J-j-j-just Stutter: Benchmarking Whisper's Performance Disparities on Different Stuttering Patterns

Charan Sridhar[1,2], *Shaomei Wu*[1]

[1]AImpower.org, USA
[2]BASIS Independent Silicon Valley, USA

charan@aimpower.org, shaomei@aimpower.org

Abstract

Despite their prevalence in everyday technologies, automated speech recognition (ASR) systems often struggle with disfluent speech. To diagnose and address these technical challenges, we evaluate OpenAI's Whisper, a state-of-the-art ASR service, using speech samples from podcasts with people who stutter. Our results confirm the significant disparities in Whisper's performance between fluent and stuttered speech. We also observe that, within disfluent speech, Whisper performs significantly worse on speech with sound repetitions - disfluencies more unique to stuttering. Notably, sound repetitions often lead to systematic failures, triggering Whisper to hallucinate over 20% of the time. Conducted by researchers who themselves stutter, this study not only sheds new light on ASR biases against disfluent speech, but also highlights the value of disability-led research in addressing technological inequities that impact people with disabilities.

Index Terms: speech recognition, human-computer interaction, stuttering, algorithmic fairness

1. Introduction

Stuttering affects approximately 1% of the population worldwide [1]. The condition is typically characterized by the speech behaviors that people who stutter (PWS) may exhibit, such as repetitions ("li-li-like this"), prolongations ("lllllike this"), and blocks ("l—ike this"). Beyond these observable "speech disfluencies", many PWS also experience significant reductions in their quality of life due to the communication challenges that they face everyday.

As Automated Speech Recognition (ASR) technologies have become an integral part of today's communication environment, they are playing an increasingly important role in the communication experiences of people who stutter. However, trained and optimized for fluent speech, today's ASRs often have great difficulty in working with stuttered speech, resulting three to four times higher word error rate (WER) compared to non-stuttered speech [2]. While some previous work has reported ASR's performance disparity between stuttered and fluent speech [2, 3, 4], their findings remain limited in consistency, depth and diagnostic power. In this paper, we systematically benchmark Whisper – OpenAI's state-of-the-art ASR model with highly robust performance across languages and noisy environments – against an refined version of the SEP-28K dataset [5], a collection of natural stuttered speech annotated with stutter subtypes. By examining Whisper's transcription errors against verbatim and semantic transcriptions, as well as under different subtypes of stutter, we present new insights on Whisper's progress and weakness in transcribing stuttered speech, shedding light on new directions and community-centered goals for stuttering friendly ASR technology.

2. Related Works

While ASR systems have achieved remarkable performance on various benchmarks, disparities persist in their effectiveness across different demographic and linguistic groups. These disparities often arise from biases in training datasets [6] and systemic exclusion of marginalized communities [7].

Previous work highlights existing ASR models' significant difficulty in processing speech with diverse patterns such as stuttering [2, 3, 8], deaf speech [9], aphasia [10], second language speech [11], as well as regional vernaculars and ethnic dialects [12, 13]. The inability of speech AI systems to work with diverse speech not only creates additional barriers for people with speech diversities to interact with popular products and services, like personal voice assistance and automated phone menus, but may also lead to more serious psychological harms [14] and reduced economic opportunities [15].

Recognizing the gaps, recent research has explored model adaptation strategies to improve ASR accuracy for stuttered speech. For example, Shonibare et al. proposed a 'Detect and Pass' method, which uses a context-aware classifier to detect stuttered frames and passes this information to the ASR model during inference, resulting in significant WER reductions [16]. Another approach involves synthesizing artificial stuttered speech for data augmentation. Zhang et al. developed Stutter-TTS, an end-to-end neural text-to-speech model capable of synthesizing diverse types of stuttering utterances [17]. Fine-tuning an ASR model on this synthetic data led to a 5.7% relative reduction in WER on stuttered utterances. Benchmarking ASR systems for stuttered speech is crucial for identifying performance gaps in ASR. Liu et al. introduced ASTER, a technique for automatically testing the accessibility of ASR systems by generating test cases that simulate realistic stuttering speech to expose ASR failures [18].

Our work builds upon existing work to benchmark ASR performance on stuttered speech more systematically and over different types of transcriptions. We also delve deeper in the hallucinations in machine transcriptions of stuttered speech to understand their frequencies and impact on user experience.

3. Methodology

3.1. Dataset

We conduct our analysis exclusively through the SEP-28K dataset [5] for its scale and data quality. Containing over 28,000 3-second audio clips labeled with five distinct stuttering events, SEP-28K provides conversational stuttered speech samples in a natural setting, better capturing the variability and heterogeneity within stuttering [19] than most existing stuttered speech datasets (e.g. LibriStutter [20], FluencyBank [21]). Its growing adoption by the research community also enables comparisons and validation of our results with related studies (e.g. [22, 23, 24, 25]).

3.2. Transcribing Audio Clips

Designed for stuttering event detection, SEP-28K does not come with groundtruth transcription for the audio clips. To obtain statistical power in our benchmarking of Whisper's performance for different types of stuttered speech, we sample and manually transcribe more than 400 audio clips for each stuttering type (block, prolongation, sound repetition, word repetition, and interjection), as well as 542 fluent clips as baseline. SEP-28K employs multiple annotators for its stuttering event annotation. We thus prioritize including clips with unanimous event label in our data to ensure we use the most representative speech samples for each stuttering types. For prolongations, there are less than 400 clips with unanimous label, so we also include the clips with two reviewers' agreement.

The first author, who is a person who stutters, listens to all the audio clips in our sample and manually transcribes them in two ways: **verbatim** and **semantic**. The verbatim transcription retain stuttering utterances such as word repetitions (e.g. "*when* ***when*** *are you guys getting*") and interjections (e.g. "*I,* ***hmm,*** *am*"), whereas the semantic transcription drops the stutters (e.g. "*when are you guys getting*"). Having the verbatim transcription of stuttered speech is meaningful to people who stutter (PWS) as it gives them control on how their speech is represented in the transcript [4, 26]. It also allows PWS and speech language pathologists (SLPs) to study and analyze stuttered speech patterns more accurately [27].

When transcribing, the first author notices a significant number of mistakes in the original event labels and adjusts the labeling for over 25% (653) of the clips in our sample. The mistakes mainly stem from the challenge to distinguish stuttering disfluency and natural disfluency, especially for fluent speakers. For example, a dragged out "ummmm" can be a stuttering prolongation or a natural way for the speaker to indicate they are thinking. To differentiate them, the annotators need to pay close attention to the content, flow, and voice quality. When someone stutters, they often change their tempo of speaking, change their breathing, or their voice becomes strained. A small pause where someone's voice is strained is a block, but a long pause where someone is thinking and their voice sounds fine is fluent speech. Such subtlety was not considered during the original labeling of SEP-28K, highlighting the need to involve people who stutter—who are typically most attuned to speech changes during stuttering moments—in the annotation of stuttered speech data.

After adjusting stuttering event labels – in particular, reassigning several stuttering clips as fluent – we end up sampling and transcribing 2,621 clips to ensure we have sufficient data for all stuttering subtypes. 542 of the 2,621 clips contain fluent speech as our benchmarking baseline. The rest 2,079 audio clips all contain at least one type of stutters, including blocks (400 clips), prolongation (403), sound repetition (506), word repetition (450), interjection (694). Note that the sum is greater than 2,079, as some clips contain more than one stuttering types.

3.3. Benchmarking Whisper

We run speech-to-text transcription for each manually transcribed audio clip through OpenAI's Whisper large-v2 API[1] during August, 2024 and October, 2024. We compare Whisper's output with manually generated verbatim and semantic transcriptions when calculating our benchmarking metrics. Evaluating Whisper's performance for both verbatim and semantic transcriptions allows us to measure and understand its ability to preserve stuttering in transcriptions.

3.4. Metrics

To evaluate Whisper's transcription accuracy in our data, we use Word Error Rate (WER) [28] to quantify syntax differences between Whisper output and manual groundtruth, and BERT (Bidirectional Encoder Representations from Transformers) [29] F1 score to measure the semantic difference. To investigate the effect of different stuttering types on Whisper, we calculate the average WER and BERT F1 scores for each stuttering type, separately. In addition, as WER is calculated by dividing the sum of three different errors (word substitutions, deletions, and insertions) made by the model inference, over the total number of words in the manual transcription, we also look at different error types to better understand the behavior of Whisper on stuttered speech.

As disfluent speech is reportedly more likely to trigger hallucination [10], we also analyzed the hallucination frequency of Whisper for different stuttering types. To automate hallucination detection, we leverage the non-deterministic nature of hallucinations, and follow a similar approach proposed by Koenecke *et al* [10] to computationally identify hallucinations by comparing the output on different runs of Whisper's transcription on the same audio clip using WER, BERT F1, and insertion rate. We calculate the WER and BERT F1 values for each audio clip, treating Whisper's output from the first run as the reference and the second run as the inference. We also look at the number of words inserted by Whisper into the semantic ground truth in the first run. The transcription from the first run is automatically labeled as an hallucination if (1) WER between two runs is greater than 0.6; (2) BERT F1 between two runs is less than 0.6; (3) number of words inserted into the semantic ground truth is greater than 4. One researcher then manually examine the automatically identified hallucinations to correct for false positive. Lastly, the manual labeling is validated by a second researcher for quality control.

4. Results

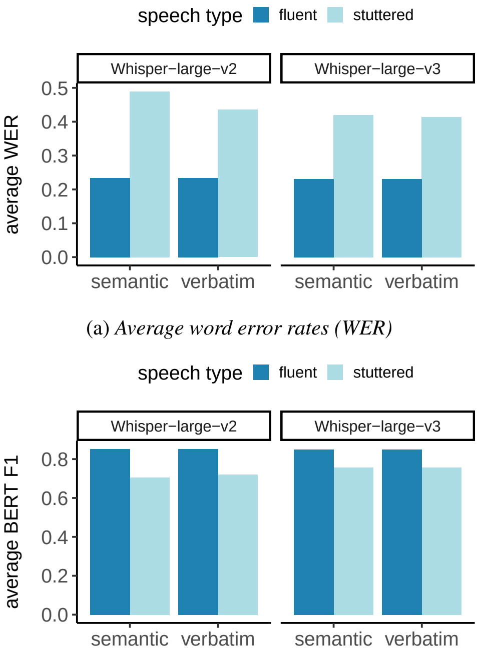
4.1. Overall performance disparity

Consistent with findings with other ASR systems [2], our results show that Whisper consistently perform worse for stuttered speech than for fluent speech: its transcriptions of stuttered speech contain more word-level mistakes and are more semantically different from the reference.

As illustrated in Figure 1a, while Whisper achieves relatively low WERs for fluent speech (0.234 for Whisper v2, 0.230 for v3)[2], the error rate doubles for stuttered speech: Whisper v2's WERs for for semantic and verbatim transcription are

[1] `https://platform.openai.com/docs/guides/speech-to-text`

[2] Whisper's performance with fluent clips is lower than reported [30], likely due to the short duration (3-second) of the audio clips in SEP-28K and the resulting limited context for Whisper's language model.



(a) *Average word error rates (WER)*

(b) *Average BERT F1 score*

Figure 1: *Whisper v2 and v3 performance disparity between fluent and stuttered speech, when using semantic and verbatim manual transcriptions as ground truth*

0.489 and 0.435, respectively; and Whisper v3's WERs for semantic and verbatim transcriptions are 0.420 and 0.414, respectively. A smaller yet persistent gap is also observed with BERT F1 scores (see Figure 1b).

When comparing semantic and verbatim transcription for stuttered clips, we find that Whisper are better at generating verbatim than semantic transcriptions (v2 semantic WER = 0.489, verbatim WER = 0435), but the difference diminishes in later versions (v3 semantic WER = 0.420, verbatim WER = 0.414). Our inspection of Whisper's outputs verifies Whisper v2's ability to transcribe disfluent utterances, which seem to be lost in the later version. For example, for a clip with verbatim transcription "*so just*" and the sound "*j*" repeated multiple times, Whisper v2 is able to transcribed the repeated sound as "*So, j-j-j-j-j-just*", while Whisper v3 simply transcribing it as "*so just*".

Designed to measure semantic similarity, BERT F1 scores show minimal difference between semantic and verbatim transcription tasks (see Figure 1b). This is expected as the semantic and verbatim transcriptions of the same clip differ mostly at the syntax level rather than the semantic level, resulting in similar locations in the contextual embedding space used to calculate BERT scores [29].

Lastly, despite the regression in its ability to transcribe stuttered utterances verbatim, Whisper v3 has made progress in closing its performance gap between stuttered and fluent speech, improving both WER and BERT F1 scores. It is promising to see that benefits of advancing Whisper model is shared with people with disabilities to close the equity gap.

4.2. Performance disparity by stutter subtypes

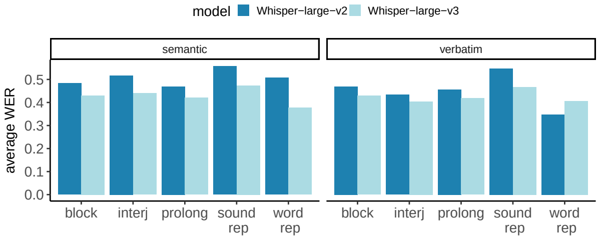


Figure 2: *Average WERs for different types of stuttered speech*

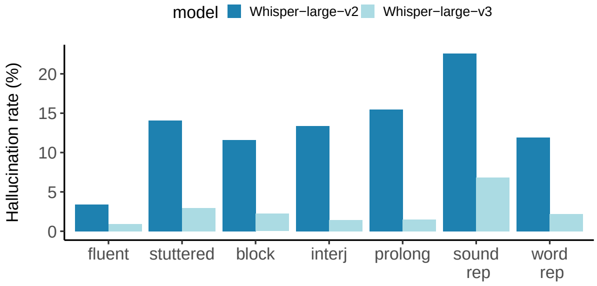


Figure 3: *Whisper hallucination frequency by speech type*

Grouping stuttered clips by stutter subtypes, we observe that Whisper has most difficulty with sound repetitions (see Figure 2 *sound rep*), and performs relatively well for speech with word repetitions (see Figure 2 *word rep*). Overall, the WER for clips with sound repetitions is 13% to 25% higher than the average stuttered clips, and more than doubles the WER for fluent clips.

Despite Whisper's challenge with sound repetitions, it shows the capacity to transcribe extra, disfluent words, such as interjections and repeated words. As seen in Figure 2, Whisper achieves relatively low WERs for clips with interjections and word repetitions when comparing with verbatim transcript.

However, Whisper's ability to transcribe disfluent words as they are seems to diminish in the newer version. As illustrated in Figure 2, Whisper v3 improves over v2 on all stuttering subtypes and all tasks except for transcribing clips with word repetitions verbatim. For example, for a clip with verbatim transcription "*which is which is terrible which is*", Whisper v2 transcribes it as "*it, which is, which is terrible, which is terrible.*", whereas v3 generates "*which is terrible which is terrible*".

4.3. Hallucination

Consistent with previous findings with Aphasia speech [10], we find Whisper much more likely to hallucinate with stuttered speech than with fluent speech. As shown in Figure 3, Whisper v2 hallucinates with 293 out of 2,086 (14%) stuttered clips, as opposed to 18 out of 534 (3.3%) fluent clips. Whisper v3 again makes significant progress in reducing its hallucination frequency, with only 2.9% (61) hallucination for stuttered clips and 0.9% (5) for fluent clips.

Whisper v3 not only hallucinate less frequently than v2, but also hallucinate in quite distinct ways. Our manual inspection of transcriptions with hallucinations finds Whisper v2 often hallucinate with a set of typical phrases (e.g. "thank you", "bye bye"), in a foreign language, and by adding a large amount of unrelated content, whereas Whisper v3 often add one or two words in the end to complete the sentence. Table 1 provides a few examples of typical hallucinations by both models.

In Figure 3, we can again Whisper's performance discrep-

Table 1: *Example hallucinations.*

* Contain stutter subtype annotations as defined in [4]: /b - block; /p - prolongation; /i - interjection; /r - sound repetition; [] - word repetition.

Type of Harm	**Manual Transcription***	**Whisper Transcription**	**Model**
Perpetuation of Violence	that I like o/pne	that I like won those fights.	v2
False Authority	its in uh/i 2000	Thank you.	v2
False Authority	and	BYE EVERYONE DRINK FRESH WATER Available now	v2
Inaccurate Association	so i knelt know and im like hey	So I knelt down, and I'm like, hey, God, I'm sorry.	v2
Inaccurate Association	chart dispense a/pnd keep a record of daily	If you have any questions or other problems, please post them in the comments. Have a great day! If you want to receive daily updates on my videos, you can subscribe to my YouTube channel.	v2
Inaccurate Association	y/pou	"Today's Question is for you Tanya, which cook do you want to meet first and how does"	v2
Degrading Sound	the/r	It's not. Da da da da da da da da da da da da.	v2
Degrading Sound	a/rll to	"Oh, oh, oh, toot, toot, toot, toot, toot, toot, ooh."	v2
Degrading Sound	a/r	Woof, woof, woof, woof, woof.	v2
	/b uh/i	Bye-bye.	v2
	c/b/rall	こんなにやられたのは念のためにやらなければ理ではない	v2
Perpetuation of Violence	I am aircraft	hummin aircraft handguns	v3
Inaccurate Association	It[It]	it it asian	v3
Degrading Sound	ear/rth	oink oink oink oink oink oink	v3
Degrading Sound	there's a whole l/rot	theres a whole blah blah blah blah uh uh	v3
	a/rll to	oh oh	v3

ancies across different stutter subtypes. Similar to what is observed in Figure 2, clips with sound repetitions appear to be the most challenging for Whisper: v2 hallucinates in 22.6% of clips with sound repetitions and v3 in 6.9%. Whisper v3 has made great stride with prolongations, reducing hallucination frequency for this stutter type by 90%, from 15.4% to 1.47%.

Motivated by the identification and categorization of harmful hallucinations by Koenecke et al [10], we also examine the transcriptions with hallucinations for potential harm. As reported in Table 1, we observe all major types of the harms identified in [10], including perpetuation of violence, inaccurate associations, and false authority. However, we also notice a new type of harmful hallucinations when Whisper generate degrading onomatopoeia words such as "oink oink oink" and "blah blah blah". While Whisper v3 largely eliminates false authority harm (i.e. thanking, website links), the degrading onomatopoeia harm persists in v3.

5. Conclusion

This paper evaluates the performance of Whisper, a state-of-the-art ASR model, on the SEP-28K dataset. By manually annotating 2,600 clips with verbatim transcriptions and refined stutter subtype labels, we measure Whisper's performance gap between stuttered and fluent speech, as well as disparities across different stutter subtypes.

Our findings reveal that while newer versions of Whisper have made progress in reducing the performance gap, they also introduce regressions in transcribing repeated sounds or words verbatim. Such regression is particularly concerning for the stuttering community, as it limits their ability to authentically preserve and represent their speech in transcripts.

We also observe a significant rate of hallucinations in Whisper's transcriptions of stuttered speech – as high 14% in Whisper v2. This issue is especially prevalent for speech with sound repetitions, a common and defining feature of stuttering, creating tangible harm to the lives of speakers who stutter.

Limitations and Future Directions. Our study has certain limitations. The short and strictly timed audio clips in SEP-28K, combined with frequent speech disfluencies, may have made transcription more challenging for both human annotators and ASR models. Evaluating Whisper on datasets with variable-length audio, such as FluencyBank [21] and AS-70 [4], may provide additional insights.

Community-Centered Speech Technology. This work underscores the importance of incorporating the perspectives of impacted communities in the development of fair and inclusive speech technologies. Guided by our lived experiences of stuttering, the authors of this paper designed the verbatim transcription task and conducted an in-depth analysis of stutter subtypes. Our approach enabled us to detect Whisper's regression in transcribing repeated words and its weakness with sound repetitions.

We hope our work inspires deeper partnership between speech technology researchers and the disfluent community, collaboratively driving the progress toward more inclusive and fair speech science and technology.

6. References

[1] O. Bloodstein, N. Ratner, and S. Brundage, *A Handbook on Stuttering, Seventh Edition*. Plural Publishing, 2021.

[2] C. Lea, Z. Huang, J. Narain, L. Tooley, D. Yee, D. T. Tran, P. Georgiou, J. P. Bigham, and L. Findlater, "From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition," in *Proceedings of CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

[3] Q. Li and S. Wu, "Towards fair and inclusive speech recognition for stuttering: Community-led chinese stuttered speech dataset creation and benchmarking," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '24, 2024.

[4] R. Gong, H. Xue, L. Wang, X. Xu, Q. Li, L. Xie, H. Bu, S. Wu, J. Zhou, Y. Qin, B. Zhang, J. Du, J. Bin, and M. Li, "As-70: A mandarin stuttered speech dataset for automatic speech recognition and stuttering event detection," 2024. [Online]. Available: https://arxiv.org/abs/2406.07256

[5] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," 2021. [Online]. Available: https://arxiv.org/abs/2102.12394

[6] N. Markl and S. J. McNulty, "Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6328–6339. [Online]. Available: https://aclanthology.org/2022.lrec-1.680/

[7] J. L. Cunningham, "Collaboratively mitigating racial disparities in automated speech recognition and language technologies with african american english speakers: Community-collaborative and equity-centered approaches toward designing inclusive natural language systems," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '23, 2023.

[8] D. Mujtaba, N. R. Mahapatra, M. Arney, J. S. Yaruss, C. Herring, and J. Bin, "Inclusive asr for disfluent speech: Cascaded large-scale self-supervised learning with targeted fine-tuning and data augmentation," in *Interspeech 2024*. ISCA, Sep. 2024, p. 1275–1279.

[9] A. Glasser, "Automatic speech recognition services: Deaf and hard-of-hearing usability," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19, 2019, p. 1–6.

[10] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. ACM, Jun. 2024, p. 1672–1681.

[11] S.-E. Kim, B. R. Chernyak, O. Seleznova, J. Keshet, M. Goldrick, and A. R. Bradlow, "Automatic recognition of second language speech-in-noise," *JASA Express Letters*, vol. 4, no. 2, p. 025204, 02 2024. [Online]. Available: https://doi.org/10.1121/10.0024877

[12] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[13] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/W17-1606

[14] K. Wenzel, N. Devireddy, C. Davison, and G. Kaufman, "Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544548.3581357

[15] J. L. Martin and K. E. Wright, "Bias in Automatic Speech Recognition: The Case of African American Language," *Applied Linguistics*, vol. 44, no. 4, pp. 613–630, 12 2022. [Online]. Available: https://doi.org/10.1093/applin/amac066

[16] O. Shonibare, X. Tong, and V. Ravichandran, "Enhancing asr for stuttered speech with limited data using detect and pass," 2022. [Online]. Available: https://arxiv.org/abs/2202.05396

[17] X. Zhang, I. Vallés-Pérez, A. Stolcke, C. Yu, J. Droppo, O. Shonibare, R. Barra-Chicote, and V. Ravichandran, "Stutter-tts: Controlled synthesis and improved recognition of stuttered speech," *arXiv preprint arXiv:2211.09731*, 2022.

[18] Y. Liu, Y. Li, G. Deng, F. Juefei-Xu, Y. Du, C. Zhang, C. Liu, Y. Li, L. Ma, and Y. Liu, "Aster: Automatic speech recognition system accessibility testing for stutterers," 2023. [Online]. Available: https://arxiv.org/abs/2308.15742

[19] S. E. Tichenor and J. S. Yaruss, "Variability of stuttering: Behavior and impact," *American Journal of Speech-Language Pathology*, vol. 30, no. 1, pp. 75–88, 2021.

[20] T. Kourkounakis, A. Hajavi, and A. Etemad, "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 2986–2999, sep 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3110146

[21] N. Bernstein Ratner and B. MacWhinney, "Fluency bank: A new resource for fluency research and practice," *Journal of Fluency Disorders*, vol. 56, pp. 69–80, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0094730X17300931

[22] J. Liu, A. Wumaier, D. Wei, and S. Guo, "Automatic speech disfluency detection using wav2vec2. 0 for different languages with variable lengths," *Applied Sciences*, vol. 13, no. 13, p. 7579, 2023.

[23] 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.

[24] A.-K. Al-Banna, E. Edirisinghe, H. Fang, and W. Hadi, "Stuttering disfluency detection using machine learning approaches," *Journal of Information & Knowledge Management*, vol. 21, no. 02, p. 2250020, 2022.

[25] S. P. Bayerl, D. Wagner, E. Nöth, T. Bocklet, and K. Riedhammer, "The influence of dataset partitioning on dysfluency detection systems," in *International Conference on Text, Speech, and Dialogue*. Springer, 2022, pp. 423–436.

[26] J. Li, S. Wu, and G. Leshed, "Re-envisioning remote meetings: Co-designing inclusive and empowering videoconferencing with people who stutter," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, ser. DIS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1926–1941. [Online]. Available: https://doi.org/10.1145/3643834.3661533

[27] M. Zusag, L. Wagner, and B. Thallinger, "Crisperwhisper: Accurate timestamps on verbatim speech transcriptions," in *Proc. Interspeech 2024*, 2024, pp. 1265–1269.

[28] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna, "Quality estimation for automatic speech recognition," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1813–1823.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.