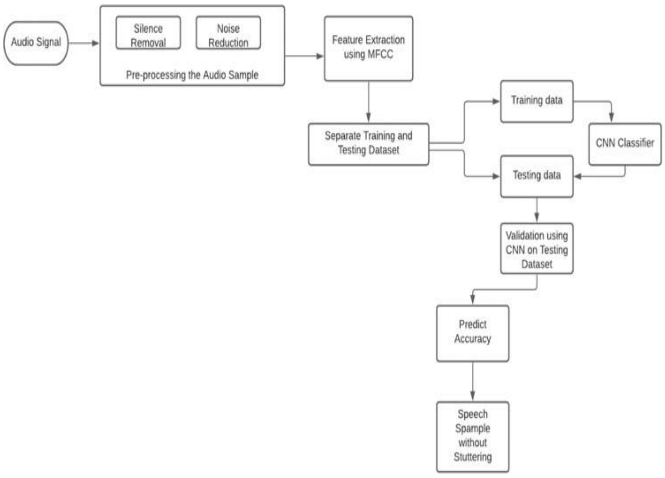

Audio Signal
Silence Removal
Noise Reduction
Pre-processing the Audio Sample
Feature Extraction using MFCC
Separate Training and Testing Dataset
Training data
CNN Classifier
Testing data
Validation using CNN on Testing Dataset
Predict Accuracy
Speech Spample without Stuttering

Fig. 1. The architecture of Stuttered Speech Recognition

C. *Feature Extraction – MFCC*

The figure 2 depicts the how MFCC extract the features from speech signal. The main purpose of extraction and processing of speech signal is to extract important features from raw audio. The proposed method uses MFCC approach to extract the features of the cepstral frequency. The most straightforward technique involves determining the average energy of the signal. This metric, along with total energy in the signal, indicates the volume of the speaker. The signal in the frequency domain through the (Fast) Fourier Transform is processed. The windowed samples is used to get accurate representations of the frequency content of the signal at different points in time. By taking the square value of the signal at each window sample, power spectrum can be derived. then the values of the power spectrum as features. The three largest frequency peaks for each window are obtained and add those to the feature vector. Mel frequency Cepstral Coefficients (MFCC) features were used for the experiments. MFCC have been shown to be an accurate representation of the spectra of speech signals and are thus highly discriminative in nature. 40 dimensional MFCCs are extracted for every frame with a 10 Ms second shift between consecutive frames.

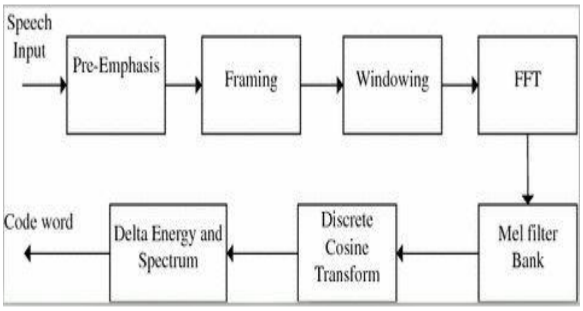


Fig. 2. Mel-frequency Cepstrum Coefficients (MFCC)

neural network data by providing the input is MFCC coeffi- cients that forecast the right signal emotion. A Convolution Layer of the kernel size 8, stage 2 and 128 function mappings are included in the network architecture. A pooling layer of size 6 follows. The two layers, each with 1024 components, are entirely linked. Finally, the categorization is based on a SoftMax layer.

D. Classifier – CNN

The figure 3 depicts the architecture diagram of Convolutional Neural Network. There are three basic layers in CNN,

- The Convolutional layer
- The Pooling layer
- The output layer Along with these three layers

there twoother layers which helps in classification. They are,

- Activation function
- Dropout layer

The input audio is passed to the first convolution layer and the convoluted output is obtained as an activation map. The filters applied in the convolution layer extract relevant features from the input image to pass further. Each filters gives a different feature to aid the correct class prediction. In case if need to retain the size of the image, use same padding, otherwise valid padding is used since it helps to reduce the number of features. The Pooling layers are then added to further reduce the number of parameters Several convolution and pooling layers are added before the prediction is made. Convolutional layer help in extracting features. As we go deeper in the network more specific features are extracted as compared to a shallow network where the features extracted are more generic. The output layer in a CNN as mentioned previously is a fully connected layer, where the input from the other layers is attended and sent so as the transform the output into the number of classes as desired by the network. The output is then generated through the output layer and is compared to the output layer for error generation. A loss function is defined in the fully connected output layer to compute the mean square loss. The gradient of error is then calculated. The error is then backpropagated to update the filter and bias values. One training cycle is completed in a single forward and backward pass.

The initial step is to turn these audio samples into trainable data from wav format. By using MFCC the coefficients are generated by Melcepst and then the coefficients are normalised, and the data randomised. Training of the Convolution

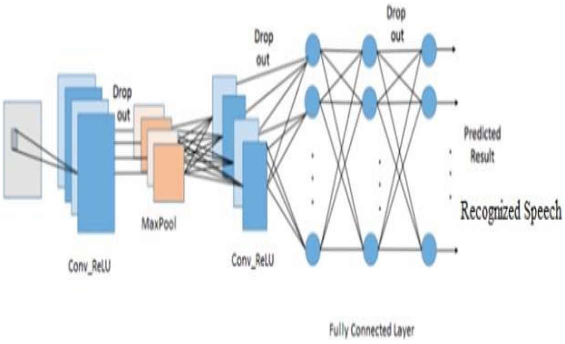


Fig. 3. Convolution Neural Network (CNN)

E. Result

In the UCLASS dataset, there are 490 audio files available, among them 392 file is used as training, 49 is used for validation and other 50 customized datasets are used for testing. Each model is trained for 10 epochs to guarantee fair comparison between all models.

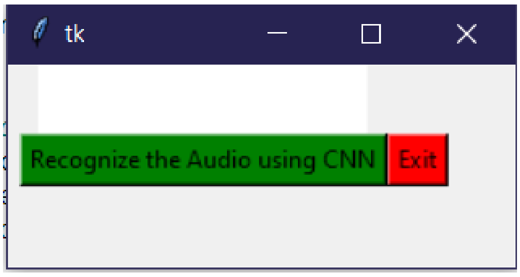
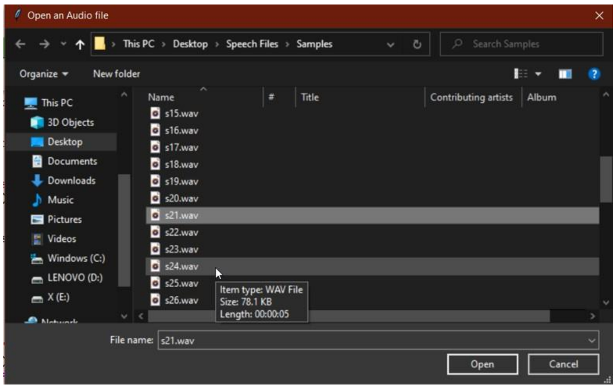


Fig. 4. TINKER GUI displaying 2 options: (a) Recognize the Audio (b) Exitthe GUI

In Fig 4. it is seen that the GUI when opened gives the user with 2 set of choices :

- To Recognize the stuttered Audio using CNN
- To Exit the GUI

The first option on selection prompts the User to select any file from the Validation list and consequently predicts the output and the second option exits from the GUI and closesthe program.



File name: s21.wav

Fig. 5. Selecting a Dataset to predict the Speech from the audio Samples

The Fig 5. shows the path where the User can select the dataset for validating the audio speech sample.

The user can select any audio from the Dataset and for every selected Audio the prediction is done and consequently the recognized speech is exerted as the final output.

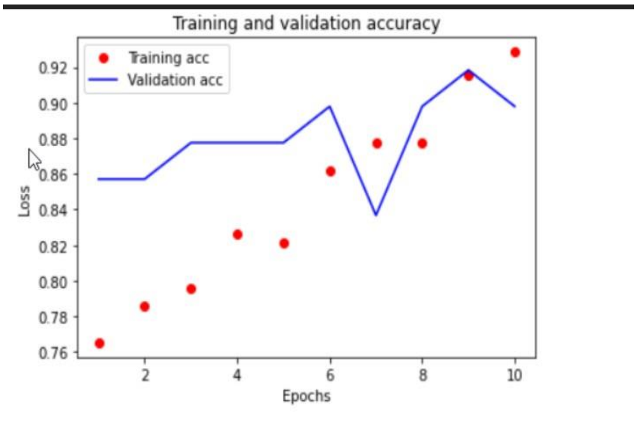


Fig. 6. Training and Validation Accuracy for CNN

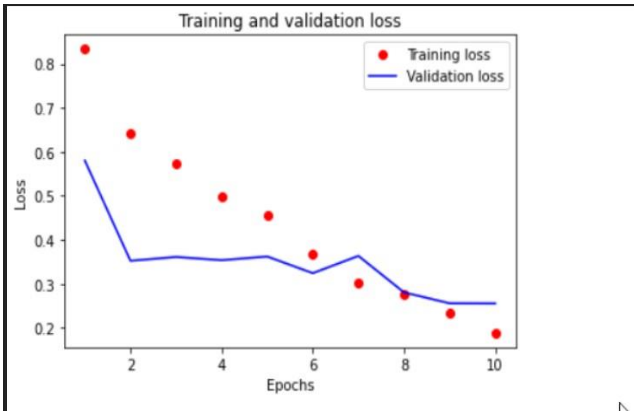
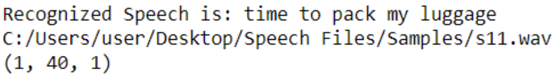


Fig. 7. Training and Validation Loss for CNN

The Figures Fig 6. and Fig 7. are the graph plots of Training and Validation Accuracy's and Losses for the Audio Datasamples.



```
Recognized Speech is: time to pack my luggage
C:/Users/user/Desktop/Speech Files/Samples/s11.wav
(1, 40, 1)
```

Fig. 8. Final Recognized Speech

The Figure 8 is the resultant text of the Recognized StutteredSpeech from the Audio Data Sample.

F. Conclusion

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formal- izations of machine learning problems driven by practical applications. In the proposed method we have investigated the scalability of a Stuttered Speech recognition based on CNN, which takes as input the raw speech signal, to large vocabulary task. In the proposed method we have achieved 92 percent accuracy using CNN with 7 percent validation loss.

Model	Dataset	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Test Accuracy
CNN	UCLASS + Customized Dataset	0.1884	0.9286	0.2555	0.8990	0.8990

Fig. 9. Final Recognized Speech

REFERENCES

[1] Arya A Surya and Surekha Mariam Varghese, "Automatic Speech Recognition System for Stuttering Disabled Persons", International Journal of Control Theory and Applications, Vol 10, 2017.

[2] Lalima Singh, "Speech Signal Analysis using FFT and LPC", In- ternational Journal of Advanced Research in Computer Engineering Technology (IJARCET), Volume 4 Issue 4, April 2015.

[3] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuth, "Voice Recogni- tion Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Volume 2, Issue 3, March 2010.

[4] Zheng Fang, Zhang Guoliang, Song Zhanjiang, "Comparison of Differ- ent Implementations of MFCC", Journal of Computer Science Technol- ogy, Vol. 16, Nov 2001.

[5] Tedd Kourkounakis, Et al., "FluentNet: End-to-End Detection of SpeechDisfluency with Deep Learning"

[6] Aweem Ashar, Et al., "Speaker Identification Using a Hybrid CNN- MFCC Approach"

[7] Sadeen Alharbi, Et al., "Automatic recognition of children's read speechfor stuttering application"

[8] R. Prabhu, Et al., "Speech Based Anti Stuttering Algorithm using Matlab", IJEAT, Volume-9 Issue-3, February 2020

[9] Ankit Dash, Et al., "Speech Recognition and Correction of a Stuttered Speech", IEEE-2018

[10] S.Girirajan, Et al., "Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory", IJRTE, Volume-8 Issue-5, January 2020

[11] Yanick Lukic, Et al., "SPEAKER IDENTIFICATION AND CLUS- TERING USING CONVOLUTIONAL NEURAL NETWORKS", IEEE-2016

[12] Wouter Gevaert, Et al., ”Neural Networks used for Speech Recognition”

[13] , VOL. 20:1-7,IEEE-2010

[14] Sakshi Gupta, Et al., ”Deep Learning Bidirectional LSTM based Detec- tion of Prolongation and Repetition in Stuttered Speech using Weighted MFCC”, IJACSA, Vol. 11, No. 9, 2020

[15] Shaswat Rajput, Et al., ”Speech Stuttering Detection and Removal”,

[16] http://biruntha-signalprocessing.blogspot.com/2015/09/frame-blocking- of-speechsignal.html

[17] https://en.wikipedia.org/wiki/Mel-frequencycepstrum

[18] https: //en.wikipedia.org/wiki/Neuralnetwork

[19] https://en.wikipedia.org/wiki/CNN