



# Towards Inclusive and Fair ASR: Insights from the SAPC Challenge for Optimizing Disordered Speech Recognition

Nada Gohider, Otman Basir

<sup>1</sup>Electrical & Computer Engineering Department, University of Waterloo, Canada

ngohider@uwaterloo.ca, otman.basir@uwaterloo.ca

## Abstract

ASR has advanced significantly, yet remains limited for impaired speakers due to data scarcity. In response to this gap, the Speech Accessibility Project (SAP) represents a significant initiative in data collection on impaired speech. This paper reports our participation in the SAPC challenge, where we leveraged SAP data to improve ASR performance for disordered speech. Our system ranked **fourth** in terms of Word Error Rate, recording values of **10.06%** and **11.8%** WER for Test1 and Test2 subsets, respectively, on the challenge leaderboard. In particular, our research examines the power of SOTA ASR models to capture contextual information in the presence of disordered speech disfluencies. We focused on two ASR architectures, ContextNet and Parakeet, based on their documented ability to efficiently and effectively handle contextual information for typical speech, utilizing distinct mechanisms. Our experiments demonstrated that Parakeet slightly outperformed ContextNet, as evidenced by WER.

**Index Terms:** disordered speech recognition, dysarthric speech recognition, inclusive ASR, dysarthria, SAP dataset, SAPC challenge

## 1. Introduction

The performance of current ASR systems has been boosted by the dominance of deep learning (DL) algorithms, which are highly dependent on the size of the training data [1][2]. Thus, a key factor in ASR advancements is the amount and quality of the data used to train the underlying machine learning (ML) models of these ASR systems. LibriSpeech [3], for instance, is a benchmark that has been extensively used to train and evaluate ASR models. Its large size of 1000 hours sampled from high-quality speech, in addition to its diversity and being open source, are key factors behind its popularity in ASR research [3]. However, these datasets are generally sampled from standard speech, relying falsely on the assumption of *independent and identically distributed (i.i.d)* as an inductive bias for the driving ML models. Consequently, these ASR models fall short when exposed to environments or settings that differ from those encountered during the training phase, causing a major setback in these systems[4][2][5]. In this context, disordered speech is a case of failure of current ASR systems, where performance can significantly deteriorate due to the distribution mismatch between typical and impaired speech.

Although speech technology should be essentially inclusive and accessible for all individuals, it can be even more critical for people with speech disorders. Speech impairments, such as dysarthria, constitute a significant barrier that can negatively impact impaired speakers in different aspects of life [6]. Furthermore, the failure of current ASR models to accommodate

impaired speakers can hinder their digital accessibility, excluding them from the benefits of ASR technology. Therefore, these challenges motivated ASR researchers to explore various approaches in search of solutions to facilitate digital accessibility of ASR applications for individuals with speech disabilities. In this context, domain adaptation and personalized ASR models have been widely studied to optimize the task of *automatic disordered speech recognition (ADSR)*. However, the lack of data on impaired speech limits the effectiveness of these models, leading to suboptimal results. A key issue that limits the development of ASR models for impaired speech is data scarcity [2][5]. In addition to the limited size of the existing datasets of pathological speech, these datasets lack the important characteristic of data diversity that is key for ML model generalization. In particular, most of the available data, such as UASpeech [7], is sampled from a small number of speakers who represent only a few cases of speech impairments, ignoring a wide spectrum of impairments that negatively impact speech representation when learning ASR models.

Responding to these needs, researchers at the University of Illinois Urbana-Champaign have initiated a great collaborative work, the **Speech Accessibility Project (SAP)**[8], which aims to address the problem of data shortage of impaired speech and ultimately bridge the performance gap of ASR systems and enable them to reach people with speech impairments. In fact, SAP data provides a level of granularity and diversity that was previously unavailable in the context of the disordered speech recognition task. Its comprehensive scope and enriched features enable more precise analyses, facilitating deeper insights into patterns and phenomena that were difficult to capture with earlier datasets. In this context, **Speech Accessibility Project Challenge (SAPC)** is a public challenge that was recently organized by the SAP team, aiming to facilitate access to the SAP data by ASR researchers and practitioners and utilize it to advance the research of ASR targeting disordered speech.

This paper reports our participation in the SAPC challenge, with the aim of contributing to the advancement of ASR research and making it accessible to individuals with speech disabilities. As SAP data offers a more representative and high-resolution perspective of speech impairments in the context of the ASR task, it encourages us to raise questions that were previously constrained by data limitations. From this perspective, this paper seeks to utilize the continuity in the SAP data and study the power of the SOTA ASR model to capture contextual information in disordered speech. We assume that the atypical patterns inherent in disordered speech can affect both local and long-term dependencies (i.e., the local and global context). Addressing these challenges necessitates architectures that can flexibly capture both short-term phonetic cues and long-term dependencies without being constrained by predefined feature

extraction mechanisms. Moreover, we believe that the benefits of ASR systems for impaired speakers can extend to a wide range of applications, including daily assistance, medical monitoring, and the early diagnosis of speech impairments. In light of this, the ability to accurately capture disfluencies in disordered speech—such as prolonged phonemes, irregular speech patterns, and filler words—is particularly crucial. This capability is essential for applications where ASR technology is employed for the early assessment of speech impairments, facilitating timely intervention and improved clinical outcomes. Therefore, our research examines the power of some SOTA ASR models to capture contextual information in the presence of disordered speech disfluencies. We focus on two ASR architectures, ContextNet [9] and FastConformer (Parakeet) [10]. We selected these architectures because of their documented ability to efficiently and effectively handle contextual information for typical speech, utilizing distinct mechanisms [9][10][11].

The remainder of this paper is organized as follows. Section 2 summarizes the research activity of related work in the literature on disordered speech recognition. Further, a brief summary of the SAPC challenge is reported in Section 3. Moreover, Section 4 and Section 5 give an overview of the ContextNet and Parakeet models, respectively. After that, the data preprocessing pipeline and the experimental setup are explained in Section 6, and results and discussion are given in Section 7. We, then, conclude the work in Section 8.

## 2. Related work

ASR researchers have explored various approaches to bridge the performance gap for impaired speech [2]. One line of research has focused on personalized models, where speaker-dependent systems are trained using data from specific target speakers [1][12][13]. Other widely studied directions are transfer learning and domain adaptation, which have been implemented through a range of techniques to enhance model generalization [14][15][16] [1]. While these research efforts have contributed to advancements in the field, their performance remains suboptimal due to persistent limitations. One of the most significant limitations is data scarcity, characterized by insufficient training samples and a lack of diversity. In light of this, a large body of ADSR research was conducted on the UASpeech dataset [7]; however, its constrained size and restricted range of etiologies present significant challenges in developing robust ASR models for disordered speech. The reliance on isolated words has constrained the ability to examine key aspects of contextual features in speech. Therefore, this research aims to shed some light on this gap by utilizing SAP data and some SOTA ASR models to study the impact of capturing contextual features in the scenario of disordered speech.

## 3. SAPC challenge

SAP [8] is an ongoing project that is actively collecting U.S. English speech from impaired speakers. To our knowledge, SAPC<sup>1</sup> is the first ASR challenge exclusively focused on disordered speech. It is an initiative to facilitate the use of SAP data by speech researchers and practitioners and ultimately enable ASR systems for people with speech disabilities. In particular, the SAP team [8] organized a challenge where registered teams were tasked with training ASR models and supported with the latest release of the SAP data (April 2024) along with a set

<sup>1</sup><https://eval.ai/web/challenges/challenge-page/2362/overview>

of data preprocessing guidelines. The submitted models were evaluated on two test subsets, Test1 and Test2, sampled from a range of speech impairments, and were withheld from participants. During the challenge, all participants were allowed a limit of submissions that were evaluated on the Test1 subset while the final evaluation was scored on the Test2 subset. For each test, the final scores reflect the best match among two used test transcripts: with- and without disfluencies. Examples of these disfluencies are filler words, false starts, repetitions of full words or phrases, and partial words. The results of both tests are publicly shared on the challenge leaderboard. The SAPC submissions were scored using two evaluation metrics: 1-word error rate (WER) and 2-semantic score (SemScore). WER measure [17] reflects the normalized edit distance between the prediction and the reference, considering three types of errors: insertions, deletions, and substitutions. On the other hand, SemScore reflects the semantic similarity between the predicted words and the ground truth on the semantic level. In particular, SemScore [18] is the weighted combination of BERTscore, phonetic distance, and natural language inference probability.

## 4. ContextNet

ContextNet [9] is composed of a fully convolutional encoder, which makes it parameter efficient, while it leverages an autoregressive decoder using a transducer/RNN module. ContextNet uses depthwise separable convolution aiming to further reduce the number of parameters, optimizing its efficiency. As CNNs can effectively capture the local context with their filters, they lack control over long-term dependencies. To overcome this problem, ContextNet utilizes squeeze-and-excitation (SE) modules to improve the representation power of the model and capture the global context in speech signals [9]. Another key advantage that features ContextNet is the use of progressive downsampling that seeks to optimize the training speed and memory without impacting the overall model performance.

An SE module seeks to learn a context-aware feature representation that efficiently captures the global context of the input speech signal [9]. This learned representation, then, supports the CNN outputs, which are augmented with local dependencies. As depicted in Figure 1, the SE module captures the long-term dependencies through a two-step process, as follows:

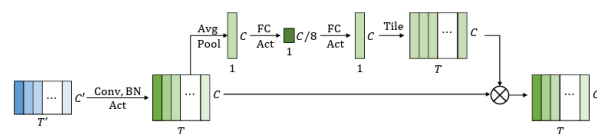


Figure 1: *Squeeze-and-Excitation Block in ContextNet Model [9]*

- **Squeeze:** This step simply encodes the global context  $\theta(x)$  by aggregating a sequence of local feature vectors using global average pooling, as depicted in Equation 1 [9].

$$\theta(x) = \text{Sigmoid}(W_2(\text{Act}(W_1\bar{x} + b_1) + b_2)) \quad (1)$$

, where  $\bar{x} = \frac{1}{T} \sum_t x_t$ , and  $[W_1, W_2]$  and  $[b_1, b_2]$  are the weight matrices and the bias vectors, respectively.

- **Excite:** Then, the excitation is realized by passing the global context vector into a sequence of local feature vectors to ex-

tract weighted features that reflect the importance of contextual information. i.e., the global channelwise weight is broadcasted on a sequence of local feature vectors using the element-wise multiplication, as described in Equation 2 [9].

$$SE(x) = \theta(x) \circ x \quad (2)$$

## 5. Parakeet

Parakeet<sup>2</sup> is a family of ASR models that share the FastConformer encoder [10], while they can have a CTC, transducer, hybrid, or TDT decoder. FastConformer is an optimized variant of the Conformer model [19]. The Conformer model integrates the attention mechanism to complement the convolutional layers merits, leveraging the advantages of both worlds. This, therefore, helps the Conformer model capture the local context through its kernels in the convolutional layers while leveraging the attention modules to learn the long-term dependencies in the speech signals [19]. FastConformer was introduced as an optimized version of the Conformer model, reducing the computational cost while preserving the recognition capacity to achieve SOTA performance. The key improvements in FastConformer architecture were realized by adding 8x depthwise-separable convolutional downsampling, modified convolution kernel size, and an efficient subsampling module [10]. In addition, FastConformer can greatly handle long-form audio (up to 11 hours) in a single pass without the need for any further segmentation or post-processing steps, as shown in Figure 2[10].

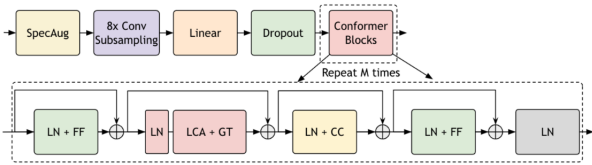


Figure 2: FastConformer [11]

## 6. Experiments

### 6.1. Data

We used the April 2024 release of SAP data, shared by the SAP team. Approximately a total of 415 hours of speech that was uttered by 524 speakers contributed to the SAP dataset. The diversity in SAP data is realized in different aspects where continuous speech was recorded using both spontaneous and reading prompts. While participants in SAP data are dominantly diagnosed with Parkinson’s disease, other disabilities are involved, such as amyotrophic lateral sclerosis (ALS), cerebral palsy (CP), Down syndrome (DS), and stroke. For data preprocessing, we followed a set of guidelines, advised by the SAPC challenge organizers, such as resampling audio to 16 kHz and basic text preprocessing (e.g., text normalization, including digital numbers, abbreviations, and special punctuations). Additionally, disfluencies resulting from speech impairments were retained in their spoken form to preserve the natural characteristics of the impaired speech. This process results in a total

<sup>2</sup><https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/asr/models.html>

of processed data of approximately 334 hours, along with their transcription. We split the data as follows: ~290h train subset, ~35h dev subset, and ~8.5h test subset. These data subsets were uttered by non-overlapping speakers. Moreover, SAP data is curated by the SAP team in such a way that it prioritizes the participants’ privacy and confidentiality and does not reveal any information about their identities. To prevent ambiguity, we designate the latter test subset as **test3** to clearly distinguish it from **Test1** and **Test2**, which were utilized by the SAPC challenge organizers for evaluating participants’ submissions.

### 6.2. Experimental setup

In addition to the results achieved in the SAPC Challenge submission, this paper also presents the performance of our trained models on the test3 subset, which we dedicate to our analysis and not used during model training. In these experiments, we employed BERTScore as a semantic similarity metric—alongside WER—to compare the utterance-level predictions of the ContextNet and Parakeet-RNNT models. BERTScore was chosen for its simplicity and interpretability relative to the more complex SemScore metric. This analysis aims to provide a deeper understanding of data trends and to inform the fine-tuning of the models’ hyperparameters. All experiments were carried out using the NeMo NVIDIA toolkit [20]. In terms of compute resources, all experiments were run on ComputeCanada<sup>3</sup> nodes, where we used one node of four GPUs of NVIDIA V100.

In terms of model initialization, we followed a domain adaptation framework where we started with publicly available pre-trained models that were trained earlier on typical speech. For ContextNet, we started with the NeMo pre-trained checkpoint<sup>4</sup> *stt\_en\_contextnet\_1024*, which consists of around 140M parameters that were trained earlier on NeMo ASRSet with over 24,500 hours of English speech. For the Parakeet model, we started with the NeMo pre-trained checkpoint<sup>5</sup> *nvidia/parakeet-rnnt-0.6b* of 600M parameters that were trained earlier on NeMo 64K hours of English speech collected in collaborative work between NVIDIA NeMo and Suno teams. Instead of reinventing the wheel for the ADSR task, we intended to leverage the knowledge embedded in those pre-trained models, as they have been trained on large-scale, diverse datasets encompassing a wide range of linguistic and acoustic conditions. These datasets include various English accents, multiple domains, and diverse environmental noise conditions, thereby enhancing the model’s robustness and generalization capabilities.

For Parakeet models, we examined the performance with both CTC and RNN/transducer decoders, using beam search without external LM, focusing solely on core acoustic modeling. For all experiments, we used SentencePiece Unigram tokenizer we trained on SAP dataset. In particular, we explored the impact of the vocabulary size relative to the model size, where we trained two tokenizers of 256 and 1024 vocab spaces. In terms of the optimizer, we used the AdamW optimizer, where we varied the learning rate in the range  $[10^{-5}, 3 \times 10^{-5}]$  and weight decay of 0.01. While we compared the models of the same training cycles/epochs, we set 10 consecutive epochs as the limit to stop training if the model was not further improving.

<sup>3</sup>[https://docs.alliancecan.ca/wiki/Technical\\_documentation](https://docs.alliancecan.ca/wiki/Technical_documentation)

<sup>4</sup>[https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt\\_en\\_contextnet\\_1024](https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_contextnet_1024)

<sup>5</sup><https://huggingface.co/nvidia/parakeet-rnnt-0.6b>

## 7. Results and Discussion

According to the public leaderboard<sup>6</sup> of the SAPC challenge, our submission achieved the fourth-lowest Word Error Rate (WER) on both challenge tests, recording values of **10.06%** and **11.8%** for Test1 and Test2, respectively. These results were accomplished with fine-tuning the pretrained Parakeet-RNNT model for 70 epochs, utilizing our trained tokenizer of 256 vocab size, and  $3 \times 10^{-5}$  learning rate, and beam search decoding. Table 1 summarizes our results on Test1 and Test2 subsets in terms of WER and SemScore as reported in the challenge submissions. The selection of hyperparameters was informed by a series of experiments conducted on test3 subset, in which we systematically varied several components, including the encoder architecture (Parakeet vs. ContextNet), loss function, and learning rate.

Table 1: *Our SAPC results of Parakeet model on Test1 and Test2 subsets*

Test Subset	WER	SemScore
Test1	<b>10.06%</b>	86.34
Test2	<b>11.8%</b>	83.29

In terms of capturing disfluencies in disordered speech, which is a key research question in this work, we have noticed that Parakeet-RNNT outperforms ContextNet across all our experiments. This advantage is particularly evident in long-form audio scenarios, where Parakeet demonstrates superior robustness in handling atypical speech patterns. While ContextNet is comparatively less effective in capturing these disfluencies, its BERTScore performance remains competitive with that of Parakeet, suggesting strong capabilities in modelling global context, as evidenced by the semantic similarity measure. Additionally, the performance gap between the two models narrows when evaluated on clean speech or shorter utterances, where both models achieve similar accuracy. Table 2 presents some examples that compare the predictions of the ContextNet and Parakeet-RNNT models on the utterance level in terms of WER and BERTScore as a semantic similarity metric. More specifically, we have noticed that across all our experiments, disfluencies commonly found in disordered speech, such as word and phrase repetitions and filler words. Although the discrepancy between the performance of the two models in terms of transcription accuracy may not significantly impact the overall performance in terms of the semantic meaning of the transcribed text, capturing these disfluencies can play a vital role in some applications, such as medical assessment scenarios.

Table 2: *Utterance-level comparison of ContextNet vs. Parakeet-rnnt*

Duration	Num words	Ground truth	ContextNet-pred	Parakeet-rnnt-pred	ContextNet WER/BertScore	Parakeet-rnnt WER/BertScore
10.35	3	"PLEASE HANG UP"	"PLEASE HANG UP"	"PLEASE HANG UP"	0.1	0.1
10.2	7	"WHERE DID DID I PAR PARK PARK"	"WHERE DID DID I PARK PARK"	"WHERE DID DID I PAR PAR PARK"	28.57(0.81)	0(1)
14.88	18	"SET A REMINDER TO TAKE POTASSIUM AT TWO P M TO ELEVEN P M A M A M"	"SET A REMINDER TO TAKE POTASSIUM FROM TWO P M"	"SET A REMINDER TO TAKE POTASSIUM FROM TWO P M TO ELEVEN P M A M A M"	50(0.83)	5.56(0.96)
16.16	14	"IS THERE A CHANCE OF RAIN RAIN ON THE TWENTH ETH ETH"	"IS THERE A CHANCE OF RAIN RAIN ON THE TWINTH"	"IS THERE A CHANCE OF RAIN ON THE TWENTH"	42.86(0.72)	28.57(0.75)
21.99	18	"RECORD PLEASE RECORD PLEASE RECORD THE TEMPERATURE IN THE LIVING ROOM WHAT'S THE TEMPERATURE IN THE LIVING ROOM"	"RECORD PLEASE RECORD WHAT THE TEMPERATURE IN THE LIVING ROOM WHAT'S THE TEMPERATURE IN THE LIVING ROOM"	"RECORD PLEASE RECORD PLEASE RECORD THE TEMPERATURE IN THE LIVING ROOM"	16.67(0.89)	0(1)
38.73	10	"STOP RECORDING ADD THIS SHER ADD THIS MUSIC TO FAVORITES"	"STOP RECORDING ADD THIS MUSIC TO FAVORITES"	"STOP RECORDING ADD THIS FAVORITES ADD THIS MUSIC TO FAVORITES"	30(0.89)	1(0.89)

<sup>6</sup><https://eval.ai/web/challenges/challenge-page/2362/leaderboard/5859>

Generally, Parakeet-RNNT achieved slightly better results than ContextNet reporting **4.9%** and **6.67%** respectively, after fine-tuning for 70 epochs on the test subset. More specifically, we found that the transducer-based variant (Parakeet-RNNT) consistently outperformed Parakeet-CTC, as depicted in Table 3. This can be attributed to the RNN-T architecture's ability to incorporate language modeling, which enhances contextual understanding. This advantage was also evident during training, where Parakeet-RNNT exhibited faster convergence compared to the CTC-based variant. Additionally, we observed that, generally, small learning rates contribute to more stable training, albeit at the cost of slower convergence. This stability is particularly beneficial for preserving the knowledge transferred from the pre-trained checkpoints while simultaneously adapting to the patterns of disordered speech. Regarding the vocabulary space of the trained tokenizer, our experiments indicate that a vocabulary size of 256 tokens yields slightly better performance than 1024 tokens. We hypothesize that for a medium-sized dataset such as SAP, a smaller vocabulary size may be more advantageous in optimizing model performance.

Table 3: *Performance of Parakeet0.6B on the test3 subset (5000samples), varying the decoder, learning rate (lr) and the tokenizer vocab size*

Decoder	Vocab Size	lr	WER
CTC	256	$10^{-5}$	11.12
	256	$3 \times 10^{-5}$	10.33
	1024	$10^{-5}$	14.89
	1024	$3 \times 10^{-5}$	13.8
RNNT	256	$10^{-5}$	5.1
	256	$3 \times 10^{-5}$	<b>4.9</b>
	1024	$10^{-5}$	6.3
	1024	$3 \times 10^{-5}$	7.1

While inference speed is not the focus of this work, ContextNet demonstrates a slight advantage over Parakeet, reflecting their difference in model size. Specifically, transcribing the test subset (5000 samples), with a batch size of 8, required approximately 3.45min and 6min for ContextNet and Parakeet, respectively, on a NVIDIA GeForce RTX 3080 GPU. Among the Parakeet variants, Parakeet-CTC showed faster inference than the transducer version, despite its slower convergence during training.

## 8. Conclusion

In this study, we present our participation in the SAPC challenge, where we leveraged the diversity of the SAP dataset to address the performance gap in the ADSR literature. Specifically, we investigated the power of some SOTA ASR models to capture the contextual information in disordered speech, which is often influenced by the disfluencies inherent in impaired speech. Our experimental findings indicate that FastConformer demonstrates greater effectiveness in capturing these disfluencies compared to ContextNet, a convolutional-based ASR model. This advantage is particularly evident in long-form speech scenarios. Additionally, our results suggest that, for future work, studying more variations of the context limit in FastConformer is a promising direction and should be further investigated for impaired speech. Moreover, our work can be further optimized by incorporating external language models to refine the acoustic model's output.

## 9. References

- [1] H. Christensen, S. P. Cunningham, C. Fox, P. D. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech." in *Interspeech*. Portland, 2012, pp. 1776–1779.
- [2] N. Gohider and O. A. Basir, "Recent advancements in automatic disordered speech recognition: A survey paper," *Natural Language Processing Journal*, p. 100110, 2024.
- [3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books;" in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [4] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.
- [5] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the cuhk dysarthric speech recognition system for the ua speech corpus." in *Interspeech*, 2018, pp. 2938–2942.
- [6] A. Craig, E. Blumgart, and Y. Tran, "The impact of stuttering on the quality of life in adults who stutter," *Journal of fluency disorders*, vol. 34, no. 2, pp. 61–71, 2009.
- [7] H. V. Sharma and M. Hasegawa-Johnson, "Universal access: Speech recognition for talkers with spastic dysarthria," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
- [8] M. Hasegawa-Johnson, X. Zheng, H. Kim, C. Mendes, M. Dickinson, E. Hege, C. Zwilling, M. M. Channell, L. Mattie, H. Hodges *et al.*, "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024.
- [9] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv preprint arXiv:2005.03191*, 2020.
- [10] D. Rekish, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [11] N. R. Koluguri, S. Kriman, G. Zelenfroind, S. Majumdar, D. Rekish, V. Noroozi, J. Balam, and B. Ginsburg, "Investigating end-to-end asr architectures for long form audio transcription," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 366–13 370.
- [12] F. Rudzicz, "Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, 2007, pp. 255–256.
- [13] F. Rudzicz, G. Hirst, and P. van Lieshout, "Vocal tract representation in the recognition of cerebral palsied speech," 2012.
- [14] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852–861, 2021.
- [15] S. R. Shahamiri, V. Lal, and D. Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.
- [16] T. Mariya Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023.
- [17] A. C. Morris, V. Maier, and P. D. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition." in *Interspeech*, 2004, pp. 2765–2768.
- [18] A. Aynedinov and A. Akbik, "Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity," *arXiv preprint arXiv:2401.17072*, 2024.
- [19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [20] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.