A Novel Self-Supervised Transformer Model for Speech Disfluency Boundary Detection and Waveform Reconstruction

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 technology is getting better at finding more than solving the problem of stuttering. Disfluency is actually fixing by remediation. The main goal of the paper is a Wav2Vec2 2.0 model fine-tuned on SEP-28k dataset. in which we use a probability threshold of 0.93 to locate boundaries in disfluency like block prolongation, repetition within a 16KHz audio stream. Once the flagged segment is caught it gets cut and the audio is stitched using a 30ms Hamming window crossfade. We run evaluation across noise conditions between 5dB to 25dB SNR. Spectral distortion at signal level is tracked by LPC analysis. For Linguistic Integrity Whisper-Base is used as a secondary auditor with Word Error Rate target below 0.15. Accuracy is 92% of detection and 99% of recall in the listening study of 40 participants. The mean of quality rating is ranging from 2.2 to 4.3 vocal identity and semantics of content remained intact.

Index Terms—Speech Remediation, Wav2Vec 2.0, Surgical Excision, SEP-28k, Disfluency Detection, Concatenative Synthesis, Assistive Technology

---

Introduction

People who stutter often face problems with technology because most systems are made for people who speak fluently [1]. Speech assistants could misinterpret their voice, phone systems can continually offer the same options, and speech-to-text software might produce inaccurate transcripts. For many stutterers, this makes everyday communication challenging. Stuttering can have an impact on mental health, education, employment prospects, social communication, and confidence. People who stammer may experience anxiety or refrain from speaking in public, at work, or in educational settings.Due to their communication issues, stutterers might experience anxiety or refrain from communicating. A speech-language pathologist (SLP) often conducts speech evaluations by listening to speech, identifying speech disorders, and then providing a report. However, this process takes time and is difficult to provide to everyone. In many areas, SLPs are limited, treatment is expensive, and people may speak more fluently in clinical settings, which can hide the real level of stuttering. Most research has focused on detecting and classifying speech disfluencies with high accuracy using machine learning and deep learning models [2]–[9]. However, most systems stop after detection [10]. They can identify speech problems or provide transcripts, but they do not improve the original speech audio. This paper presents a system that tries to improve Identify applicable funding agency here. If none, delete this. speech instead of only detecting problems. Each repetition, prolongation, or speech block is treated as a small problem in the speech signal with a clear start and end point. To precisely identify these speech issue locations, a refined Wav2Vec 2.0

model [11] is implemented. To render speech sound natural and continuous, the detected portion is removed, and the resulting speech is joined using concatenative synthesis with a smooth connection. Two techniques are used to test the system. While Whisper Base [12] determines whether the corrected speech has its original meaning, LPC analysis determines whether speech quality is still good after eliminating disfluencies at various noise levels from 5dB to 25dB SNR. Both tests are crucial since speech that sounds consistent but alters meaning cannot be rated an acceptable response.

Related work

Using Long-Term Average Spectrum (LTAS) analysis, Narasinga et al. (2025) developed a stutter detection system to better accurately identify speech disfluencies. The study extracted speech features like Constant-Q Transform (CQT) and Gammatone from datasets like SEP-28k and FluencyBank in order to investigate stuttering patterns in speech. A Bi-LSTM model was then trained using these features to gradually understand speech patterns. The outcomes outperformed previous short-term speech features. However, the method needed more computation, used only audio information without language information, and did not support real-time use or speech correction. [9]

Maity et al. (2025) developed a deep learning model using CNN, LSTM, and an attention mechanism to improve stuttering detection. They used the SEP-28k dataset and extracted Mel-spectrogram features from speech.Speech patterns were identified by CNN layers while LSTM layers learned how speech changes over time. The attention mechanism helped the model focus more on important stuttering parts in speech, which improved accuracy compared to other models. However, the system mainly focused on two-class classification and did not include real-time testing or testing in multiple languages, which limited its wider use. [8]

Romana et al. (2024) presented a speech disfluency detection method that used only audio features without using speech transcription. The study used the Switchboard dataset and applied WavLM-based speech features with BLSTM layers to learn speech patterns over time. This method reduced the need for speech-to-text systems and avoided transcription errors. The results showed good accuracy for detecting different types of disfluencies. However, the model required high memory, was tested mainly on English speech, and showed weaker performance for restart detection. Also, real-time use was not studied. [6]

Rohit Waddar et al. (2024) proposed a CNN-based model for stutter detection using MFCC speech features from the Libri-Stutter dataset. The system used a one-dimensional CNN model with binary cross-entropy loss to classify speech as fluent or disfluent. The model showed better performance than traditional machine learning methods. However, the study focused only on two-class classification and used a small dataset. It also did not test real-time performance or include speech improvement methods. [5]

Narasinga et al. (2025) developed a stutter detection system using Long-Term Average Spectrum (LTAS) analysis to better identify speech disfluencies. The study used datasets like SEP- 28k and FluencyBank and extracted speech features such as Constant-Q Transform (CQT) and Gammatone to understand stuttering patterns in speech. These features were then used in a Bi-LSTM model to learn speech patterns over time. The results showed better performance than older short-term speech features. However, the method needed more computation, used only audio information without language information, and did not support real-time use or speech correction. [4]

Passali et al. (2025) proposed a method to create simulated disfluencies in order to advance text-based disfluency detecting systems. Using datasets such as Switchboard and Disfl-QA, the study developed the LARD method to incorporate speech disfluencies into clean text data. As a result, more training data for

supervised learning was generated. The technique focused on text patterns such repetitions, filler words, and corrections to improve detection in language processing tasks. [3]

The study makes use of Twitter emotion data and Indian raga music data.While Wav2Vec embeddings were used to understand audio data, TF-IDF and SVM were utilized to identify emotions in text. The invention made music recommendations based on a person's sentiments by integrating speech and text emotions. The dataset was limited and contained only two raga classes, despite the method's usefulness and uniqueness. Also, the study did not focus on speech disfluency detection or real-time performance. [2]

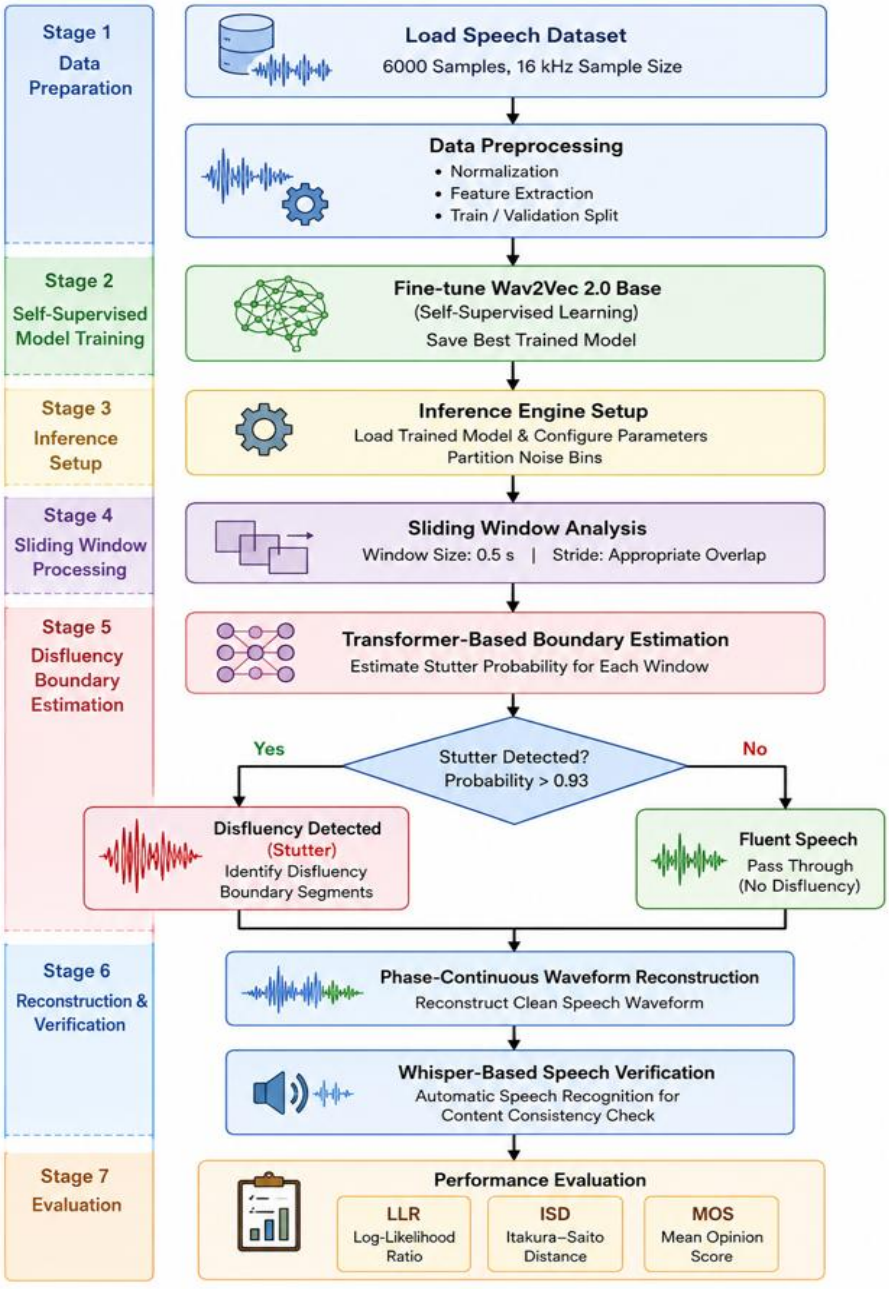
Maity et al. (2025) developed a deep learning model using CNN, LSTM, and an attention mechanism to improve stuttering detection. They used the SEP-28k dataset and extracted Mel-spectrogram features from speech. Speech patterns were identified by CNN layers while LSTM layers learned how speech changes over time. The attention mechanism helped the model focus more on important stuttering parts in speech, which improved accuracy compared to other models. However, the system mainly focused on two-class classification and did not include real-time testing or testing in multiple languages, which limited its wider use. [1]

Romana et al. (2024) presented a speech disfluency detection method that used only audio features without using speech transcription. The study used the Switchboard dataset and applied WavLM-based speech features with BLSTM layers to learn speech patterns over time. This method reduced the need for speech-to-text systems and avoided transcription errors. The results showed good accuracy for detecting different types of disfluencies. However, the model required high memory, was tested mainly on English speech, and showed weaker performance for restart detection. Also, real-time use was not studied. [6]

Key Contribution

Aspect	**Limitation**
Dataset Size and Diversity	Numerous research use synthetic corpora or constrained or domain-specific datasets like UCLASS, SEP-28k, or FluencyBank. There is still little linguistic coverage and little cross-dataset examination.
Language Coverage	Most systems are trained and assessed on English speech, while other research projects investigate Mandarin, Italian, or bilingual data. There is inadequate multilingual benchmarking.
Synthetic Data Dependency	Several approaches depend on synthetic datasets (e.g., LibriStutter) or oversampling techniques such as SMOTE, which may not fully capture natural disfluency variability.
Binary Classification Bias	Many models perform binary classification (fluent vs. disfluent) rather than multi-class detection, limiting fine-grained disfluency modeling.
Real-Time Processing	Most systems do not provide latency analysis, streaming evaluation, or real-time deployment validation despite claiming suitability for practical applications.
Remediation and Therapy Support	All reviewed works focus on detection, classification, or transcription. None integrate corrective feedback, fluency improvement strategies, or therapeutic guidance mechanisms.

Proposed Architecture Flow



The objective of the proposed research is to construct an end-to-end system that can detect speech disfluency accurately and offer successful rehabilitation with respect to computational efficiency, privacy protection, and multilingual adaptability. The methodolgy uses a six-step pipeline that is structured:

Step 1: Data Ingestion and Signal Standardization: Obtaining raw Speech data from the SEP-28k dataset is the initial step in the method. To ensure real-time operability, speech signals endure substantial conditioning. [4]. To introduce varying noise samples based on the quantile distribution, a set-seed shuffle is used to separate a subset of speech sample 6,000 records into five equal samples: Qk = binshuffle(X, seed = 42, k ∈ {5, 10, 15, 20, 25}dB SNR). For processing procedure on the NVIDIA RTX 4050, waveforms are resampled to 16 kHz and normalized to precisely 80,000 samples (5 seconds).

Step 2: Latent Feature Encoding: A multilayer CNN encoder analyzes the standardized speech signal. In order to transformer samples to process longer window seize without getting over memory restrictions, this samples compresses the data by mapping the raw temporal signal into a high-dimensional latent space Z.

Step 3: Neural Detection (Wav2Vec 2.0 Transformer) : The system analyzes long-range dependencies to find samples, prolongations, and repetitions employing an improved Wav2Vec 2.0 Transformer with 12 samples. The classification head employs mean pooling and a Softmax activation to compute the probability of disfluency [6].

Step 4: Surgical Excision & Remediation Logic: The system is design for restoration using a sliding window (0.5s window, 0.1s stride) when confidence probability exceeded the threshold of 0.93 then a surgical trigger is activated.

Step 5: Concatenative Waveform Synthesis: To reconstruct audio without phase discontinuities, fluent segments are joined using a 30ms Hanning windowed cross-fade.

Step 6: Dual-Layer Performance Evaluation

1. Objective Acoustic Physics (12th-Order LPC): Spectral integrity is enhanced using 12th-order Linear Predictive Coding(LPC) [9].
2. Log-Likelihood Ratio (LLR) and Itakura-Saito Distance (ISD) is used to measure distortion of spectral.

Conclusions

Our proposed solution improves the speech audio instead of detecting problems. Our system uses a fine-tuned Wav2Vec 2.0 model with a probability threshold of 0.93. Once the detection
is done, the problem is removed at a very accurate time, 30 ms Hamming-window crossfade is used to joined nearby speech parts using concatenative synthesis.

References

1. C. Constantino, P. Campbell, and S. Simpson, "Stuttering and the social model," Journal of Communication Disorders, vol. 96, p. 106200, 2022, https://doi.org/10.1016/j.jcomdis.2022.106200.
2. S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Stutternet: Stuttering detection using time delay neural network," in 2021 29th European Signal Processing Conference (EUSIPCO). IEEE, 2021, pp. 426–430, https://doi.org/10.23919/eusipco54536.2021.9616063.
3. T. Kourkounakis, A. Hajavi, and A. Etemad, "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2986–2999, 2021, https://doi.org/10.1109/taslp.2021.3110146.
4. M. Jouaiti and K. Dautenhahn, "Dysfluency classification in stuttered speech using deep learning for real-time applications," in ICASSP 2022– 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 6482–6486, https://doi.org/10.1109/icassp43922.2022.9746638.
5. J. Liu, A. Wumaier, D. Wei, and S. Guo, "Automatic speech disfluency detection using wav2vec 2.0 for different languages with variable lengths," Applied Sciences, vol. 13, no. 13, p. 7579, 2023, https://doi.org/10.3390/app13137579.
6. A. Romana, K. Koishida, and E. M. Provost, "Automatic disfluency detection from untranscribed speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 4727–4740, 2024, https://doi.org/10.1109/taslp.2024.3485465.
7. R. Waddar, V. Rathod, S. Chikkamath, N. SR, and S. V. Budihal, "A cnn-based stutter detection using mfcc features with binary cross-entropy loss function," in 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), vol. 1. IEEE, 2024, pp. 1–6, https://doi.org/10.1109/inc460750.2024.10649122.
8. M. Maity, I. Gupta, and D. K. Vishwakarma, "Stuttering detection using lstm and lstm-attention based convolutional neural network," in 2025 10th International Conference on Signal Processing and Communication (ICSC). IEEE, 2025, pp. 579–583, https://doi.org/10.1109/icsc64553.2025.10968952.
9. V. Narasinga, P. Kommagouni, S. Vanga, K. Motepalli, P. Barche, and A. K. Vuppala, "Enhancing stutter detection using long-term average spectrum values," in ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025,pp. 1–5, https://doi.org/10.1109/icassp49660.2025.10888625.
10. I. Sindhu and M. S. Sainin, "Automatic speech and voice disorder detection using deep learning—a systematic literature review," IEEE Access, vol. 12, pp. 49 667–49 681, 2024, https://doi.org/10.1109/access.2024.3371713.