Interspeech 2024
1-5 September 2024, Kos, Greece

MMSD-Net: Towards Multi-modal Stuttering Detection

Liangyu Nie[1], *Sudarsana Reddy Kadiri*[2], *Ruchit Agrawal*[3]

[1]University of Texas at Dallas, USA
[2]University of Southern California, USA
[3]University of Birmingham, UAE

liangyu.nie@utdallas.edu,skadiri@usc.edu,r.r.agrawal@bham.ac.uk

Abstract

Stuttering is a common speech impediment that is caused by irregular disruptions in speech production, affecting over 70 million people across the world. Standard automatic speech processing tools do not take speech ailments into account and are thereby not able to generate meaningful results when presented with stuttered speech as input. The automatic detection of stuttering is an integral step towards building efficient, context-aware speech processing systems. While previous approaches explore both statistical and neural approaches for stuttering detection, all of these methods are uni-modal in nature. This paper presents MMSD-Net, the first multi-modal neural framework for stuttering detection. Experiments and results demonstrate that incorporating the visual signal significantly aids stuttering detection, and our model yields an improvement of 2-17% in the F1-score over existing state-of-the-art uni-modal approaches.
Index Terms: Speech disorders, Stuttering detection, Multi-modal neural networks, Transformer

1. Introduction

Recent advancements in machine learning have enabled a wide range of AI applications across a myriad of sectors such as government, industry, healthcare and transportation. The automatic recognition and transcription of speech is an integral task in machine learning, which enables users to interact seamlessly with machines, lending itself applicable to a variety of tasks such as automatic translation of speech, subtitling of multimedia content, audio signal analysis and editing, and so on. In particular, the massive developments in speech processing have given rise to an unprecedented surge of virtual/digital assistants such as Siri and Alexa, used by millions of users across the globe. However, these modern speech processing tools are not perfect, and are unable to deal with a range of speech impediments.

Stuttered speech is one such common speech impairment, wherein the production of fluent speech is hindered by a malfunctioning of the central nervous system of the afflicted person. This makes it highly challenging for people who stutter to access popular speech recognition tools such as Siri and Alexa. As an example, Apple's Siri reported an accuracy ranging from 18.2-73% when given stuttered speech as input, compared to a high accuracy of 92% when presented with normal speech as input [1], which makes it impractical to be used by people who stutter. Stuttering can manifest itself in various types of disfluencies, such as sound repetition, part-word repetition, word repetition, phrase repetition, revision, interjection, prolongation and block [2]. The lack of robust speech recognition of stuttered speech is an unfair consequence for a significant portion of the world population, with over 70 million people being affected by this condition [3].

Automatic stuttering detection is an integral step towards mitigating this limitation of existing speech processing tools. While a number of approaches have been proposed in the recent years for stuttered speech detection, these are either audio-based or text-based, and therefore *uni-modal* in nature. The application of multi-modal neural networks for stuttering detection remains unexplored. This paper aims to bridge this gap and presents the first multi-modal neural network framework for stuttered speech detection. The primary motivation behind our proposed multi-modal approach is that cues to detect stuttering could be found not only in the audio signal, but also on the speakers' faces. Our hypothesis is that these visual signals contain relevant information pertitent to the SD task. We validate our hypothesis by conducting experiments across a range of settings and demonstrating results on publicly available datasets.

The primary contributions of this paper are summarized below:

- We present MMSD-Net, the first multi-modal neural network for automatic stuttering detection.
- Our proposed architecture effectively integrates audio, video, and language data using a novel multi-modal fusion mechanism, which enhances feature fusion for superior multi-modal task performance.
- We conduct experimental studies using publicly available datasets and present extensive comparisons of the results obtained by our model against the state-of-the-art uni-modal methods for stuttering detection.
- We demonstrate that MMSD-Net outperforms state-of-the-art uni-modal methods by 2-17 % on F1-score.
- We release the code for pre-processing, post-processing as well as the neural network models publicly to ensure reproducibility of our research.

2. Related Work

The automatic detection of stuttering has been approached through various methodologies, primarily divided into statistical approaches and deep learning approaches. Statistical methods rely on feature extraction from speech signals and subsequent classification using machine learning algorithms. These approaches often employ features such as autocorrelation functions, spectral measures, and Mel-frequency cepstral coefficients (MFCCs), utilizing classifiers like support vector machines (SVM), hidden Markov models (HMM), and artificial neural networks (ANN) [4, 5, 6, 7]. While statistical methods have shown promising results, they often require manual feature engineering and may not generalize well across different

10.21437/Interspeech.2024-1497

datasets due to their dependency on specific feature sets and classifiers. For example, SVMs may struggle with datasets containing imbalanced classes or noisy features, limiting their performance in real-world scenarios [8]. Similarly, HMMs have been successful in modeling temporal sequences in speech, but they may struggle with capturing complex dependencies in stuttered speech, particularly in distinguishing between different types of disfluencies [7].

The advancements in deep learning have led to the exploration of deep neural networks for stutter detection [9, 10, 11]. Recent approaches have explored the usage of the perceptron model [10] and autoregressive models such as Long-Short Term Memory networks (LSTMs) [11, 12]. While the former employs LSTMs with integer linear programming [13], the latter employs bidirectional LSTMs with attention using the MFCC features. Authors in [9] propose a CNN-based model to learn stutter-related features, formulating SD as a binary classification problem. The only input features used in this study are the spectrograms. Another approach called FluentNet [14] builds upon this method and explores the use of a residual network along with an LSTM network to learn frame-level representations. However, these methods only consider a small subset of disfluent speakers in their studies, and are not tested exhaustively on their ability to generalize well to a variety of stuttered speakers.

With the advent of the Transformer model [15, 16] propose using a controllable time-delay transformer architecture, but only for Chinese data. A similar approach proposed recently, called StutterNet [17], employs a time delay neural network and formulates the stutter detection task as a multi-class classification problem. However, this method only consider limited disfluent behaviours (blocks, repetition, and prolongation) in addition to fluent speech segments.

A major challenge to train deep neural networks is the availability of large-scale annotated datasets. To address this problem, [18] curate a large-scale dataset for stuttering detection, called SEP-28k. They also present experiments using the ConvLSTM model and demonstrate results on the FluencyBank as well as SEP-28k datasets. More recently, [19] employ the Wav2Vec model for stuttering detection and explore self-supervised learning for this task. The closest to our work is the method called *FluentSpeech* proposed by [20]. However, it is uni-modal and does not leverage the visual signal. Additionally, it considers pauses, non-lexical vocalisations and interjections such as *so, hmmm, umm, like* in speech as stuttered segments [21]. We disagree from their method in that such normal disfluencies could correspond to useful pauses in speech wherein the speaker can plan their upcoming discourse; and are notably different from *stuttering*, which is a neuro-developmental speech disorder that is characterized by core behaviour and corresponds to abnormally persistent stoppages in normal speech, often accompanied by unusual behaviours such as quick eye blinks, lip tremors and nodding of head [22].

It must be noted that while there have been several approaches proposed for stuttering detection, these rely on classical machine learning (typically on the audio signal) or unimodal deep learning methods. The application of multi-modal deep learning remains unexplored for this task. We bridge this gap and present MMSD-Net, the first multi-modal deep learning method for stutter event detection.

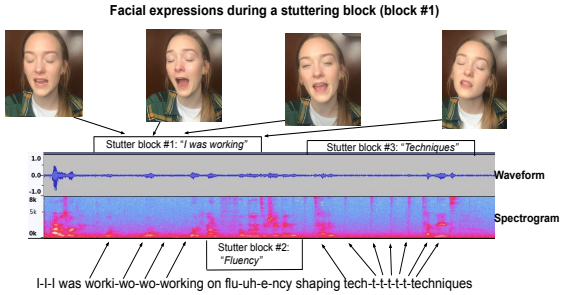

Figure 1: *Illustration of multi-modal cues during stuttering. Sentence: "I was working on fluency shaping techniques."*

3. Proposed Methodology

This section presents a detailed overview of MMSD-Net. Technically, stutter presents itself as an audiovisual problem. Figure 1 demonstrates a data sample from the *Adults Who Stutter* dataset [23], depicting a snippet of the video, and the corresponding waveform, spectrogram and textual transcription respectively. The primary motivation of our method is that cues for stuttering can be found from the visual signal as well as the audio signal, as illustrated in Figure 1. It can be observed that facial expressions change during stuttering and can thereby provide important contributions for the automatic stutter detection task. To this end, we employ the visual signal in addition to the audio signal, and present MMSD-Net, the first multi-modal neural approach towards stuttering detection.

Model architecture: Figure 2 presents an overview of our model architecture. MMSD-Net comprises three primary modules, described in the subsequent subsections:

3.1. Multi-encoder module

Current Multimodal Language Models are mainly geared towards understanding video and text-based content. MMSD-Net incorporates specialized modality encoders to process not just video and text, but also auditory data. These encoders are designed to extract the most relevant features from each modality. This upgrade significantly enhances MMSD-Net's ability to process and interpret information across various modalities efficiently. Unlike traditional models that might rely on Convolutional Neural Networks (CNNs), MMSD-Net leverages three transformer encoders for video, audio and text data respectively. Transformers excel at capturing long-range dependencies within sequences, making them well-suited for stuttering detection. By capturing these nuanced details, MMSD-Net gains a richer understanding of the input data corresponding to different modalities, allowing for more comprehensive reasoning. The specialized encoders process the video (**v**), audio (**a**) and textual (**t**) inputs as follows:

$$\mathbf{h}_v = \text{Model}(\mathbf{v}) \in \mathbb{R}^{d_v} \quad (1)$$

$$\mathbf{h}_a = \text{Model}(\mathbf{a}) \in \mathbb{R}^{d_a} \quad (2)$$

$$\mathbf{h}_t = \text{Model}(\mathbf{t}) \in \mathbb{R}^{d_t} \quad (3)$$

where $\mathbf{h}_v$, $\mathbf{h}_a$, and $\mathbf{h}_t$ represent the extracted features for video, and audio, respectively. d_v, d_a, and d_t denote the dimensionality of the features for each modality.

To reduce computational costs and minimize the number of tokens in the prefix, we employ a 1-D convolutional layer

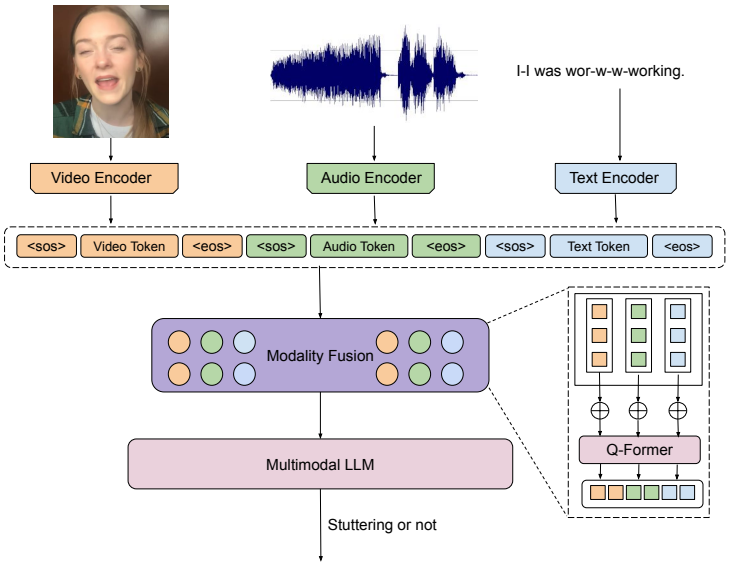

Figure 2: *Detailed architecture of our proposed model MMSD-Net*

to compress the length of the multi-modal features to a smaller value. Subsequently, a linear layer is employed to adjust the hidden size of the features, fusing it with the size of the MLMs' embeddings as follows:

$$\mathbf{h}'_v = \text{Linear}(\text{Conv1D}(\mathbf{h}_v)) \quad (4)$$

$$\mathbf{h}'_a = \text{Linear}(\text{Conv1D}(\mathbf{h}_a)) \quad (5)$$

$$\mathbf{h}'_t = \text{Linear}(\text{Conv1D}(\mathbf{h}_t)) \quad (6)$$

where $\mathbf{h}'_v, \mathbf{h}'_a, \mathbf{h}'_t$ are the transformed features with a fixed length of L' and an embedding dimension of d_e. The value of L' is significantly smaller than L_v, L_a, and L_t, while d_e corresponds to the dimensionality of the embedding matrix $\mathbf{E} \in \mathbb{R}^{v \times d_e}$ associated with the MMSD-Net.

3.2. Multimodality Fusion Module

Modality encoders are often trained independently, which can result in differences in the representations they produce. Therefore, it is essential to merge these distinct representations into a unified space. We design our fusion strategy based on the Q-Former [24], and modify it to efficiently integrate video and audio features with textual features, thereby facilitating more rapid adaptation. In this context, we designate the video features derived from our visual modality encoder by $h_v \in \mathbb{R}^{L_v \times d_v}$, and the audio features extracted from the audio modality encoder are denoted as $h_a \in \mathbb{R}^{L_a \times d_a}$.

In order to fuse the distinct representations learnt separately by the individual modality encoders, we consider the transformed visual and audio modality representations obtained in Equation 4, 5 and 6 as the soft tokens of our MLM module. The visual and audio representations are fused with the textual embedding space using the attention mechanism from Equation 1, as follows:

$$\mathbf{h}^a = \text{Attn}(\mathbf{h}', \mathbf{E}, \mathbf{E}) \quad (7)$$

where $\mathbf{h}'$ is the modality representation obtained in Equations 4, 5, and 6, $\mathbf{E}$ is the embedding matrix as defined in Section 3.1, and $\mathbf{h}^a$ is the corresponding fused representation, specifically $\mathbf{h}^a_v$, $\mathbf{h}^a_a$, and $\mathbf{h}^a_t$. After this fusion operation facilitated by the attention mechanism, the MMSD can seamlessly process the representations from various modalities.

In order to integrate the fused modality representations with the instruction information, we employ the concatenation operation. Given the fused modality representations, we define the integration as follows

$$x = [h^a_v : h^a_a : h^a_t : Embed(x_t)] \quad (8)$$

where $[:]$ represents the concatenation operation, x represents the multi-modal instruction, x_t represents the sequence of tokens in the textual instruction, and $Embed\ (x_t)$ represents the sequence of embeddings of x_t.

3.3. MLM module

Multimodal Language Models (MLM) [25] have demonstrated exceptional aptitude in comprehending and executing human directives. In particular, cross-modal approaches have demonstrated improved performance for audio synchronization tasks [26], [27]. In MMSD-Net, we utilize pretrained MLMs as the core modules, establishing the basis of MMSD-Net's functionality. The pretrained MLM network processes three primary inputs: the query vector $Q \in \mathbb{R}^{n_q \times d_q}$, the key vector $K \in \mathbb{R}^{n_k \times d_k}$, and the value vector $V \in \mathbb{R}^{n_v \times d_v}$. Through the scaled dot-product attention, the model evaluates attention scores by comparing each query in Q with all keys in K. These scores are then utilized to refine the query representations by generating a weighted sum of the values in V. The operation is mathematically represented as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Here, d_k denotes the size of the key and query vectors, while n_q and n_k represent the counts of queries and keys, respectively.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$

We use stacked multi-head attention as well as positional encodings to model the complete sequential information encoded

by the inputs. The decoder employs masked multi-head attention followed by softmax normalization for binary classification. The positional encodings are added to the input as well as output embeddings, enabling the model to capture the sequentiality of the input sentence without having recurrence. The encodings are computed from the position (pos) and the dimension (i) as follows:

$$PE_{(pos,2i)} = sin(pos/10000^{(2i/d_{model})}) \quad (9)$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{(2i/d_{model})}) \quad (10)$$

where PE stands for positional encodings and d_{model} is the dimensionality of the vectors resulting from the embeddings learned from the input and output tokens.

4. Experiments and Results

We conduct experiments for the detection of stuttered speech using our proposed model MMSD-Net, and compare the results obtained by our model with four state-of-the art baseline models, i.e. FluentSpeech [28], ResNet+BiLSTM [9], ConvLSTM [18] and StutterNet [17].

4.1. Datasets

This study uses four publicly available datasets (1 audio-modality and 3 audio-visual modalities) for the experiments. The details of modalities present in each dataset, along with the amount of data used for training and testing in each of our experiments, are described in Table 1.

Table 1: *Details of datasets used for our experiments.*

Dataset Name	Modality	Content	Training set	Testing set
Sept28K [18]	Audio	Normal Speech	28,000	0
FluencyBank [23]	Video, Audio	Normal +stuttered speech	52,000	0
Adults Who Stutter [23]	Video, Audio	Stuttered Speech	200	500
SpeakingFaces [29]	Video, Audio	Normal speech	200	500

4.2. Experimental Setup and Hyperparameters

For training efficiency, we leverage LoRA [30] for optimization on 4 Nvidia A100 GPUs. Each GPU handles a batch size of 4, with gradients accumulated for 5 steps before updating the model. The training process lasts for 10 epochs, employing a cosine learning rate scheduler with an initial learning rate of 5×10^{-10} and a warmup ratio of 0.02. To promote efficiency, FP16 precision is used for both training and inference. The maximum sequence length is capped at 512 tokens.

4.3. Results and Discussion

Table 2 gives the results (in terms of precision, recall and F1-score) of the proposed MMSD-Net method along with the four baseline methods. The results are presented for the dataset obtained by combining the FluencyBank [23], Sept28K [18] and Adults Who Stutter [23] datasets as described in Table 1. In addition to model performance, the comparison of modalities used by each method is also presented in the Table 2. Among the four baseline methods, StutterNet [17] achieved better results compared to other three methods (FluentSpeech [28], ResNet+BiLSTM [9] and ConvLSTM [18]), and ResNet+BiLSTM [9] achieved the lowest scores among the baseline methods. While both StutterNet [17] and ResNet+BiLSTM [9] are specifically trained on audio data, StutterNet achieves superior performance using MFCC features extracted from the audio samples. This suggests that MFCC features are more suitable for stutter detection compared to the features employed by ResNet+BiLSTM, i.e. the spectrograms.

Our proposed MMSD-Net outperforms all other methods in terms of precision, recall, and F1-score, achieving the highest scores across all metrics, which demonstrates its superiority in stuttering detection. Our findings demonstrate that the fusion module can effectively combine information from three different modalities. Quantitatively, the proposed MMSD-Net gave an absolute improvement of 2% in the F1-score (and 3% in the Precision) over the best baseline method (StutterNet [17]) and 16% in the F1-score (and 17% in the Precision) over ResNet+BiLSTM [9]. These results validate our hypothesis that facial expressions serve as an important cues to detect stuttered speech, and employing the visual signal as part of a multi-modal framework improves the performance of automatic stuttering detection.

Table 2: *Comparison of results obtained by the proposed MMSD-Net with the four baseline methods. Best result highlighted in bold, second best underlined.* ***P*** *= Precision,* ***R*** *= Recall,* ***F1*** *= F1-score.*

.

Model Name	Audio	Video	P	R	F1
FluentSpeech [28]	✓	✗	85.73	81.82	83.72
ResNet+BiLSTMs [9]	✓	✗	75.28	72.73	73.98
StutterNet [17]	✓	✗	<u>89.41</u>	<u>87.10</u>	<u>88.23</u>
ConvLSTM [18]	✓	✗	82.63	78.32	80.41
MMSD-Net	✓	✓	**92.58**	**87.93**	**90.19**

5. Conclusion

This study presents MMSD-Net, the first multi-modal neural framework crafted explicitly for stuttered speech detection. We conducted experiments on publicly available datasets and performed studies comparing against four existing uni-modal baselines. Our findings showcase noteworthy improvements in stuttered speech detection accuracy, with enhancements ranging from 2-17% in F1-score over established baseline models, indicating the effectiveness of our multi-modal approach for stuttering detection. This paper signifies a major stride towards augmenting the efficacy of stuttered speech detection, and highlights the complementarity of multiple modalities for this task. In particular, it demonstrates that the visual signal carries relevant information for the stuttering detection task, and lays the groundwork for the development of further advancements in speech processing tools catered to individuals suffering from speech impediments. In the future, we would like to extend our experimentation to larger datasets and conduct a qualitative analysis of the impact of multi-modality on handling various kinds of stuttering.

6. References

[1] E. Mullin, "Why Siri won't listen to millions of people with disabilities," *Scientific American. Retrieved January*, vol. 8, p. 2018, 2016.

[2] J. E. Prasse and G. E. Kikano, "Stuttering: an overview," *American family physician*, vol. 77, no. 9, pp. 1271–1276, 2008.

[3] E. Yairi and N. Ambrose, "Epidemiology of stuttering: 21st century advances," *Journal of Fluency Disorders*, vol. 38, no. 2, pp. 66–87, 2013.

[4] P. Howell and S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech," in *Proceedings of the first World Congress on fluency disorders*, vol. 2. University Press Nijmegen Nijmegen, The Netherlands, 1995, pp. 372–374.

[5] P. Howell, S. Sackin, and K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: ANN recognition of repetitions and prolongations with supplied word segment markers," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 5, pp. 1085–1096, 1997.

[6] T.-S. Tan, A. Ariff, C.-M. Ting, S.-H. Salleh *et al.*, "Application of malay speech technology in malay speech therapy assistance tools," in *International Conference on Intelligent and Advanced Systems*. IEEE, 2007, pp. 330–334.

[7] K. Ravikumar, R. Rajagopal, and H. Nagaraj, "An approach for objective assessment of stuttered speech using MFCC," in *The international congress for global science and technology*, vol. 19, 2009.

[8] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[9] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6089–6093.

[10] B. Villegas, K. M. Flores, K. J. Acuña, K. Pacheco-Barrios, and D. Elias, "A novel stuttering disfluency classification system based on respiratory biosignals," in *IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 4660–4663.

[11] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional lstm," *Interspeech 2016*, 2016.

[12] J. Santoso, T. Yamada, and S. Makino, "Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 302–306.

[13] K. Georgila, "Using integer linear programming for detecting speech disfluencies," in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009, pp. 109–112.

[14] T. Kourkounakis 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.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Q. Chen, M. Chen, B. Li, and W. Wang, "Controllable time-delay transformer for real-time punctuation prediction and disfluency detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8069–8073.

[17] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "StutterNet: Stuttering detection using time delay neural network," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 426–430.

[18] 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," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6798–6802.

[19] S. P. Bayerl, D. Wagner, E. Nöth, and K. Riedhammer, "Detecting dysfluencies in stuttering therapy using wav2vec 2.0," *arXiv preprint arXiv:2204.03417*, 2022.

[20] Z. Jiang, Q. Yang, J. Zuo, Z. Ye, R. Huang, Y. Ren, and Z. Zhao, "Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models," *arXiv preprint arXiv:2305.13612*, 2023.

[21] P. M. Roberts, A. Meltzer, and J. Wilding, "Disfluencies in non-stuttering adults across sample lengths and topics," *Journal of communication disorders*, vol. 42, no. 6, pp. 414–427, 2009.

[22] P. Riva-Posse, L. Busto-Marolt, Á. Schteinschnaider, L. Martinez-Echenique, Á. Cammarota, and M. Merello, "Phenomenology of abnormal movements in stuttering," *Parkinsonism & related disorders*, vol. 14, no. 5, pp. 415–419, 2008.

[23] N. B. Ratner and B. MacWhinney, "Fluency Bank: A new resource for fluency research and practice," *Journal of Fluency Disorders*, vol. 56, pp. 69–80, 2018.

[24] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023.

[25] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," 2019.

[26] R. Agrawal, D. Wolff, and S. Dixon, "Structure-aware audio-to-score alignment using progressively dilated convolutional neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 571–575.

[27] ——, "A convolutional-attentional neural framework for structure-aware performance-score synchronization," *IEEE Signal Processing Letters*, vol. 29, pp. 344–348, 2021.

[28] Z. Jiang, Q. Yang, J. Zuo, Z. Ye, R. Huang, Y. Ren, and Z. Zhao, "FluentSpeech: Stutter-oriented automatic speech editing with context-aware diffusion models," in *Findings of the Association for Computational Linguistics (ACL)*, Toronto, Canada, Jul. 2023.

[29] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, p. 3465, 2021.

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021.