Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 16, Number 1 (2023) pp. 45-53
© Research India Publications
https://dx.doi.org/10.37622/ACST/16.1.2023.45-53

Comparison of Machine Learning Algorithms for Detection of Stuttering in Speech

Dr. P.Sunitha Devi, Sarvagna Gudlavalleti, Ramyasri Lakka, Rithika Kuchanpally
Asst.Professors, Department of Computer Science and Engineering
G Narayanamma Institute of Technology and Science Hyderabad, India
sunitha@gnits.ac.in, sarvagnagudlavalleti@gmail.com
lakkaramyasri@gmail.com, rithikakuchanpally@gmail.com

Sai Sonali Dudekula
Department of Electronics and Telecommunication Engineering
G Narayanamma Institute of Technology and Science, Hyderabad, India
saisonali0506@gmail.com

Abstract

Stuttering is a speech disorder characterized by repetition of sounds, syllables, or words;prolongation of sounds;An individual who stutters exactly knows what he or she would like to say but has trouble producing a normal flow of speech.This project aims to use machine learning algorithms to detect stuttering behavior in Telugu language speech samples. Stuttering is a speech disorder that affects the flow of speech, making it difficult to communicate with others. The study will collect voice samples from individuals reading the same Telugu script and generate 8KHz and 16KHz samples in .wav files. The data will be annotated and coded on a 0-8 scale and will be generated as a data set. Machine Learning algorithms will be used to build a model for detecting stuttering patterns in the audio files.Finally the model will be integrated with a user interface.

I. INTRODUCTION

The speech disorder known as stuttering, which is also known as stammering or childhood-onset fluency disorder, disrupts the normal rhythm and fluency of speech on a regular basis [1], [2]. People who stutter have trouble articulating their thoughts and feelings in speech, often by prolonging or repeating a consonant, vowel, or syllable. When they come across a difficult word or sound in conversation, they may pause for a moment.

Young children, whose linguistic skills are still maturing, frequently stutter. Young children typically grow out of this kind of stuttering as they develop their language skills. However, some people struggle with stuttering throughout their entire lives. People who stutter often exhibit additional behaviors alongside their primary stuttering, such as grimacing, blinking, and other tics. This kind of stuttering can have a negative effect on confidence and social interactions.

Stuttering treatment options include speech therapy, fluencyenhancing technology, and cognitive behavioral therapy for both children and adults. Early stuttering diagnosis is difficult for speech therapists because there are so many potential triggers.

II. MOTIVATION AND RELATED WORK

A. Motivation

Detecting stuttering behavior at an early stage is crucial for proper assessment and treatment of speech disorders. Therefore, the main goal of this research is to develop a reliable stuttering behavior detection model that can accurately identify the extent of stammering behavior in individuals. By using machine learning algorithms and analyzing speech samples, this model can help speech pathologists to establish a more accurate assessment of stuttering behavior without the need for manual counting of speech disfluencies.

Speech disorders or speech impairments can take many forms, including stuttering, lisping, and other conditions that make it difficult for individuals to create or form normal speech sounds. These disorders can impact an individual's ability to communicate effectively with others, leading to social and psychological challenges. Therefore, accurate diagnosis and treatment of speech disorders are critical for ensuring individuals can communicate effectively and lead fulfilling lives.

By developing an effective stuttering behavior detection model, this research could significantly improve the lives of individuals who struggle with speech disorders. The model could help to identify stuttering behavior at an early stage, allowing for more targeted and effective interventions to be implemented. Furthermore, it could reduce the need for manual assessment, making the assessment process quicker and more efficient for both clinicians and patients.

B. Related Work

There is a wide range of ongoing research projects (in terms of acoustic feature extraction and classification methods) aimed at developing automatic tools for the detection and identification of stuttering, which is an interdisciplinary research problem. The majority of the existing work uses language models or automatic speech recognition (ASR) systems [3] to detect and identify stuttering. These systems work by first converting audio signals into textual form, which can then be analyzed using language models. This section offers a thorough analysis of the many stuttering identification methods that have been implemented using acoustic-based feature extraction and machine learning [4].

Speech signal processing tools have been traditionally utilized for speech detection but cater to only problems such as dysarthria and hearing impairment, however, very little

data is available with respect to the other speech related problems such as stuttering which is what this paper aims at addressing.

The entire system is divided into two phases which are the training and testing phase where feature extraction is performed in both the phases [5].In the training phase, we want to capture the patterns found in the feature vectors extracted from the disfluency utterance while in the testing phase ,the same feature extraction is performed on a unknown disfluency , these features are then passed to the classifier , which matches it to the closest matching disfluency and classifies the sample data as one of the disfluencies we are using for stuttering data ,which are, syllable repetition, word repetition, prolongation, and interjection.

The method only provides four classifications for the disfluencies which on the other hand other disfluencies such as involuntary silences ,blocks in communication are not considered . Since, only audio sample is utilized as the input data , when there are blocks in communication, it would be difficult to differentiate between blocks and purposeful pauses given by the speaker .

Another proposed research presents an automated method for evaluating stuttered speech using a combination of the Weighted Mel Frequency Cepstral Coefficients (WMFCC) feature extraction method and the deep-learning-based classification method Bidirectional Long-Short Term Memory (BiLSTM). The efficacy of this model was tested against other classification models, and the results showed that the proposed model had a better classification accuracy of 96.67 percent for detecting disfluencies like prolongation and repetition of syllables, words, and phrases.

The proposed model uses 14-dimensional WMFCC feature vectors that can extract both static and dynamic acoustic features, leading to enhanced detection accuracy of stuttered events while reducing computational overhead. The Bi-LSTM model used in the proposed method can learn long dependencies and take full account of disfluency patterns in speech frames. The experiments conducted on the proposed model demonstrated its effectiveness in improving the recognition accuracy of stuttered events.

All the prior works mainly focused on studying the disfluencies in British and American English but the work is not done in any available Indian language.So, our work was mainly focused on generating a dataset in Telugu language.We collected speech samples from individuals with stuttering disorders in Telugu language, which we then labeled and preprocessed for feature extraction. We used various acousticbased features, such as Mel Frequency Cepstral Coefficients (MFCCs), to extract information about the speech signal that could be used to identify disfluencies.

We then trained and tested our machine learning models using these features,our results showed that our proposed method was effective in detecting stuttering in Telugu language, achieving high accuracy rates in both training and testing phases. This research has significant implications for the development of automatic stuttering detection tools for Indian languages, which could potentially improve the diagnosis and treatment of stuttering disorders in the region.

Overall, our research contributes to the broader body of literature on stuttering detection and has important implications for speech pathology and clinical practice in India. It highlights the need for further research in this area, particularly in other Indian

languages, and demonstrates the potential of acoustic-based feature extraction and machine learning techniques for developing automated stuttering detection tools.

III. PROPOSED FRAMEWORK

Proposed framework mainly includes construction of a dataset for facilitating studies on speech disfluencies in Telugu language primarily.It then goes on to building machine learning models using four different machine learning algorithms and test the models with other voice samples and verifying the efficiency of the model by measuring accuracy and F1 score.The final stage includes integrating the model with a user interface for general purpose use.

A. Construction of the dataset

To create a reliable dataset for building a model that can analyze stuttering in children's speech, a multi-step process was undertaken. First, voice clips were collected from several schools, and then they were preprocessed to eliminate any background noise and filtered to optimize the quality of the recordings. Next, the voice clips were uploaded into Praat, a software commonly used for speech analysis, and segmented into syllables as shown in Fig. 1. Each syllable was then marked according to its degree of stuttering, which was classified into a range of 0-7 as shown in Fig. 2 . This approach allows for a more nuanced understanding of the severity of stuttering in children's speech, as the range enables the identification of both mild and severe instances of stuttering [6]. Finally, all of the annotated information was organized and recorded into a CSV file, which will be used as the input dataset for the model building process. The approach of segmenting, labeling, and annotating the voice samples in this manner ensures a high-quality dataset that can be used to build a reliable and accurate model for analyzing stuttering in children's speech.

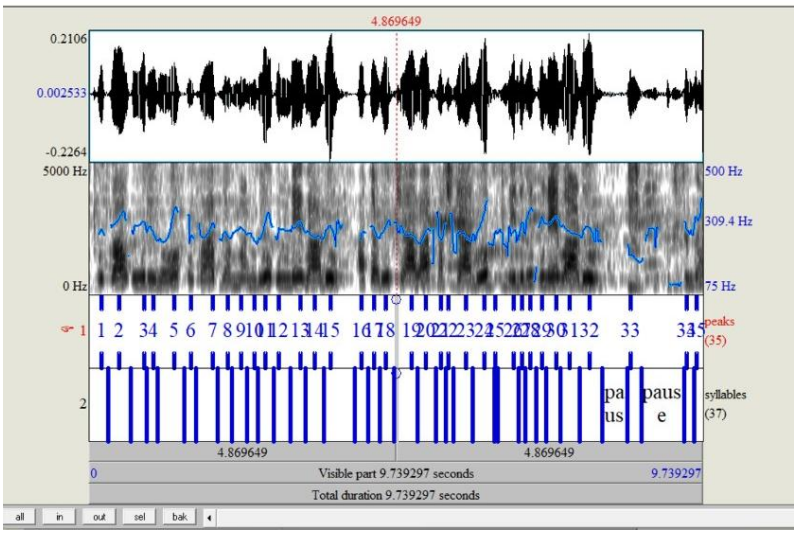

Fig. 1. Annotating the voice clip in Praat Software

I14 fx

	A	B	C	D	E
1	Syllable	Start Time	End Time	Total Duration	Class
2	/o/	0.976264	1.151783	0.175518	0
3	/ka/	1.152288	1.584288	0.432	0
4	/U/	2.048882	2.341414	0.292532	0
5	/ri/	2.341414	2.575439	0.234025	0
6	/lO/	2.575439	3	0.424561	0
7	/o/	3.160504	3.355525	0.195021	0
8	/ka/	3.355525	3.531044	0.175519	0
9	/brAH/	3.531044	3.921087	0.390043	0
10	/ma/	3.921087	4.096606	0.175519	0
11	/Nu/	4.096606	4.272126	0.17552	0
12	/du/	4.272126	4.681671	0.409545	6
13	/un/	5.208228	5.461756	0.253528	0
14	/de/	5.481258	5.617773	0.136515	0
15	/vA/	5.617773	5.988314	0.370541	0
16	/du/	5.988314	6.339352	0.351038	0
17	/a/	7.197446	8.77712	1.579674	3
18	/thA/	7.197446	8.77712	1.579674	3
19	/du/	7.197446	8.77712	1.579674	3
20	/chA/	8.855128	9.167163	0.312035	0
21	/lA/	9.167163	9.537703	0.37054	0
22	/kha/	9.849737	9.986252	0.136515	0
23	/stA/	9.986252	10.376295	0.390043	0
24	/lu/	10.376295	10.746836	0.370541	0
25	/ye/	11.05887	13.204105	2.145235	3
26	/du/	11.05887	13.204105	2.145235	3
27	/ru/	11.05887	13.204105	2.145235	3
28	/ku/	13.204105	13.36022	0.156115	0
29	/nna/	13.36022	13.691659	0.331439	0
30	/du/	13.691659	13.925684	0.234025	0

Fig. 2. Syllable level division of voice sample

B. Data preprocessing

A set of 39 MFCC coefficients are obtained from the voice sample input and their frame length is computed [7]. These coefficients are utilized as the X input for the model. Next, time variables t1 and t2 are defined, where t1 is set to 0 and t2 is calculated as the

duration of the input voice sample divided by the frame length. These variables are then checked against four constraints, which include scenarios where t1 is less than the start time and t2 is less than the start time, t1 is greater than or equal to the start time and t2 is less than or equal to the end time, t1 is less than the start time and t2 is greater than or equal to the start time, and t1 is less than the end time and t2 is greater than the end time, and t2 is less than the start time of the next value. These constraints represent all possible scenarios of where t1 and t2 can exist between. Subsequently, the Y input for the model is generated by appending the t values to the Y list, where t is computed as the ratio of the audio clip duration to the number of frames.

C. Model Building

After preprocessing the data and obtaining the X and Y labels, they are then fed into the machine learning models for further analysis. In order to assess the performance of these models, 80 percent of the data is utilized for training the model while the remaining 20 percent is reserved for testing the model's accuracy [8].

Four distinct machine learning models are created for the purpose of comparing their respective accuracies. These models are constructed using a variety of algorithms, such as the Random Forest Algorithm, Decision Tree Algorithm, Naive Bayes Algorithm, and Logistic Regression Algorithm. By leveraging these algorithms, each model is able to learn and extract meaningful patterns and relationships from the data, with the aim of accurately classifying a given voice sample as either stuttered or normal speech.

During testing, the trained models are validated by comparing their performance against the training data. The outputs generated by the models are then analyzed to determine whether a given voice sample is stuttered or normal speech. To further evaluate the efficiency of the model's output, various metrics such as accuracy, F1 score, precision, and recall values are calculated. These metrics provide a comprehensive assessment of the model's overall performance, helping to identify any potential areas of improvement and optimize the model for future use.

IV. RESULTS AND DISCUSSIONS

Out of the four models built, Random Forest Algorithm gave the highest amount of accuracy of 86.97 percent and an F1 score of 0.62.The next highest accuracy was obtained from Logistic Regression Algorithm which was 82.81 percent.All the results are mentioned below as given in Table I.

A. Model Integration with User Interface

The Random Forest Algorithm was identified as the optimal choice for generating the model, given its highest accuracy score [9]. To facilitate ease of access to the end user, a web application was developed and integrated with the model. The front-end of the web page was created using HTML and CSS, and included several features such as the ability to upload a voice clip in the form of a .wav file.

TABLE I: RESULTS OBTAINED FROM ALL THE MODELS

Model	Accuracy	Precision	Recall	F1 score
Random Forest	86.97	1.0	0.45	0.62
Logistic Regression	82.81	0.68	0.5	0.57
Decision Tree	78.12	0.55	0.47	0.51
Naive Bayes	74.47	0.46	0.5	0.48

In addition, the web page also provided an option for recording a voice clip and saving it to the desktop in cases where a voice clip was not readily available. To integrate the model with the web application's user interface, a flask server was employed [10].. This allowed for the input from the webpage to be directly fed into the machine learning model, which would then be run in the backend. Once processed, the final output would be displayed on the web page, indicating whether the given voice clip was indicative of stuttered speech or normal speech as shown in Fig. 3 and Fig. 4.

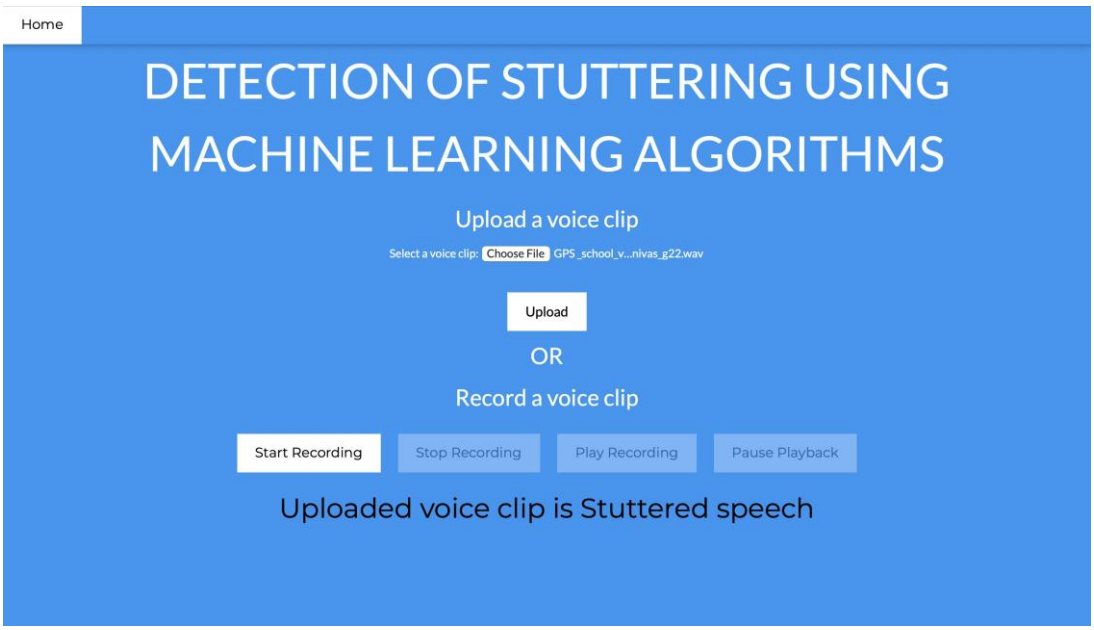

Fig. 3. Result of stuttered voice sample

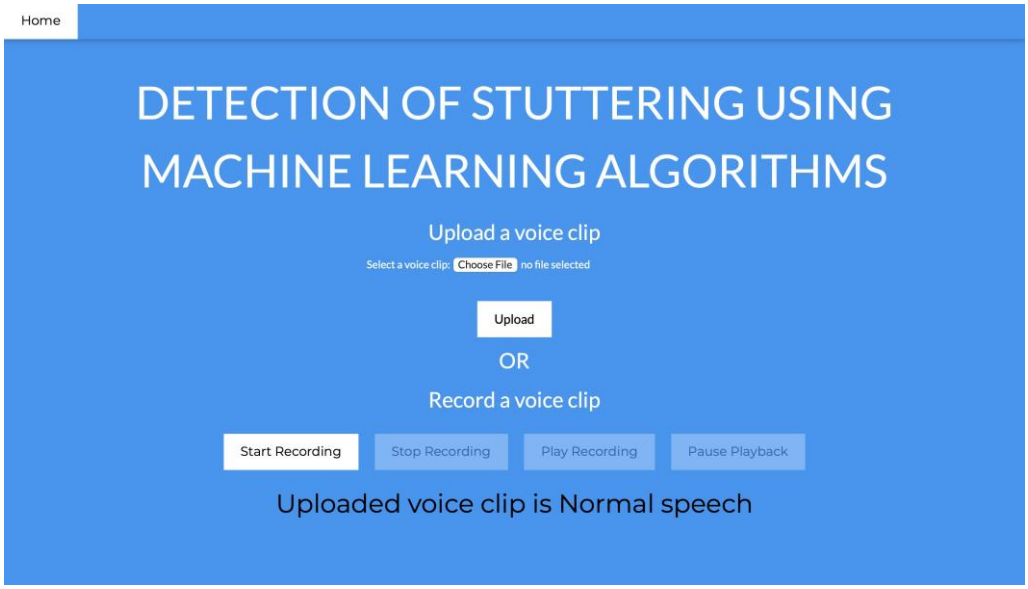

Fig. 4. Result of Normal voice sample

V. CONCLUSION

In this study, a dataset of voice clips from several children was collected in the form of a .wav file. To prepare the data for modeling, the voice clips were preprocessed using the Praat software [11], which involved marking regions by syllables and creating csv files as an input dataset to the machine learning models. Four different models were trained and tested using the dataset, and the Random Forest Model proved to be the most accurate with an impressive 86.97 percent accuracy. To make the model more accessible to end-users, a web page was created using HTML and CSS, with the backend integrated using Flask server. The user interface was designed to allow for easy uploading of a voice clip in .wav format, with the output displayed on the web page indicating whether the given voice sample was stuttered or not. This streamlined approach to utilizing the machine learning model increases the potential for widespread use and impact.

REFERENCES

[1] E. Shriberg, "To'errrr'is human: ecology and acoustics of speech disfluencies," Journal of the International Phonetic Association, pp. 153–169, 2001.

[2] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," Language and Linguistics Compass, vol. 2, no. 4, pp. 589–602, 2008.

[3] S. R. Maskey, Y. Gao, and B. Zhou, "Disfluency detection for a speechto-speech translation system using phrase-level machine translation with weighted finite state transducers," Dec. 28 2010, uS Patent 7,860,719.

[4] Sakshi Gupta, Ravi S., Rajesh K., Rajesh Verma. ”Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC”, International Journal of Advanced Computer Science and Applications, 2020

[5] P. Mahesha, D.S. Vinod. ”Gaussian Mixture Model Based Classification of Stuttering Dysfluencies”, Journal of Intelligent Systems, 2016

[6] P. Howell, S. Davis, and J. Bartrip, “The university college london archive of stuttered speech (uclass),” 2009.

[7] R.Riad,A.-C.Bachoud-Le vi,F.Rudzicz,andE.Dupoux,“Identification of primary and collateral tracks in stuttered speech,” arXiv preprint arXiv:2003.01018, 2020.

[8] A. Montplaisir-Gonclaves, N.Ezzati-Jivan, F.Wininger, M.R. Dagenais. ”State History Tree:An Incremental Disk-Based Data Structure for Very Large Interval Data”, 2013 International Conference on Social Computing, 2013

[9] Ali, Jehad Khan, Rehanullah Ahmad, Nasir Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[10] Aslam, Fankar Mohammed, Hawa Lokhande, Prashant. (2015). Efficient Way Of Web Development Using Python And Flask.. International Journal of Advanced Research in Computer Science. 6.

[11] Boersma, Paul Weenink, David. (2001). PRAAT, a system for doing phonetics by computer. Glot international. 5. 341-345.