Automatic Speech Recognition Models for Pathological Speech: Challenges and Insights

Kesego Mokgosi[1,*], Cathy Ennis[1] and Robert Ross[1,2]

[1]Technological University Dublin, Dublin,Ireland
[2]ADAPT Research Centre, Dublin,Ireland

Abstract

Conversational avatars provide innovative platforms for enhancing therapist-patient interactions in speech therapy by offering real-time feedback. However, the performance of Automatic Speech Recognition (ASR) models on disordered speech, such as dysarthria and stuttering, remains underexplored. The effectiveness of these systems hinges on the accuracy and processing speed of ASR models when transcribing pathological speech, particularly in real-time scenarios. This study evaluates several pre-trained ASR models, including Whisper-large-v3-turbo, Canary, DistilWhisper, and NVIDIA's stt-en-fastconformer-ctc-large across three datasets: Common Voice (standard speech), TORGO (dysarthric speech), and UCLASS (stuttered speech). We assess the models using Word Error Rate (WER), Real-Time Factor (RTF), and BERTScore to measure transcription accuracy, computational efficiency, and semantic congruence. The stt-en-fastconformer-ctc-large model demonstrates the fastest processing speeds, achieving the lowest WER and highest BERTScores on both the Common Voice and TORGO datasets, making it highly suitable for real-time therapeutic applications. However, all models struggle with accurately transcribing stuttered speech from the UCLASS dataset. These results highlight the need for ASR improvements for disordered speech, focusing on edge deployment to reduce latency and enhance accuracy with multimodal inputs.

Keywords

Automatic Speech Recognition, Disordered Speech, Conversational Avatars, Speech Therapy

1. Introduction

Conversational avatars are an innovative platform for therapist-patient interactions that hold significant potential to improve speech therapy outcomes. By providing real-time feedback, these avatars aim to revolutionise speech dysfunction therapy and enhance patient engagement [1]. However, the efficacy of these avatars critically depends on the accuracy and speed of voice transcription, which is a challenging task when dealing with disordered speech [2]. Disordered speech, which affects phonation, articulation, and fluency, poses significant challenges for Automatic Speech Recognition (ASR) systems [2]. In the context of conversational avatars, ASR serves as the foundational technology that enables the avatar to understand and respond to the patient's speech. Therefore, the performance of ASR systems is crucial for seamless transcription and uninterrupted therapy sessions [3].

According to Georgescu et al. [4], ASR models in conversational therapy avatars need to balance transcription accuracy with hardware constraints, such as processor speed and memory usage. Transformer-based models like Whisper [5] and Canary [6] have shown effectiveness in several speech datasets by enhancing computational efficiency through parallel processing. However, their capacity to handle disordered speech in therapeutic contexts remains underexplored. Speech disorders like apraxia and dysarthria present unique challenges for ASR systems due to irregular pronunciation, distorted phonemes, and variable speech rates [7], [8], often leading to transcription errors and delays that disrupt the therapeutic interaction. These interruptions can hinder the effectiveness of therapy sessions [8], indicating the need for fast and accurate ASR inputs to maintain patient engagement. Additionally,

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
*Corresponding author.
d23126641@mytudublin.ie (K. Mokgosi); cathy.ennis@tudublin.ie (C. Ennis); robert.ross@tudublin.ie (R. Ross)
https://github.com/kesbeast23/ (K. Mokgosi)
0009-0000-5713-2002 (K. Mokgosi)

edge-based solutions are becoming increasingly essential in speech therapy applications as they process and store data closer to the source. This reduces latency and ensures data privacy, which are critical factors in therapeutic settings [9]. Deploying ASR models locally enables real-time processing without the delays typically associated with cloud-based services, while also addressing privacy concerns by keeping sensitive patient data on-site [10]. This justifies our focus on testing ASR models through local deployments rather than relying on hosted services.

While prior research has examined ASR performance on disordered speech, dialectal speech, low-resource languages, and children's speech, they rarely address specific conversational challenges in therapeutic applications, such as real-time dialogue management, error correction, and maintaining patient engagement [7]. In this paper, we carefully benchmark state-of-the-art ASR models against speech from patients with speech impairments, testing their feasibility in high-stakes, real-time speech therapy scenarios. We provide a comprehensive evaluation of pre-trained ASR models on disordered speech without fine-tuning, assessing their transcription accuracy, processing speed and semantic congruence within the context of speech therapy applications. By highlighting the limitations of current ASR models in handling different types of speech disorders, we analyse how these limitations impact real-time therapeutic interactions. Furthermore, we emphasise the importance of edge-based solutions for ASR in speech therapy, justifying the use of local model deployments to reduce latency and address privacy concerns in clinical settings. Based on our findings, we offer insights and recommendations for speech disorder ASR models and the integration of multimodal inputs, aiming to enhance semantic comprehension and efficacy in conversational speech therapy applications.

2. Related Work

Automatic Speech Recognition (ASR) systems face significant challenges when processing disordered speech, such as stuttering, dysarthria, and apraxia, due to variations in fluency, articulation, and phonation. These differences hinder the performance of conventional ASR models, which are typically trained on fluent speech data. In exploratory research, Green et al. [11] developed ASR models specifically tailored for individuals with speech impairments using the TORGO [12] and UCLASS [13] datasets. While these models showed improved comprehension in controlled environments—such as isolated word or phrase recognition, they struggled significantly in conversational scenarios where speech is more spontaneous and variable, limiting their practical application in real-world settings [8]. ASR systems have also been studied in the context of speech therapy applications, particularly with conversational avatars and voice assistants. Mitra et al. [14] explored the impact of fluency issues on ASR performance within a voice assistant system and found that traditional ASR models struggled to accurately process stuttered speech, resulting in misunderstandings and communication breakdowns. By employing hybrid ASR models with modified decoding parameters, they improved transcription accuracy for moderate to severe stuttering, enhancing the usability of voice assistants for individuals with fluency disorders. Mulfari and Villari [15] focused on optimising ASR for edge devices for users with speech impairments. They fine-tuned models like Whisper for edge computing nodes, improving ASR performance for impaired speech by enabling more efficient on-device processing, which is critical for real-time applications and ensuring user privacy.

Clinically, transcription precision is critical for conveying speaker intent and enabling meaningful use of assistive technologies. Tobin et al. [16] explored the use of conversational avatars for providing real-time therapy feedback and found that while personalised ASR models performed well with short, scripted phrases, they faced challenges with spontaneous conversational speech, particularly for individuals with severe speech impairments. Mitra et al. [14] and Mulfari et al. [15] highlighted the potential to optimise ASR systems for specific applications without requiring extensive retraining. By adjusting decoding parameters for Whisper [5] and utilising edge computing, these models can achieve a balance between accuracy. Recent research has explored fine-tuning models like Whisper and Conformer on disordered speech datasets using techniques such as data augmentation and transfer learning. While models like Whisper [5] and Canary [6] excel at transcribing standard speech due to their capability to

model long-range dependencies in audio sequences, their performance on disordered speech remains underexplored. Studies by Mitra et al. [14] and Tomanek et al. [17] suggest that adaptation strategies can improve their effectiveness in this domain.

Despite advancements, significant limitations remain in the current state of ASR technology. Existing models often struggle to accurately transcribe disordered speech in real-time therapeutic contexts, where extensive fine-tuning with disorder-specific data is impractical due to data scarcity and privacy concerns. The lack of generalisable models that perform well without fine-tuning limits ASR accessibility in therapy. Additionally, balancing processing speed and accuracy on edge devices remains a challenge, hindering the deployment of efficient real-time solutions in clinical environments. These challenges highlight the need for comprehensive evaluations of pre-trained ASR models on disordered speech to identify models suitable for therapy and guide the development of effective, practical solutions.

3. Datasets

To evaluate the performance of ASR models on both standard and disordered speech in therapeutic applications, the Common Voice dataset [18] for standard speech, the TORGO dataset for dysarthric speech [12], and the University College London Archive of Stuttered Speech (UCLASS) dataset [13] for stuttered speech were selected. While other datasets of disordered speech exist, such as the Nemours Database of Dysarthric Speech [19] and the USC-TIMIT database [20], we selected TORGO and UCLASS because of their extensive use in prior research and the availability of annotations necessary for our analysis.

The Common Voice dataset [18] is a massively multilingual collection of transcribed speech intended for speech technology research and development. For this study, we extracted 600 samples to serve as a benchmark for standard speech. This subset allows us to compare ASR model performance on typical speech patterns against their performance on disordered speech, providing a baseline for evaluations while ensuring computational feasibility.

The TORGO dataset [12] and the University College London Archive of Stuttered Speech (UCLASS) dataset [13] provide speech recordings from individuals with dysarthria and stuttering, respectively. We utilised 600 samples from TORGO, which includes a variety of speaking activities, from single words to spontaneous speech. This subset captures the influence of neurological disabilities on speech patterns and facilitates the evaluation of ASR models' ability to transcribe dysarthric speech accurately. From UCLASS, we used 31 long-form audio samples featuring speech characteristics such as prolongations and repetitions across reading tasks, conversations, and monologues. The smaller number of samples from UCLASS reflects the longer duration and complexity of the recordings, which are essential for analysing ASR performance on stuttered speech in more naturalistic settings. In order to mitigate potential distortions in WER calculations caused by repeated phrases or words that could unfairly inflate error rates, repetitive transcriptions were filtered out during analysis. Both datasets provide complete annotations and quality recordings, making them suitable for our analysis of ASR models on disordered speech in therapeutic contexts.

4. Evaluation Metrics

To evaluate the performance of Automatic Speech Recognition (ASR) models on pathological speech, we employed three key metrics: Word Error Rate (WER), Real-Time Factor (RTF), and BERTScore. These metrics provide insights into transcription accuracy, computational efficiency, and semantic similarity, all crucial in applications where both precision and speed are vital. Word Error Rate (WER) is a standard metric for assessing transcription accuracy by comparing the ASR output to a reference transcription. It is calculated as:

$$\text{WER} = \frac{S + D + I}{N} \tag{1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription. Lower WER values indicate better transcription accuracy, which is particularly important for speech conditions such as stuttering and dysarthria.

Real-Time Factor (RTF) measures the processing speed of an ASR system relative to the duration of the input audio. An RTF value less than 1 indicates that the system processes audio faster than real-time, which is crucial for applications that require immediate feedback. Lower RTF values are preferred, as they signify higher computational efficiency and faster processing.

$$\text{RTF} = \frac{\text{Processing Time}}{\text{Audio Duration}} \tag{2}$$

While WER and RTF assess transcription accuracy and processing speed, they do not capture the semantic similarity between the ASR output and the intended message. This is particularly relevant in disordered speech applications, where conveying the correct meaning is more important than exact word matching. To address this, we employed BERTScore as an additional evaluation metric. BERTScore leverages contextual embeddings from the Bidirectional Encoder Representations from Transformers (BERT) model [21] to evaluate semantic similarity between the predicted transcription and the reference. In the context of disordered speech, traditional metrics like WER might penalise transcriptions where the wording differs but the meaning is preserved. BERTScore addresses this by capturing semantic content, making it relevant for assessing ASR systems handling non-fluent speech [22]. BERTScore operates by generating contextual embeddings for each token in both the candidate (ASR output) and the reference transcription using a pre-trained BERT model. Higher BERTScore values indicate greater semantic similarity between the ASR output and the reference. The similarity between tokens is measured using cosine similarity:

$$\text{sim}(x_i, y_j) = \frac{\mathbf{x}_i \cdot \mathbf{y}_j}{|\mathbf{x}_i|, |\mathbf{y}_j|} \tag{3}$$

where $\mathbf{x}_i$ and $\mathbf{y}_j$ are the embeddings of the i-th and j-th tokens in the candidate and reference, respectively. For each token in the candidate, the most similar token in the reference is identified, and vice versa. Precision (P) and recall (R) are computed as the average of these maximum similarities:

$$P = \frac{1}{|C|} \sum_{x_i \in C} \max_{y_j \in R} \text{sim}(x_i, y_j) \tag{4}$$

$$R = \frac{1}{|R|} \sum_{y_j \in R} \max_{x_i \in C} \text{sim}(y_j, x_i) \tag{5}$$

where C and R are the sets of tokens in the candidate and reference transcriptions, respectively. The final BERTScore is the F1 score combining precision and recall.

$$\text{BERTScore} = 2 \times \frac{P \times R}{P + R} \tag{6}$$

5. Experimental Setup

Experiments were conducted on a Linux virtual machine equipped with an NVIDIA A100 GPU, 83.5 GB of system RAM, and 40.0 GB of GPU RAM, using CUDA 12.1. We utilised PyTorch, Hugging Face Transformers, and librosa for audio processing. The Word Error Rate (WER) was calculated using the jiwer library, while the BERTScore was calculated using the bert-score library. The pre-trained ASR models evaluated range from large transformer-based architectures, Whisper-large-v3-turbo and DistilWhisper, as well as conformer-based models, NVIDIA's Canary and stt-en-fastconformer-ctc-large.

The datasets used were Common Voice for standard speech, TORGO for dysarthric speech, and UCLASS for stuttered speech. Preprocessing was minimal, involving conversion of audio arrays and

bytes audio data to WAV format using librosa for compatibility with the ASR models. The aim was to assess the models' out-of-the-box performance on raw audio data. To simulate real-time conditions, each audio file was processed individually rather than in batches. This approach mirrors live therapeutic settings where immediate transcription is required without the benefits of batch and enables an in-depth assessment of each audio. We measured transcription accuracy (WER) and processing speed (RTF) under these conditions to objectively compare model performance across the datasets. Finally, the BERTScore for each model across the different datasets was calculated.

6. Results and Analysis

The Word Error Rate (WER) and Real-Time Factor (RTF) for each model across the datasets are presented in Table 1. On the TORGO dataset, which contains dysarthric speech, the stt-en-fastconformer-ctc-large model achieved the lowest WER of 0.49, indicating better transcription accuracy for dysarthric speech compared to the other models. DistilWhisper and Whisper-large-v3-turbo had slaightly higher WERs of 0.64 and 0.65, respectively, while Canary had the highest WER at 0.70, suggesting significant challenges in transcribing dysarthric speech. After filtering out audio files with repetitive transcriptions identified in the TORGO dataset to ensure a fairer evaluation, such as instances where the ASR model generated repeated phrases, if included, would artificially inflate the WER and distort the model's true transcription capabilities.

Removing these files allows for a more accurate assessment of the ASR system's performance by focusing on its ability to transcribe natural speech patterns effectively without being skewed by redundant transcription errors. Canary had the second lowest WER at 0.63 after stt-en-fastconformer-ctc-large. On the UCLASS dataset, featuring stuttered speech, Whisper-large-v3-turbo and DistilWhisper achieved the lowest WER of 0.48. The stt-en-fastconformer-ctc-large had a higher WER of 0.61, and Canary again had the highest WER at 0.84. This indicates that the Whisper-based models performed better on stuttered speech. On the Common Voice dataset, representing standard speech, Canary performed best with a WER of 0.14, while the other models had slightly higher WERs of 0.18. This demonstrates their effectiveness in transcribing typical speech, with Canary showing a slight advantage.

Regarding processing speed, the stt-en-fastconformer-ctc-large consistently demonstrated the lowest RTF across all datasets, with values significantly less than 1, indicating it processes audio faster than real-time. For instance, it achieved an RTF of 0.05 on the UCLASS dataset and 0.02 on the Common Voice dataset. This efficiency makes it suitable for real-time applications. The WER for the TORGO dataset (dysarthric speech) was higher than for UCLASS (stuttered speech), yet the BERTScores for TORGO were significantly better. This difference reflects how WER focuses on surface-level transcription accuracy, while BERTScore measures semantic similarity. In the TORGO dataset, despite higher WERs, the ASR models were still able to capture the overall meaning of the speech, resulting in better BERTScores while on the UCLASS dataset, the models had lower WERs but struggled to maintain semantic accuracy, likely due to disfluencies like repetitions and prolongations that characterise stuttered speech.

Table 1
WER and Average RTF Across Datasets

Model	**TORGO**			**UCLASS**		**Common Voice**	
	WER	Filtered WER	RTF	WER	RTF	WER	RTF
Whisper-large-v3-turbo	0.65	0.64	0.11	**0.48**	0.02	0.18	0.05
Canary	0.70	0.63	0.86	0.84	0.06	**0.14**	0.58
DistilWhisper	0.64	0.64	0.13	**0.48**	**0.01**	0.18	0.06
stt_en_fastconformer_ctc_large	**0.49**	**0.49**	**0.04**	0.61	0.05	0.18	**0.02**

The BERTScore analysis offers insights into the semantic similarity between the ASR outputs and the reference transcriptions. Table 2 presents the average BERTScores (F1 scores) for each model across the datasets. On the TORGO dataset, the stt-en-fastconformer-ctc-large model achieved the

highest BERTScore (0.7496), indicating a strong ability to preserve semantic meaning despite potential word-level errors.Whisper-large-v3-turbo and DistilWhisper had moderate scores of 0.6395, while Canary had a slightly higher score of 0.6746. On the UCLASS dataset, BERTScores were much lower across all models, with negative values for Canary (-0.0357), highlighting the difficulty of capturing semantic content in stuttered speech. This result reflects the challenges in transcribing speech that include frequent disfluencies like repetitions and prolongations. In contrast, on the Common Voice dataset, all models performed well, achieving high BERTScores, with Canary leading at 0.8719. These high scores, consistent with the low WERs, indicate that the models were able to accurately transcribe and preserve semantic content in standard speech.

Table 2
Average BERTScores for Each Model and Dataset

Dataset	Model	Precision	Recall	F1 Score
TORGO	Whisper-large-v3-turbo	0.6829	0.6006	0.6395
	Canary	0.7182	0.6363	0.6746
	DistilWhisper	0.6829	0.6006	0.6395
	stt-en-fastconformer-ctc-large	**0.7374**	**0.7654**	**0.7496**
UCLASS	Whisper-large-v3-turbo	**0.4609**	**0.0764**	**0.2600**
	Canary	0.1836	-0.2309	-0.0357
	DistilWhisper	0.4525	0.0740	0.2552
	stt-en-fastconformer-ctc-large	0.1792	-0.0740	0.0492
Common Voice	Whisper-large-v3-turbo	0.8250	0.8165	0.8208
	Canary	**0.8784**	**0.8655**	**0.8719**
	DistilWhisper	0.8250	0.8165	0.8208
	stt-en-fastconformer-ctc-large	0.7210	0.7405	0.7307

Table 3: Sentence-level BERTScores for ASR Models on Different Datasets

Dataset	Model	Prediction and Scores
Torgo	**Reference:**	except in the winter when the ooze or snow or ice prevents
	Whisper-large-v3-turbo	I'm still in the reading. The rules were still what I'd command. **Precision:** 0.0651, **Recall:** -0.1521, **F1 Score:** -0.0442
	Canary	Except in the winning, we also use the experiments. **Precision:** 0.1439, **Recall:** -0.1320, **F1 Score:** 0.0037
	DistilWhisper	I'm still in the reading. The rules were still what I'd command. **Precision:** 0.0651, **Recall:** -0.1521, **F1 Score:** -0.0442
	stt-en-fastconformer-ctc-large	except in the wining we also po less human **Precision:** -0.0958, **Recall:** -0.2074, **F1 Score:** -0.1504
Common Voice	**Reference:**	Wait for the end of the war.
	Whisper-large-v3-turbo	Wait for the end of the wall. **Precision:** 0.6265, **Recall:** 0.6266, **F1 Score:** 0.6272

Continued on next page

Table 3 – continued from previous page

Dataset	Model	Prediction and Scores
	Canary	Wait for the end of the world. **Precision:** 0.7143, **Recall:** 0.7144, **F1 Score:** 0.7148
	DistilWhisper	Wait for the end of the wall. **Precision:** 0.6265, **Recall:** 0.6266, **F1 Score:** 0.6272
	stt-en-fastconformer-ctc-large	wait for the end of the war **Precision:** 0.8996, **Recall:** 0.8647, **F1 Score:** 0.8823
UClass	**Reference:**	Um, U I go to Trinthoe School, and I like doing football, art, design, technology, maths and, lots of other subjects. My best hobby is football, I scored 18 goals um er last year, and. Er. (Interviewer): Tell me about your friends at schools, your teachers, things you like to do at school. Yeah. {Interviewer}. My best friend's XXXX XXXXXX. Um. Went round to his house, um, {block}about two days ago. Um. Er We wernt out on our bikes, Um. M. (Interviewer). (S): Lost World. Um. Lost World is my best film. Um, {block}IT was quite scary, um. About dinosaurs. (I): What happened? (S): U it's sequel to Jrassic Park ANNNNN, there some dinosaurs escaped and, well actually came back. Um. (I): What were they trying to do? (S): Well they trying to catch them and, and going put them back um in the right place. (I): How do they, what, how do they discover? (S): They walking through the forest and they SSSS{block}- heard noises and saw them. (I): Right, OK, that's great.
	Whisper-large-v3-turbo	I go to Trinidad School and I like doing football, art, design technology, maths and a lot of other subjects. My best hobby is football. I scored 18 goals last year. Tell me about your friends at school and teachers. Yeah. What do you like to do in school? My best friend, Sean Waters, went around to his house about two days ago. What did you do then? I went out on the bikes. Um... OK, tell me about Lost World. Lost World. Um... Lost World's my best film. Um... It was quite scary, um... About dinosaurs. What happened? It was a sequel to Jurassic Park and... ...and some dinosaurs escaped and... ...when she came back. Um... What were they trying to do? What was going on? We were trying to catch them and... I think I put them back in the right place. How did they discover them? They were walking through the forest and they just heard noises and saw them. Right. Okay, that's great. Tell me about what you've been doing today. I went out for the bandstand. **Precision:** 0.4458, **Recall:** 0.2156, **F1 Score:** 0.3293
	Canary	I go to Chinnanhoe School and I like doing football, art, design, technology, maths and a lot of other subjects. My best friend Sean Waters, my best hobby is football. I scored eighteen goals last year. Tell me about what you've been doing today. **Precision:** 0.5696, **Recall:** -0.1899, **F1 Score:** 0.1631

Continued on next page

Table 3 – continued from previous page

Dataset	Model	Prediction and Scores
	DistilWhisper	I go to Trinidad School and I like doing football, art, design technology, maths and a lot of other subjects. My best hobby is football. I scored 18 goals last year. Tell me about your friends at school and teachers and things you like doing in school. My best friend, Sean Waters, went around to his house about two days ago. I went out on the bikes. I went out on the bikes. OK, tell me about Lost World. Lost World. Lost World is my best film. It was quite scary. About dinosaurs. What happened? I was sequel to Jurassic Park and some dinosaurs back in the right place. How did they discover them? They were walking through the forest and they just heard noises and saw them. Right. Okay, that's great. Tell me about what you've been doing today. I went out for the band's town **Precision:** 0.4827, **Recall:** 0.0843, **F1 Score:** 0.2771
	stt-en-fastconformer-ctc-large	oh got ch school and i love doing football design technology math and subjects my best hobby is football a sco at eighteen goals last year yeah the best friend showan waters went to his house about two days ago bent on the bike los what my best film oh is quite scary about doneances secre to jusic park ands escaped and where she came back trying to catch them and put them back on in our place i look at through to foolish the he noises and soom i went out for the band stand **Precision:** 0.0006, **Recall:** -0.2668, **F1 Score:** -0.1350

7. Discussion

The evaluation results demonstrate that while all models perform well on standard speech, their effectiveness on disordered speech varies significantly. The stt-en-fastconformer-ctc-large model stands out in its performance on dysarthric speech in the TORGO dataset, achieving the lowest WER and highest BERTScore, indicating a strong ability to capture semantic content. Its capability to process audio faster than real-time (low RTF) makes it particularly suitable for real-time therapeutic applications, where timely feedback is essential for maintaining patient engagement and supporting therapeutic progress. However, the stt-en-fastconformer-ctc-large struggles more with stuttered speech in the UCLASS dataset, showing higher WER and lower BERTScore compared to the Whisper-based models, which performed better in this context. This reflects the unique challenges posed by different speech disorders, suggesting that no single model may be universally effective across all types of disordered speech. It is important to note that the differences across datasets may also be partly attributed to dialectal variations, such as British or American English, as well as other factors like recording conditions, thus impacting ASR model performance.

The trade-off between transcription accuracy and processing speed is a key consideration. While the Whisper models offer superior accuracy for stuttered speech, the stt-en-fastconformer-ctc-large provides a better balance of accuracy and speed, making it more appropriate for dysarthric speech, where preserving real-time interaction is critical. The relatively higher speed and competitive accuracy of the stt-en-fastconformer-ctc-large across datasets emphasise its hardware efficiency, making it a compelling option in scenarios constrained by processing power and latency requirements. Moreover, the discrepancies between WER and BERTScore, particularly in the TORGO dataset, highlight the importance of semantic evaluation in ASR systems for speech therapy. While WER is valuable for measuring surface-level accuracy, BERTScore provides a more nuanced understanding of how well the model captures the intended meaning, which is especially important in therapeutic contexts.

8. Conclusion and Future Work

This study evaluated pre-trained ASR models:Whisper-large-v3-turbo, Canary, DistilWhisper and stt-en-fastconformer-ctc-large on both standard and disordered speech, focusing on WER, RTF and BERTScore. The stt-en-fastconformer-ctc-large model excelled in dysarthric speech (TORGO dataset), achieving the lowest WER and highest BERTScore while processing audio faster than real-time. This balance between speed and accuracy makes it highly suitable for real-time therapeutic applications. In contrast, Whisper-large-v3-turbo outperformed the other models on stuttered speech (UCLASS dataset), achieving the lowest WER, although it was slower than stt-en-fastconformer-ctc-large on the TORGO and Common Voice datasets.

These results highlight the need to select ASR models based on the specific speech disorder. While stt-en-fastconformer-ctc-large performed well with dysarthric speech, Whisper-based models were more effective for stuttered speech, though with higher computational demands. This suggests no single model is ideal for all speech disorders, requiring a balance between accuracy and processing speed in real-time applications. Future work should focus on fine-tuning models with disorder-specific data, improving generalisation across different disorders and environments, and optimising for edge computing to reduce latency and provide privacy required for therapeutic settings. Integrating multimodal inputs, like visual cues, may also enhance recognition accuracy.

Acknowledgments

The Science Foundation Ireland Centre for Research Training in Digitally Enhanced Reality (d-real) under Grant Nos. 18/CRT/6224 and 19/FFP/6917 as well as the ADAPT SFI Research Centre for AI-Driven Digital Content Technology under Grant No. 13/RC/2106-P2 supported this research. The author has designated a CC BY public copyright license for any author-accepted manuscript resulting from this submission, in accordance with Open Access principles.

References

[1] J. Fruitet, M. Fouillen, V. Facque, H. Chainay, S. D. Chalvron, F. Tarpin-Bernard, Engaging with an embodied conversational agent in a computerized cognitive training: an acceptability study with the elderly, in: ACM International Conference Proceeding Series, Association for Computing Machinery, 2023, pp. 359–362. doi:10.1145/3610661.3616130.

[2] P. Kulkarni, O. Duffy, J. Synnott, W. G. Kernohan, R. McNaney, Speech and language practitioners' experiences of commercially available voice-assisted technology: Web-based survey study, JMIR Rehabilitation and Assistive Technologies 9 (2022). doi:10.2196/29249.

[3] J. Cleland, S. Lloyd, L. Campbell, L. Crampin, J. P. Palo, E. Sugden, A. Wrench, N. Zharkova, The impact of real-time articulatory information on phonetic transcription: Ultrasound-aided transcription in cleft lip and palate speech, Folia Phoniatrica et Logopaedica 72 (2020) 120–130. doi:10.1159/000499753.

[4] A. L. Georgescu, A. Pappalardo, H. Cucu, M. Blott, Performance vs. hardware requirements in state-of-the-art automatic speech recognition, 2021. doi:10.1186/s13636-021-00217-4.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.

[6] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, 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.

[7] M. Pernon, F. Assal, I. Kodrasi, M. Laganaro, Perceptual classification of motor speech disorders:

The role of severity, speech task, and listener's expertise, Journal of Speech, Language, and Hearing Research 65 (2022) 2727–2747. doi:10.1044/2022_JSLHR-21-00519.

[8] H. P. Rowe, S. E. Gutz, M. F. Maffei, K. Tomanek, J. R. Green, Characterizing dysarthria diversity for automatic speech recognition: A tutorial from the clinical perspective, Frontiers in Computer Science 4 (2022). doi:10.3389/fcomp.2022.770210.

[9] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: Vision and challenges, IEEE internet of things journal 3 (2016) 637–646.

[10] M. Satyanarayanan, The emergence of edge computing, Computer 50 (2017) 30–39.

[11] J. R. Green, R. L. MacDonald, P. P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner, P. C. Nelson, K. Tomanek, Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 4, International Speech Communication Association, 2021, pp. 3051–3055. doi:10.21437/Interspeech.2021-1384.

[12] F. Rudzicz, A. K. Namasivayam, T. Wolff, The torgo database of acoustic and articulatory speech from speakers with dysarthria, Language Resources and Evaluation 46 (2012) 523–541. doi:10.1007/s10579-011-9145-0.

[13] P. Howell, S. Davis, J. Bartrip, The university college london archive of stuttered speech (uclass), Journal of Speech, Language, and Hearing Research 52 (2009) 556–569. doi:10.1044/1092-4388(07-0129).

[14] V. Mitra, Z. Huang, C. Lea, L. Tooley, S. Wu, D. Botten, A. Palekar, S. Thelapurath, P. Georgiou, S. Kajarekar, J. Bigham, Analysis and tuning of a voice assistant system for dysfluent speech, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 4, International Speech Communication Association, 2021, pp. 3086–3090. doi:10.21437/Interspeech.2021-2006.

[15] D. Mulfari, M. Villari, A voice user interface on the edge for people with speech impairments, Electronics (Switzerland) 13 (2024). doi:10.3390/electronics13071389.

[16] J. Tobin, K. Tomanek, Personalized automatic speech recognition trained on small disordered speech datasets, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, volume 2022-May, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6637–6641. doi:10.1109/ICASSP43922.2022.9747516.

[17] K. Tomanek, J. Tobin, S. Venugopalan, R. Cave, K. Seaver, J. R. Green, R. Heywood, Large language models as a proxy for human evaluation in assessing the comprehensibility of disordered speech transcription, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 10846–10850. doi:10.1109/ICASSP48485.2024.10447177.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, arXiv preprint arXiv:1912.06670 (2019).

[19] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, H. T. Bunnell, The nemours database of dysarthric speech, in: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, volume 3, IEEE, 1996, pp. 1962–1965.

[20] J. S. Garofolo, Timit acoustic phonetic continuous speech corpus, Linguistic Data Consortium, 1993 (1993).

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[22] J. Tobin, Q. Li, S. Venugopalan, K. Seaver, R. Cave, K. Tomanek, Assessing asr model quality on disordered speech using bertscore, International Speech Communication Association, 2022, pp. 26–30. doi:10.21437/s4sg.2022-6.