Literature Review: Speech Disfluency Detection and Remediation Using Deep Learning Algorithms.

Jagadish S Jakati[1,2], Abeer Aljohan[1,3].

[1] Lincoln Global Postdoctoral Research (LGPR) Program, Lincoln University College, Petaling Jaya, Malaysia

[2]Postdoc Research Scholar & Senior Assistant professor D Y Patil International University Akurdi, Pune, Maharashtra, India.

[3]Postdoc Research Supervisor & Professor Applied College, Taibah University, Madinah, Saudi Arabia

Email ID.jagadishjs30@gmail.com, aahjohani@taibahu.edu.sa

Abstract: Speech disfluency, including stuttering, repetitions, prolongations, and involuntary pauses, significantly affects verbal communication and quality of life. Recent advances in deep learning have enabled automated techniques for speech disfluency detection using acoustic and linguistic features. However, most existing research focuses primarily on detection accuracy, with limited attention to remediation, real-time deployment, multilingual support, and computational efficiency. This paper presents a systematic review and limitation analysis of existing research works related to speech disfluency detection and remediation. Each study is analyzed based on dataset characteristics, methodology, performance metrics, and limitations. The objective of this review is to identify key research gaps and motivate the development of a comprehensive deep learning framework for real-time speech disfluency detection and remediation.

Keywords: UCLASS stutter; SVM; SEP-28k; StutterNet; wav2vec2

Introduction

Speech disfluency refers to disruptions in the natural flow of speech, such as repetitions, prolongations, and speech blocks. These disorders can have a profound impact on communication effectiveness, emotional well-being, and social interaction. Early and accurate identification of speech disfluency is essential for effective speech therapy and rehabilitation. Conventional assessment methods rely on manual evaluation by speech-language pathologists, which can be subjective, time-consuming, and resource-intensive. With the emergence of deep learning, automated speech analysis systems have been developed to detect disfluency patterns using acoustic and temporal features. Despite promising results, existing approaches often lack real-time capabilities, remediation strategies, personalization, and multilingual applicability. This literature review aims to critically analyze existing research and highlight their limitations to justify the need for a robust deep learning-based speech disfluency detection and remediation framework. [1].

Related work

Payal Mohapatra et al[1]. (2022), proposes a method to improve automatic detection of speech disfluencies such as repetitions, prolongations, and blocks. The study uses the SEP-28k dataset, which contains about 28,000 short speech segments from people who stutter, and evaluates generalization using FluencyBank. To address the problem of noisy annotations and limited high-quality labeled data, the authors apply a data distillation strategy that keeps only samples with full annotator agreement. Contextual speech features are extracted using the pretrained wav2vec 2.0 model, and these embeddings are classified using a CNN-based architecture called DisfluencyNet. The proposed approach improves the robustness and accuracy of speech disfluency detection, particularly in low-resource conditions with limited reliable training data.

Huaijin Deng et al[2]. (2020), presented how different types of speech disfluencies affect automatic fluency evaluation. The study uses the Corpus of Spontaneous Japanese dataset, which contains spontaneous speech recordings with detailed annotations such as filled pauses, word fragments, silence duration, and timing information. Fluency evaluation is formulated as a binary classification task using a Support Vector Machine (SVM) model with prosodic and disfluency-based handcrafted features. Experimental results show that word fragment-based features are more effective than filled pause features for detecting disfluent speech, and combining them with prosodic features further improves performance. However, the approach is limited to Japanese speech data, relies on manual annotations and handcrafted features, and uses traditional machine learning rather than deep learning methods.

Vamshiraghusimha Narasinga et al[3]. (2025), proposes a stutter detection approach that uses Long-Term Average Spectrum (LTAS) features to capture long-term spectral characteristics of speech, which are often missed by short-term features like Mel Frequency Cepstral Coefficients. Experiments are conducted on the SEP-28k dataset and validated using FluencyBank for cross-dataset generalization. The method extracts LTAS features using different filter banks such as Gammatone, Constant-Q Transform, and Single-Frequency filters, and classification is performed using SVM, LSTM, and Bi-LSTM models. Results show that LTAS-based representations outperform traditional MFCC features, with Bi-LSTM combined with CQT-LTAS achieving the best performance, demonstrating improved detection of multiple stuttering types and better generalization across datasets, although the approach focuses only on acoustic features and involves higher computational complexity.

Tedd Kourkounakis et al[4]. (2020), proposes an end-to-end deep learning framework for detecting multiple speech disfluency types directly from acoustic signals without relying on speech-to-text systems. The model is evaluated on the UCLASS dataset and a synthetic dataset called LibriStutter, which is generated from LibriSpeech by inserting artificial disfluencies to increase training diversity. The proposed FluentNet architecture processes Short-Time Fourier Transform spectrograms and combines a Squeeze-and-Excitation Residual Network for spectral feature extraction, Bidirectional LSTM layers for temporal modeling, and a global attention mechanism to highlight important speech segments. Experimental results show that FluentNet achieves state-of-the-art performance with high accuracy and low miss rates, demonstrating strong generalization across real and synthetic datasets, although the approach is limited to English datasets and increases complexity by using separate classifiers for different disfluency types.

Shakeel A. Sheikh et al[5]. (2021) presents an acoustic-based deep learning approach for detecting multiple stuttering behaviors without relying on Automatic Speech Recognition systems. The model is evaluated on the UCLASS stuttering speech dataset and uses Time Delay Neural Network layers with Mel Frequency Cepstral Coefficients features to capture temporal patterns in speech. The proposed StutterNet architecture models long-range temporal dependencies and performs multiclass classification of stuttering types such as repetitions, prolongations, blocks, and fluent speech. Experimental results show improved accuracy and generalization compared to baseline models while using fewer parameters, demonstrating the effectiveness of TDNN-based architectures for stuttering detection.

Amrit Romana, et al. (2024)[6], proposes methods for detecting speech disfluencies directly from audio without relying on transcripts or Automatic Speech Recognition. Experiments are conducted on the Switchboard conversational speech dataset, and three approaches are compared: ASR-based text models, acoustic-only models using pretrained representations such as WavLM, and a multimodal fusion approach combining acoustic and linguistic features. Results show that acoustic-only models outperform ASR pipelines in some cases, while the multimodal BLSTM-based system achieves the best overall performance and robustness under noisy conditions, though it requires higher computational resources and the study is limited to English conversational speech data.

Peter Mihajlik et al[7]. (2024), investigates how different labeling strategies for disfluencies and non-lexical sounds affect the performance of end-to-end Automatic Speech Recognition systems. Experiments are conducted on the multilingual conversational datasets BEA-Base V2 and GRASS using models such as Fast-Conformer and wav2vec2-large. The study compares approaches including removal, unified labeling, and separation of fillers and backchannels. Results show that

removing disfluencies improves recognition accuracy for Hungarian speech, while retaining meaningful backchannels such as “mhm” performs better for Austrian German. Although the proposed labeling strategies improve ASR performance compared to baseline systems, lexical disfluencies remain difficult to recognize and improvements in overall word error rate are relatively modest.

Peter Mihajlik et al.[8], 2024, Proposes multiple disfluency and nonlexical labeling strategies are evaluated within CTC-based end-to-end ASR systems, including removal, unified labeling, and separation of fillers and backchannels. Models are trained using Fast-Conformer for Hungarian and wav2vec2- large for Austrian German. A reference-token-number-aware WER metric is introduced to fairly compare systems producing different output lengths, along with separate error rates for lexical words, disfluencies, and non-lexical sounds. Overall WER improvements are modest, lexical disfluencies remain challenging, and performance varies across languages. Removal-based methods lose conversational meaning. The study focuses on ASR accuracy only and does not address real-time deployment or therapeutic feedback.

Table 1. Comparative Summary of Existing Research Works

Authors	Title	Dataset	Methodology	Limitations
Mohapatra et al[1]., 2022	Speech Disfluency Detection with Contextual Representation and Data Distillation	SEP-28k, FluencyBank	wav2vec 2.0 embeddings + CNN (DisfluencyNet) with data distillation	High computation, weak block detection, no real-time analysis
Deng et al.[2], 2020	Automatic Fluency Evaluation of Spontaneous Speech Using Disfluency-Based Features	CSJ (Japanese)	Disfluency + prosodic features with SVM classifier	Language dependent, transcription-based, no deep learning
Narasinga et al[3]., 2025	Enhancing Stutter Detection using Long-Term Average Spectrum Values	SEP-28k, Fluency Bank	LTAS features (CQT, Gammatone) + Bi-LSTM classification	High computation, acousticonly, no real-time or remediation
Passali et al[4]., 2025	Artificial Disfluency Detection: Uh No, Disfluency Generation for the Masses	Switchboard, Disfl-QA	LARD artificial disfluency generation + supervised learning	Text-only, English-only, no acoustic modeling or remediation

Kourkounakis et al[5]., 2020	FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning	UCLASS, LibriStutter	SE-ResNet + BLSTM + Attention on spectrograms	English-only, synthetic data, no remediation or real-time analysis
Sheikh et al[6]., 2021	StutterNet: Stuttering Detection Using Time Delay Neural Network	UCLASS	TDNN on MFCCs with statistical pooling	Single dataset, fixed window assumption, weak block detection
Romana et al[7]., 2024	Automatic Disfluency Detection From Untranscribed Speech	Switchboard	WavLM acoustic model + BLSTM multimodal fusion	High memory, English-only, restart detection weak
Mihajlik et al[8]., 2024	On Disfluency and Non-lexical Sound Labeling for End-to-End ASR	BEA-Base V2, GRASS	CTC-based ASR with disfluency/non-lexical labeling strategies	CTC-based ASR with disfluency/non-lexical labeling strategies

Key Contribution

The limitations and research gaps identified in the existing literature emphasize the need for a scalable, real-time, and therapy-oriented deep learning framework. The proposed research aims to develop an end-to-end system capable of accurate speech disfluency detection and effective remediation while ensuring computational efficiency, privacy preservation, and multilingual adaptability.

Discussions

Experiments are conducted on the various dataset. The corpus includes monologue recordings from 139 speakers aged between 8 and 18 years, of which 25 speakers are used in this study. Audio is segmented into 4- second clips, resulting in approximately 800 samples. Each segment is manually annotated into six stutter categories: sound repetition, word repetition, phrase repetition, revision, interjection, and prolongation. All audio is resampled to 16 kHz and converted into log-Mel spectrogram representations. The datasets provide both single-language and multilingual scenarios for evaluating editing quality and ASR improvement.

Conclusions

This paper presented a comprehensive review and limitation analysis of existing deep learning-based speech disfluency detection and remediation approaches. By systematically evaluating the shortcomings of current methods, this study establishes a strong foundation for future research focused on developing a robust, real-time, and adaptive speech disfluency detection and remediation framework.

References

1. P. Mohapatra et al., "Speech disfluency detection with contextual representation and data distillation," in ACM IASA, 2022. *https://doi.org/0.1145/3539490.3539601*
2. H. Deng et al., "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in IEEE ICASSP, 2020. *https://doi.org/10.1109/ICASSP40776.2020.9053452*
3. V. Narasinga et al., "Enhancing stutter detection using long-term average spectrum values," in IEEE ICASSP, 2025. *https://doi.org/10.1109/ICASSP49660.2025.10888625*
4. T. Passali et al., "Artificial disfluency detection: Uh no, disfluency generation for the masses," Computer Speech and Language, 2025. *https://doi.org/10.1016/j.csl.2024.101711*
5. T. Kourkounakis et al., "Fluentnet: End-to-end detection of speech disfluency with deep learning," in IEEE ICASSP, 2020. *https://doi.org/10.1109/TASLP.2021.3110146*
6. S. A. Sheikh et al., "Stutternet: Stuttering detection using time delay neural network," arXiv preprint, 2021.*https://doi.org/10.23919/EUSIPCO54536.2021.9616063*
7. A. Romana, K. Koishida, and E. M. Provost, "Automatic disfluency detection from untranscribed speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.*https://doi.org/ 10.1109/TASLP.2024.3485465*
8. P. Mihajlik et al., "On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition," in Interspeech, 2024. *https://doi.org/10.21437/Interspeech.2024-2157*