Automated Stuttering Classification

Subjects: Computer Science, Artificial Intelligence | Biology

Contributor: Krishna Basak , Nilamadhab Mishra , Hsien-Tsung Chang

Speech disfluency, particularly stuttering, can have a significant impact on effective communication. Stuttering is a speech disorder characterized by repetitions, prolongations, and blocks in the flow of speech, which can result in communication difficulties, social isolation, and low self-esteem. Stuttering can also lead to negative reactions from listeners, such as impatience or frustration, which can further exacerbate communication difficulties.

stuttered speech | speech disfluency | multi-head self-attention

1. Background: Stuttering and Its Types

Stuttering is not a disease but a disorder that can be cured through proper consultation [1]. It has many types depending upon how they hinder fluent speech. A summary of these stuttering types is given in **Table 1** and further explained below. Block (BL) refers to sudden pauses in vocal utterances in between the speech. For example—I want [...] pizza—where there is a distinct gap between the speech. This pause is involuntary and hard to detect from audio signals only.

Table 1. Types of Stuttering Disfluency with Descriptions.

Label	Dysfluency Type	Description
BL	Block	Involuntary pause
PR	Prolongation	Elongated syllable
SR	Sound repetition	Repeated syllable
WR	Word repetition	Repeated word
IJ	Interjection	Added filler word
PhR	Phrase Repetition	Repeated phrase
ND	No dysfluency	Fluent speech

Prolongation (PR) happens when the speaker elongates a syllable/phoneme of any word during speaking. The duration of such elongations varies according to the severity of dysfluency and is often accompanied by high pitch. An example of this is—Have a ni[iiii]ce day.

In stuttering disfluency, Repetition is stated as the quick repetition of a part of speech. It is further classified into different categories. Sound Repetition (SR) happens when only a small sound is repeated. For example—I am

re[re-re-re]ady, where their sound is repeated more than once. In Word Repetition (WR), the speaker repeats a complete word as in I am [am] fine. Phrase Repetition (PhR), as the name suggests, is the repetition of a phrase while speaking. An example of this is—He was [he was] there.

The last stuttering type is Interjection (IJ), in which the speaker utters some filler words/exclamations that do not belong to the spoken phrase. Some common filler words are ‘um’, ‘uh’, ‘like’, ‘you know’, etc. The No Dysfluency (ND) in **Table 1** does not refer to any stuttering type. It is for when someone/some audio clip does not have any stuttering problems. In this research, the researchers focus on detecting the following stuttering types: BL, PR, SR, WR, and IJ. These 5 stuttering types are the most common and implemented in most of the research work.

2. Stutter Classification Using Classic Machine Learning

The paper [2] focused on the use of Linear Predictive Cepstral Coefficient (LPCC) to identify prolongations and repetitions in speech signals. The authors of the paper manually annotated 10 audio clips from University College London’s Archive of Stuttered Speech Dataset (UCLASS)—a single clip from each of the 8 male and 2 female speakers. They then extracted the LPCC feature from the clips by representing the Linear Predictive Coefficient (LPC) in the cepstrum domain [3] using auto-correlation analysis. Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (k-NN) algorithms were used to classify the clips. The authors obtained 89.77% accuracy while using k-NN with k = 4 and 87.5% accuracy using the LDA approach.

Mel-Frequency Cepstral Coefficients (MFCCs) are used as the speech feature in [4] to determine if an audio clip has repetition or not. The authors employed the Support Vector Machine (SVM) algorithm as a classifier in the attempt to identify diffluent speech from 15 audio samples. Their approach resulted in 94.35% average accuracy. The paper [5] also emphasized using MFCCs and obtained an average of 87% accuracy using Euclidean Distance as the classification algorithm.

The work undertaken in [6] explored the applicability of the Gaussian Mixture Model (GMM) for stuttering disfluency recognition. They curated a dataset containing 200 audio clips from 40 male and and 10 female speakers and annotated each clip with one of the following stuttering types—SR, WR, PR, and IJ. The authors extracted MFCCs from each of the clips and trained the model. They achieved the highest average accuracy of 96.43% when using 39 MFCC parameters and 64 mixture components.

The work [7] suggested that Speech Therapy has a significant effect on curing stuttered speech. In this research, the authors introduced the Kassel State of Fluency Dataset (KSoF) containing audio clips from PWS and underwent speech therapy. KSoF contains 5500 audio clips of 6 different stuttering events—BL, PR, SR, WR, IJ, and therapy-specific speech modifications. The authors extracted ComParE 2016 [8] features using OpenSMILE [9] and wav2vec 2.0 (W2V2) [10] and then trained an SVM classifier with a Gaussian kernel. The model produced a 48.17% average F1 Score.

Table 2 provides a summary of different ML methods used for stutter classification. Most of the works that utilize classical ML methods have used less number of audio clips—often curated by the authors themselves. Given the variability of stuttering disfluency, these small datasets neither represent a wide range of speakers nor have much data available for the ML models to get trained properly. This might cause the models to be biased.

Table 2. Summary of Prior Machine Learning Approaches for Stuttered Speech Classification.

Paper	Dataset	Feature	Model/Method	Results
[2]	UCLASS	LPCC	k-NN and LDA	Acc. 89.27% for k-NN and 87.5% for LDA
[4]	Custom	MFCC	SVM	Avg. Acc. 94.35%
[5]	Custom	MFCC	Euclidean Distance	Avg. Acc. 87%
[6]	Custom	MFCC	GMM	Avg. Acc. 96.43%
[7]	KSoF	OpenSMILE and wav2vec 2.0	SVM	Avg. F1 48.17%

3. Stutter Classification Using Deep Learning

The work performed in [11] explores the usage of respiratory bio-signals to differentiate between BL and non-BL speech. The authors carried out the research where a total of 68 speakers (36 Adult Who Stutter (AWS) and 33 Adult Who Do Not Stutter (AWNS)) were given a speech-related task and their respiratory patterns and pulse were recorded. Various features were extracted from the bio-signals and a Multi-Layer Perceptron (MLP) was trained to classify them. Their approach resulted in 82.6% accuracy.

In the paper [12], the authors explored Residual Networks (ResNet) [13] and Bidirectional Long Short-Term Memory (Bi-LSTM) [14]. Long Short-Term Memory (LSTM) is used in speech processing and natural Language Processing (NLP) and it is effective for classifying sequential data [15]. The authors manually annotated a total of 800 audio clips from UCLASS [16] to train the model and obtained a 91.15% average accuracy.

The FluentNet architecture suggested in [17] is a successor of the previous paper, where the authors upgraded the normal ResNet to a Squeeze-and-Excitation Residual Network (SE-ResNet) [18] and added an extra layer of Attention Mechanism (Attention) to focus on the important parts of speech. The experiments were performed using UCLASS and LibriStutter—a synthetic dataset built using clips from LibriSpeech ASR Corpus [19]. They obtained an average accuracy of 91.75% and 86.7% after training the FluentNet using Spectrogram (Spec) obtained from UCLASS and LibriStutter, respectively.

The study [20] used a Controllable Time-delay Transformer (CT-Transformer) to detect speech disfluencies and correct punctuation in real time. In this research, the authors first created the transcripts for each audio clip [21] and then speech Words and Positional Embed were generated from each transcript. In this way, a CT-Transformer was

trained on the IWSLT 2011 [22] dataset and an in-house Chinese dataset. The model obtained an overall 70.5% F1 Score for disfluency detection using the in-house Chinese Corpus.

One of the recent deep learning (DL) models for stutter classification is StutterNet [23]. The authors used the Time-Delay Neural Network (TDNN) model and trained it using MFCC input obtained from UCLASS. The optimized StutterNet resulted in 50.79% total accuracy while classifying stutter types—BL, PR, ND, and Repetition.

In **Table 3**, a summary is given of existing DL models for stuttered speech classification. Also, the paper [24] conducted a comprehensive examination of the various techniques used for stuttering classification, including acoustic features, statistical methods, and DL methods. Additionally, the authors highlighted some of the challenges associated with these methods and suggested potential future avenues for research.

Table 3. A Summary of Previous Deep Learning Methods for Stuttered Speech Classification.

Paper	Dataset	Feature	Model/Method	Results
[11]	Custom	Respiratory Bio-signals	MLP	Acc. 82.6%
[12]	UCLASS	Spectrogram	ResNet + Bi-LSTM	Avg. Acc. 91.15%
[17]	UCLASS + LibriStutter	Spectrogram	FluentNet	Avg. Acc. 91.75% and 86.7%
[20]	In-house Chinese Corpus	Word and Position Embedding	CT Transformer	F1 70.5%
[23]	UCLASS	MFCC	StutterNet	Acc. 50.79%

References

1. Dalton, P. Approaches to the Treatment of Stuttering; Routledge: Abingdon-on-Thames, UK, 2018.
2. Chee, L.S.; Ai, O.C.; Hariharan, M.; Yaacob, S. Automatic detection of prolongations and repetitions using LPCC. In Proceedings of the 2009 International Conference for Technical Postgraduates (TECHPOS), Kuala Lumpur, Malaysia, 14–15 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1–4.
3. Wong, E.; Sridharan, S. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In Proceedings of the 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489), Hong Kong, China, 4 May 2001; IEEE: Piscataway, NJ, USA, 2001; pp. 95–98.
4. Ravikumar, K.; Rajagopal, R.; Nagaraj, H. An Approach for Objective Assessment of Stuttered Speech Using MFCC. International Congress for Global Science and Technology. 2009; Volume

19. Available online: http://www.itie.in/Ravi_Paper_itie_ICGST.pdf (accessed on 20 September 2023).

5. Jhawar, G.; Nagraj, P.; Mahalakshmi, P. Speech disorder recognition using MFCC. In Proceedings of the 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 6–8 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 246–250.
6. Mahesha, P.; Vinod, D. Gaussian mixture model based classification of stuttering dysfluencies. J. Intell. Syst. 2016, 25, 387–399.
7. Bayerl, S.P.; von Gudenberg, A.W.; Hönig, F.; Nöth, E.; Riedhammer, K. KSoF: The Kassel State of Fluency Dataset—A Therapy Centered Dataset of Stuttering. arXiv 2022, arXiv:2203.05383.
8. Schuller, B.; Steidl, S.; Batliner, A.; Hirschberg, J.; Burgoon, J.K.; Baird, A.; Elkins, A.; Zhang, Y.; Coutinho, E.; Evanini, K. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In Proceedings of the 17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, CA, USA, 8–12 September 2016; 2020; Volumes 1–5, pp. 2001–2005.
9. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
10. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 2020, 33, 12449–12460.
11. Villegas, B.; Flores, K.M.; Acuña, K.J.; Pacheco-Barrios, K.; Elias, D. A novel stuttering disfluency classification system based on respiratory biosignals. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4660–4663.
12. Kourkounakis, T.; Hajavi, A.; Etemad, A. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6089–6093.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
14. Scarpiniti, M.; Comminiello, D.; Uncini, A.; Lee, Y.-C. Deep recurrent neural networks for audio classification in construction sites. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 810–814.
15. Hochreiter, S.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.

16. Howell, P.; Davis, S.; Bartrip, J. The university college london archive of stuttered speech (uclass). J. Speech Lang. Hear. Res. 2009, 52, 556–569.
17. Kourkounakis, T.; Hajavi, A.; Etemad, A. FluentNet: End-to-end detection of speech disfluency with deep learning. arXiv 2020, arXiv:2009.11394.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
19. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5206–5210.
20. Chen, Q.; Chen, M.; Li, B.; Wang, W. Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 8069–8073.
21. Neubig, G.; Akita, Y.; Mori, S.; Kawahara, T. A monotonic statistical machine translation approach to speaking style transformation. Comput. Speech Lang. 2012, 26, 349–370.
22. Federico, M.; Hwang, M.-Y.; Rödder, M.; Stüker, S. International Workshop on Spoken Language Translation. 2011. Available online: https://aclanthology.org/www.mt-archive.info/10/IWSLT-2011-TOC.htm (accessed on 20 September 2023).
23. Sheikh, S.A.; Sahidullah, M.; Hirsch, F.; Ouni, S. Stutternet: Stuttering detection using time delay neural network. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 426–430.
24. Sheikh, S.A.; Sahidullah, M.; Hirsch, F.; Ouni, S. Machine learning for stuttering identification: Review, challenges and future directions. Neurocomputing 2022, 514, 385–402.