Investigation of Self-Supervised Speech Models for Stuttered Speech Detection

Md Dilwar Alam[1], Deepti Gupta[2]
[1]University Institute of Engineering and Technology
[2]Panjab University Chandigarh Chandigarh, India

Abstract—A speech condition called stuttering is typified by irregularities in speech fluency, such as repetitions, blocks, and prolongations. Speech-language pathologists' (SLPs') manual evaluations, which take a lot of time and need specialized knowledge, are a major component of traditional diagnosis. This study explores utterance-level stuttering detection using self-supervised learning (SSL) models to facilitate automated evaluation. We specifically assess how well a number of pretrained SSL speech models perform on utterance-level stuttering categorization tasks: WavLM Base, HuBERT Base, Wav2Vec 2.0 Base, WavLM Large, HuBERT Large, and Wav2Vec 2.0 Large. The Kassel State of Fluency (KSoF) dataset, FluencyBank, and SEP-28K are used for independent testing, and the models are refined using these datasets. F1 scores for various stuttering types are used to gauge performance. All three test sets (SEP-28K, FluencyBank, and KSoF) have the following F1 values: WavLM Base (0.797, 0.800, 0.772), HuBERT Base (0.790, 0.790, 0.766), Wav2Vec 2.0 Base (0.778, 0.782, 0.758), WavLM Large (0.832, 0.832, 0.758), HuBERT Large (0.817, 0.816, 0.788), and Wav2Vec 2.0 Large (0.804, 0.803, 0.779).WavLM Large continuously performs the best on utterance-level benchmarks out of all the models. This comparison study demonstrates how well SSL models identify stuttering and offers information about how they may be used in actual speech pathology and fluency disorder evaluation.

Keywords—Sel-Supervised Learning, Utteranc-Level Stuttering Detection, Feature Extraction.

1. INTRODUCTION

A neurodevelopmental speech disease, stuttering causes involuntary repetitions, prolongations, blocks, and interjections that interfere with speech's natural flow. For millions of people globally, it has a major impact on professional development, academic achievement, and social engagement. Stuttering affects about 80 million people worldwide, but access to clinical evaluation is still restricted because there aren't enough qualified speech-language pathologists (SLPs) [1,2]. The necessity for automated screening methods to enhance early identification and management is further highlighted by the fact that many SLPs express a lack of confidence in their ability to cure fluency issues [2].

Stuttering event detection and classification have advanced significantly as a result of the use of machine learning (ML) in speech pathology. Early attempts used manually extracted characteristics like linear prediction cepstral coefficients (LPCCs) or Mel-frequency cepstral coefficients (MFCCs) [3]. The generalizability of these conventional approaches across speakers and languages is constrained, especially when it comes to spontaneous or therapy-influenced speech [4].

Speech processing challenges have been transformed by recent developments in self-supervised learning (SSL). By learning contextualized speech representations from unlabeled data, models like Wav2Vec 2.0, HuBERT, and WavLM have shown higher performance in speaker verification, automatic speech recognition (ASR), and fluency analysis [5,6]. Additionally, research has demonstrated that SSL characteristics are more effective than conventional acoustic features at detecting stuttering, especially in noisy or low-resource environments [7, 8].

Although ASR and utterance-level classification have been used in previous research to detect stuttering [9], word-level annotation is necessary in clinical situations because clinicians usually evaluate stuttering at the level of individual words or syllables[2]. Additionally, multi-label classification is a more realistic technique because real-world speech frequently involves overlapping or co-occurring stuttering kinds [10].

This study examines the effectiveness of six SSL models— WavLM, HuBERT, and Wav2Vec 2.0—for utterance-level stuttering detection in both Base and

Large configurations in order to overcome these constraints. Block, Prolongation, Sound Repetition, Word Repetition, and Interjection are the five typically annotated stuttering kinds that we concentrate on. These types are labeled in benchmark datasets as SEP- 28k [11], FluencyBank Timestamped [9], and KSoF [4]. We assess generalization across SEP-28k, FluencyBank, and KSoF separately and refine each model on a combined training set of SEP-28k and FluencyBank.

This study compares self-supervised learning models (Wav2Vec 2.0, HuBERT, and WavLM in base and large configurations) for stuttering detection in order to fill in these gaps. This work's primary contributions are:

1. Model Comparison: We assess how well six SSL-based methods detect stuttering.
2. Dataset Combination: To evaluate generalizability, we test on SEP-28k, FluencyBank, and KSoF after training on a combined dataset (SEP-28k + FluencyBank).
3. Fine-grained Detection: To enhance clinical relevance, we categorize five different types of disfluencies and present cross-type performance.

2. RELATED WORK

Research on automatic stuttering detection has progressed from rule-based techniques to deep learning and data-driven strategies. MFCC and LPCC, two manually constructed acoustic features, were crucial to early systems' ability to identify speech dysfluencies [12,13]. Although successful in controlled settings, these techniques had issues with scalability and generalization.

The field has made great progress since the introduction of big annotated datasets. More than 28,000 clips with annotations for five different disfluency types—blocks, prolongations, sound repetitions, word repetitions, and interjections—were added by SEP-28k [11] and its extension [14]. Previous datasets were expanded by FluencyBank Timestamped [9], which offered word-level timing data crucial for clinical applications. Furthermore, therapy-based recordings with thorough labels, including speech alterations pertinent to tracking therapeutic progress, were provided by KSoF [4].

A varied stuttering speech dataset comprising Indian languages with demographic metadata and both read and spontaneous speech recordings was introduced by Project Boli in order to close the multilingual gap [1]. Training more reliable stuttering detection models has been made possible by these datasets.

Deep learning models like Bi-LSTMs have outperformed conventional machine learning techniques like SVMs and random forests in terms of modeling methods because they are better at capturing temporal relationships in speech. To classify four dysfluency classes in real-time, for example, [15] integrated MFCCs with phoneme probabilities.

The detection of stuttering has changed with the advent of self-supervised learning (SSL). Wav2Vec 2.0, HuBERT, and WavLM are examples of SSL models that are refined for downstream tasks after being trained on vast amounts of unlabeled speech. A word-level SSL-based model was presented by Shih et al. [2] and performed better than earlier utterance-level methods. On the KSoF dataset, Sheikh et al. [8] also used Wav2Vec 2.0 embeddings and showed better accuracy than conventional baseline models.

A 10% relative gain in ASR accuracy was attained by Arunkumar et al. [5] by the use of ensemble approaches that included features from Wav2Vec 2.0, HuBERT, and WavLM in order to further improve performance. In another work, Javanmardi et al. [16] improved the classification of dysarthria severity using early and final layer embeddings of Wav2Vec 2.0, demonstrating the adaptability of SSL characteristics for various speech disorders.

Since stuttering kinds frequently co-occur, multi-label classification has also drawn interest. An attention-based head Wav2Vec 2.0 system was presented by Bayerl et al. [10] for multi-label stuttering detection across several languages and corpora. By hierarchically combining transcription and detection, Lian et al. [17] suggested an unconstrained dysfluency modeling (UDM) framework that lessens the need for manual annotation.

Multilingual and cross-corpus evaluation are still essential. Wav2Vec 2.0 was fine-tuned on therapy-altered speech by Bayerl et al. [18], who showed better generalization across situations. In order to capture speaker-specific patterns and disfluency boundaries, Mohapatra et al. [19] employed CNNs and bidirectional RNNs to concentrate on contextual cues. These patterns are emphasized in thorough evaluations like Sheikh et al. [12], which also advocate for more uniform standards.

In this study, we expand on these foundations by testing three benchmark datasets for utterance-level identification of five stuttering kinds using three different SSL models: WavLM, HuBERT, and Wav2Vec 2.0. In an effort to increase practical applicability, we also do ablation analysis to comprehend the impact of various model characteristics and dataset combinations.

3. PROPOSED METHOD

This study investigates the use of self-supervised learning (SSL) models for stuttering detection on three important datasets: KSoF, FluencyBank, and SEP-28k. Every dataset makes a distinct contribution to the creation and assessment of stuttering detection systems. SEP-28k offers comprehensive annotations of English speech stutter kinds. The model's generalizability is improved by FluencyBank's timestamped, real-world speech data from adults and children. KSoF adds linguistic diversity and robustness to the evaluation process; it was created for Arabic stuttering detection.

The study makes use of SSL models in base and large configurations, including WavLM, HuBERT, and Wav2Vec 2.0. These models are optimized for stuttering classification tasks after being pre-trained on a significant amount of unlabeled speech. The study assesses generalizability and adaptability by comparing model performance across datasets. Techniques for feature extraction are essential to the methodology, especially when utilizing the averaged hidden states or the final hidden layer.

Based on a thorough literature review, this study contrasts more contemporary SSL-based approaches with more conventional techniques like CNN/LSTM-based architectures, ZCR, and MFCCs. This study assesses generalization across languages and domains in a new way, in contrast to previous publications that usually concentrate on a single language or dataset. It proves the effectiveness of SSL in this area by showing that models such as WavLM-Large get higher F1 scores and greater generalization on unknown data.

Utterance Level Stuttering Classification

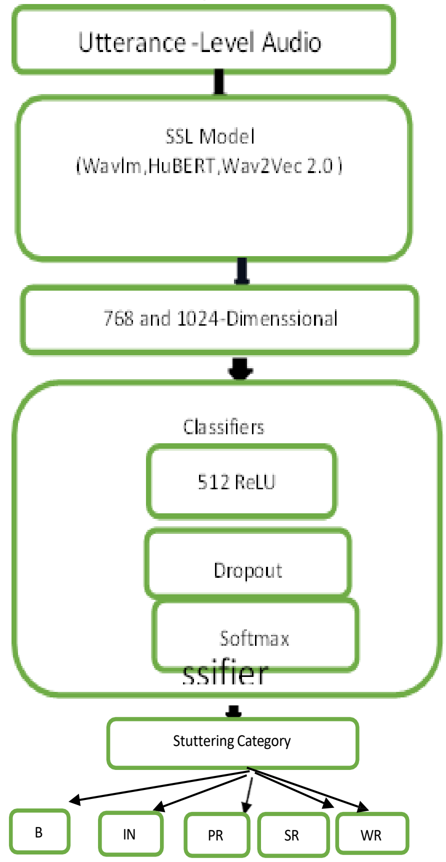

Figure 1 Flowchart of stutter type classification

The findings highlight how crucial it is to develop more inclusive and reliable speech technology by utilizing big, varied datasets and contemporary designs. Additionally, integrating cross-linguistic datasets like KSoF demonstrates how these models can be used in multilingual real-world applications. This study helps create more accessible speech interfaces for PWS in addition to advancing automatic stuttering detection.

4 RESULTS AND DISCUSSIONS

The performance of several self-supervised speech models on the stuttering event detection challenge is presented in this section. The SEP-28k and FluencyBank datasets were combined to train each model. The SEP-28k, FluencyBank, and KSoF datasets were tested independently to assess the trained models' capacity for generalization.

4.1 Using WavLM-Base

The SEP-28k and FluencyBank datasets, which together comprise 32,321 labeled audio segments, were used to train the WavLM-Base model. Three test sets—

SEP-28k, FluencyBank, and KSoF—were used to evaluate the trained model's in-domain performance as well as its cross-dataset generalization capabilities. The WavLM-Base model had high detection skills on data comparable to the training distribution, as evidenced by the experimental findings, which demonstrated an overall F1- score of 0.80 on the SEP-28k test set. The model demonstrated its resilience in managing a variety of speech patterns by maintaining a similar performance with an F1- score of 0.80 when evaluated on FluencyBank. Despite a minor performance decline that was anticipated given the variations in recording conditions and speaker demographics, the model nevertheless did rather well on the cross-dataset test using the KSoF corpus, achieving an overall F1-score of 0.77.

In every stuttering category—blocks, interjections, prolongations, sound repetitions, and word repetitions—the model continuously demonstrated strong performance. The fact that this consistency was seen in all three datasets. Table 1 shows a breakdown of the F1-scores for each type of stuttering as well as overall F1-score.

4.2 Using HuBERT-Base*:*

The combined SEP-28k and FluencyBank datasets, which contained 32,321 tagged audio segments, were used to train the HuBERT-Base model. SEP-28k, FluencyBank, and KSoF were the three different testing datasets used to assess this model's cross-corpus generalization and in-domain classification performance.

The model's total F1-score on both the SEP-28k and FluencyBank test sets was 0.79, indicating consistent performance. When trained on a sufficiently large and diverse dataset, this shows that the model can generalize across recordings from various sources. The HuBERT-Base model demonstrated good generalization across variations in recording settings and speaker populations when tested on the unseen KSoF dataset, achieving an F1-score of 0.77.

In every stuttering category—blocks, interjections, prolongations, sound repetitions, and word repetitions—the model continuously demonstrated performance. The fact that this consistency was seen in all three test datasets. Table 1 shows a breakdown of the F1-scores for each type of stuttering as well as overall F1-score.

4.3 Using Wav2Vec 2.0 -Base

The combined SEP-28k and FluencyBank datasets, which included 32,321 tagged stuttering and non-stuttering segments, were used to train the Wav2Vec 2.0 Base model. The model's performance within the dataset and its ability to generalize to new data were subsequently assessed using three testing datasets: SEP-28k, FluencyBank, and the external KSoF corpus. The model consistently identified stuttering events across recordings with identical labeling structures, as evidenced by its F1-score of 0.78 on the SEP-28k and FluencyBank testing datasets. The model retained a reasonable F1 score of 0.76 for the external KSoF dataset, indicating that it can generalize even to audio recorded in various situations.

For Block, Prolongation, Interjection, Sound Repetition, and Word Repetition events, the model continuously demonstrated high detection capabilities in terms of stuttering type-specific performance across all three datasets. This demonstrates the model's ability to handle a variety of stuttering patterns with resilience and its potential for use in automated stuttering detection systems. Table 1 shows a breakdown of the F1-scores for each type of stuttering as well as overall F1-score.

Table 1:Performance of Stuttering Detection for Base Model

Training data:SEP-28K+FluencyBank (size:32321)							
		F1-Score					
MODEL	Test Dataset	Over All	B	IR	PR	SR	WR
WavLm	SEP-28k	0.797	0.82	0.78	0.81	0.71	0.74
	FluencyBank	0.800	0.83	0.83	0.80	0.80	0.78
	KSoF	0.772	0.80	0.80	0.78	0.70	0.71
HuBERT	SEP-28K	0.790	0.82	0.78	0.79	0.78	0.74
	FLuencyBank	0.790	0.80	0.77	0.80	0.80	0.76
	KSoF	0.766	0.80	0.79	0.77	0.71	0.71
Wav2v ec 2.0	SEP-28K	0.778	0.80	0.76	0.79	0.76	0.72
	FluencyBank	0.782	0.80	0.76	0.78	0.77	0.78
	KSOF	0.758	0.79	0.78	0.75	0.69	0.71

4.4 Using WavLM-Large*:*

The WavLM-Large model was trained using 32,321 labeled speech samples from the SEP-28K and FluencyBank combined dataset. SEP-28K, FluencyBank, and KSOF were the three different testing datasets used to assess the model's performance.

According to the evaluation results, the model obtained an F1-score of 0.832 on the FluencyBank dataset, 0.808 on the KSOF dataset, and 0.832 on the SEP-28K dataset. The model's remarkable generalization skills across in-domain and out-of-domain datasets are demonstrated by these results.

The WavLM-Large model consistently demonstrated good F1 scores for stuttering type-wise performance across all three datasets, especially for Block, Interjection, Prolongation, Sound Repetition, and Word Repetition stuttering events. The model's ability to handle a range of stuttering traits in real-world speech samples is further supported by this. The breakdown of these results is presented in Table 2.

4.5 Using HuBERT-Large:

A combined dataset of SEP-28K and FluencyBank, which included 32,321 audio samples with stuttering labels, was used to train the HuBERT-Large model. Three datasets— SEP-28K, FluencyBank, and KSOF—were independently tested to evaluate the model's capacity for generalization.

On the SEP-28K dataset, the model's F1-score was 0.817; on FluencyBank, it was 0.816; and on the KSOF dataset, it was 0.788. These findings show that the model continues to function dependably when tested on additional datasets like KSOF in addition to the observed data distributions.

For the basic stuttering categories of Block, Interjection, Prolongation, Sound Repetition, and Word Repetition, the HuBERT Large model consistently shown strong identification abilities across all three datasets, attaining high F1-scores, as show results in Table 2.

4.6 Using Wav2Vec 2.0-Large:

The Wav2Vec 2.0-Large model was trained using 32,321 labeled samples from a combined dataset that included the SEP-28K and FluencyBank corpora. To determine how well the model generalized across various speech sources, it was independently tested on three test datasets: SEP-28K, FluencyBank, and KSOF.

The model's F1-scores on the SEP-28K test set were 0.804, FluencyBank was 0.803, and KSOF was 0.779. The aforementioned findings validate that the Wav2Vec 2.0- Large model exhibits consistent and dependable performance on both in-domain and external datasets, including KSOF.

The model also achieved strong F1-scores and demonstrated dependable detection performance across the primary stuttering types, including Block, Interjection, Prolongation, Sound Repetition, and Word Repetition. The results in Table 2 demonstrate its effectiveness in real-world stuttering identification tasks.

Table 2: Performance of Stuttering Detection for Large Model

Training data:SEP-28K+FluencyBank (size:32321)							
		F1-Score					
MODEL	Test Dataset	Over All	B	IR	PR	SR	WR
WavLm	SEP-28k	0.832	0.85	0.82	0.84	0.81	0.78
	FluencyBank	0.832	0.87	0.82	0.83	0.83	0.81
	KSoF	0.808	0.84	0.83	0.81	0.73	0.75
ıBER T	SEP-28K	0.817	0.84	0.80	0.83	0.80	0.76
	FLuencyBank	0.816	0.85	0.80	0.82	0.81	0.79
	KSoF	0.788	0.82	0.81	0.79	0.77	0.73
Wav2vec 2.0	SEP-28K	0.804	0.83	0.79	0.81	0.79	0.75
	FluencyBank	0.803	0.83	0.79	0.80	0.80	0.78
	KSOF	0.779	0.81	0.79	0.78	0.71	0.72

A comparison of deep learning models for stuttering detection across five disfluency classes is presented in this article. To improve diversity and robustness, models were trained using the combined SEP-28k and FluencyBank datasets. To evaluate generalizability, the SEP-28k, FluencyBank, and KSoF test sets were used. Because of its balance, the F1 score served as the main performance indicator. The findings indicate that WavLM Large fared better than the other models, particularly when it came to identifying word repetition and extension. However, because they are nuanced and speaker-specific, interjections and sound repeats presented difficulties. The significance of model scale and pretraining data in managing stuttering variability is demonstrated by the consistent superior performance of larger SSL models, such as HuBERT Large and Wav2Vec 2.0 Large, over their base equivalents.

5. CONCLUSIONS

In this work, six self-supervised learning (SSL) models for utterance-level stuttering detection—WavLM, HuBERT, and Wav2Vec 2.0 in Base and Large configurations—are thoroughly evaluated. To verify model performance, we used lightweight classifiers to evaluate representations taken from pretrained SSL models on three benchmark datasets: KSoF, FluencyBank, and SEP-28k. WavLM Large continuously produced the highest F1 scores of any model, exhibiting excellent generalization in both in-domain and cross-domain contexts. The findings demonstrate that, even in the absence of fine-tuning, massive SSL designs have the capacity to record strong acoustic patterns associated with speech disfluencies. This study provides a solid foundation for upcoming cross-corpus stuttering research and highlights the efficacy of SSL-based methods in the identification of fluency disorders. Future research will include multilingual adaptability, frame-level disfluency detection, and integration with clinical evaluation.

REFERENCE

[1] A. Batra, M. narang, N. K. Sharma, and P. K. Das, "Boli: A dataset for understanding stuttering experience and analyzing stuttered speech," 2025, [Online]. Available: http://arxiv.org/abs/2501.15877

[2] Y.-J. Shih, Z. Gkalitsiou, A. G. Dimakis, and D. Harwath, "Self-superviesed Speech Models for Word-Level Stuttered Speech Detection," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.10704

[3] O. Chia Ai, M. Hariharan, S. Yaacob, and L. Sin Chee, "Classification of speech dysfluencies with MFCC and LPCC features," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2157–2165, 2012, doi: 10.1016/j.eswa.2011.07.065.

[4] S. P. Bayerl, A. W. von Gudenberg, F. Hönig, E. Nöth, and K. Riedhammer, "KSoF: The Kassel State of Fluency Dataset - A Therapy Centered Dataset of Stuttering," *2022 Lang. Resour. Eval. Conf. Lr. 2022*, no. June, pp. 1780–1787, 2022.

[5] A. Arunkumar, V. N. Sukhadia, and S. Umesh, "Investigation of Ensemble features of Self-Supervised Pretrained Models for Automatic Speech Recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, pp. 5145–5149, 2022, doi: 10.21437/Interspeech.2022-11376.

[6] S. W. Yang *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 4, pp. 3161–3165, 2021, doi:10.21437/Interspeech.2021-1775.

[7] R. V, R. Chainani, S. Mehrotra, S. Sah, and S. Mahajan, "Evaluative comparison of machine learning algorithms for stutter detection and classification," *MethodsX*, vol. 13, no. October, p. 103050, 2024, doi: 10.1016/j.mex.2024.103050.

[8] S. A. Sheikh, M. Sahidullah, S. Ouni, and F. Hirsch, "End-to-End and Self-Supervised Learning for ComParE 2022 Stuttering Sub-Challenge," *MM 2022 - Proc. 30th ACM Int. Conf. Multimed.*, pp. 7104–7108, 2022, doi: 10.1145/3503161.3551588.

[9] A. Romana, M. Niu, M. Perez, and E. M. Provost, "FluencyBank Timestamped: An Updated Data Set for Disfluency Detection and Automatic Intended Speech Recognition," *J. Speech. Lang. Hear. Res.*, vol. 67, no. 11, pp. 4203–4215, 2024, doi: 10.1044/2024_JSLHR-24-00070.

[10] S. P. Bayerl *et al.*, "A Stutter Seldom Comes Alone-Cross-Corpus Stuttering Detection as a Multi-label Problem," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-Augus, no. August 2023, pp. 1538–1542, 2023, doi: 10.21437/Interspeech.2023-2026.

[11] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28K: A dataset for stuttering event detection from podcasts with people who stutter," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, no. Lm, pp. 6798–6802, 2021, doi: 10.1109/ICASSP39728.2021.9413520.

[12] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Machine learning for stuttering identification: Review, challenges and future directions," *Neurocomputing*, vol. 514, pp. 385–402, 2022, doi: 10.1016/j.neucom.2022.10.015.

[13] T. Kourkounakis, A. Hajavi, and A. Etemad, "FluentNet: End-to-End Detection of Stuttered Speech Disfluencies with Deep Learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2986–2999, 2021, doi: 10.1109/TASLP.2021.3110146.

[14] S. P. Bayerl, D. Wagner, E. Nöth, T. Bocklet, and K. Riedhammer, "The Influence of Dataset Partitioning on Dysfluency Detection Systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13502 LNAI, pp. 423–436, 2022, doi: 10.1007/978-3-031-16270-1_35.

[15] M. Jouaiti and K. Dautenhahn, "Dysfluency Classification in Stuttered Speech Using Deep Learning for Real-Time Applications," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, no. February, pp. 6482–6486, 2022, doi: 10.1109/ICASSP43922.2022.9746638.

[16] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-Based Detection and Severity Level Classification of Dysarthria From Speech," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2023-June, no. 1, 2023, doi: 10.1109/ICASSP49357.2023.10094857.

[17] J. Lian *et al.*, "Unconstrained Dysfluency Modeling for Dysfluent Speech Transcription and Detection," *2023 IEEE Autom. Speech Recognit. Underst. Work. ASRU 2023*, pp. 1–8, 2023, doi: 10.1109/ASRU57964.2023.10389771.

[18] S. P. Bayerl, D. Wagner, E. Nöth, and K. Riedhammer, "Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2022-Septe, pp. 2868–2872, 2022, doi:10.21437/Interspeech.2022-10908.

[19] P. Mohapatra, A. Pandey, B. Islam, and Q. Zhu, "Speech Disfluency Detection with Contextual Representation and Data Distillation," *IASA 2022 - Proc. 2022 1st ACM Int. Work. Intell. Acoust. Syst. Appl.*, pp. 19–24, 2022, doi: 10.1145/3539490.3539601.