



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

Charan Sridhar^{1,2}, Shaomei Wu¹

¹AImpower.org, USA

²BASIS Independent Silicon Valley, USA

charansr@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 model, using speech samples from podcasts with people who stutter. Our results show significant disparities in Whisper’s performance between fluent and stuttered speech. Within disfluent speech, Whisper performs significantly worse on speech with sound repetitions - a disfluency more unique to stuttering. Notably, sound repetitions not only lead to transcription mistakes but also trigger Whisper to hallucinate over 20% of the time. Conducted by researchers who stutter, this study brings new insights on ASR biases against disfluent speech and highlights the value of disability-led research in addressing technological inequities affecting people with disabilities.

Index Terms: speech recognition, hallucination, stuttering, speech disfluency, algorithmic fairness

1. Introduction

Stuttering affects approximately 1% of world population [1]. While it is typically characterized by observable and involuntary “speech disfluencies”, many people who stutter (PWS) also experience significant reductions in their quality of life due to the communication challenges that they face everyday [2].

As Automated Speech Recognition (ASR) technologies 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 [3]. While some previous work has reported ASR’s performance disparity between stuttered and fluent speech [3, 4, 5], their findings remain limited in consistency, depth and a direct connection to the stuttering experience. Conducted by two researchers who stutter, this work systematically benchmark Whisper – OpenAI’s state-of-the-art ASR model with highly robust performance across languages and noisy environments – against a refined version of the SEP-28K dataset [6], 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 [7] and systemic exclusion of marginalized communities [8].

Previous work has shown existing ASR models’ poor performance with diverse speech patterns including stuttering [3, 4, 9], deaf speech [10], aphasia [11], second language speech [12], and regional and ethnic dialects [13, 14]. The inability of speech AI systems to work with diverse speech not only creates barriers for people with speech diversities to access mainstream products and services – such as voice assistants and automated phone menus, but can also lead to psychological harm [15] and reduced economic opportunities [16].

Recognizing these gaps, recent research has explored different 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 reduction in WER [17]. Another approach involves data augmentation with synthesized stuttered speech. Zhang *et al.* developed Stutter-TTS, a neural text-to-speech model capable of synthesizing diverse types of stuttering utterances [18]. 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 stuttered speech to expose ASR failures [19].

Building on prior work, this work provides a more systematic benchmarking of ASR performance on stuttered speech over different types of transcriptions and subtypes of stutters. We also analyze hallucinations in the resulting ASR outputs to understand their frequency and impact on user experience.

3. Methodology

3.1. Dataset

We leverage the SEP-28K dataset [6] for this study due to its scale and quality. With over 28,000 3-second audio clips labeled with five distinct stuttering subtypes, SEP-28K captures conversational stuttered speech in natural settings (i.e. podcasts), offering greater variability and heterogeneity within stuttering [20] than most existing stuttered speech datasets (e.g. LibriStutter [21], FluencyBank [22]). Its growing adoption by the research community also enables comparisons and validation of our results with related studies (e.g. [23, 24, 25, 26]).

3.2. Transcribing Audio Clips

Designed for stuttering event detection, SEP-28K does not include ground truth transcriptions for its audio clips. To ensure sufficient statistical power in benchmarking Whisper’s performance across different stuttering subtypes, we sample and manually transcribe over 400 audio clips for each stuttering subtype – block, prolongation, sound repetition, word repetition, and interjection – as well as 542 fluent clips for baseline comparison. Since SEP-28K uses multiple annotators for each clip, we prioritize using clips with unanimous agreement on the stuttering event subtype label to obtain the most representative audio clips for each stuttering subtype. For prolongations, where there are less than 400 clips with unanimous agreement in SEP-28K, we also include those with agreement between two annotators.

The first author, who is a person who stutters, listens to all the audio clips in our sample and manually transcribes them in two formats: **verbatim** and **semantic**. Verbatim transcriptions preserve stuttered utterances such as word repetitions (e.g. “when **when** are you guys getting”) and interjections (e.g. “I, **hmm**, am”), while semantic transcriptions omit the disfluencies (e.g. “when are you guys getting”). Having verbatim transcriptions is meaningful to PWS as it gives them agency over how their speech is represented in the transcript [5, 27]. It also enables PWS and speech language pathologists (SLPs) to analyze and understand stuttered speech patterns more accurately [28].

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

OpenAI’s Whisper is a Transform-based speech recognition model trained on 680,000 hours of labeled audio data collected from the web [29]. Approximately 117,000 hours of the audio is non-English while rest is primarily English. Using a convolutional encoder to convert log-mel spectrograms into embeddings and a decoder-only transformer to auto-regressively generate text, Whisper was a multi-task model trained to predict the next token in the transcription or other task-specific output sequence. Whisper large-v2 version was released on December 8, 2022, and large-V3 version was released on November 6, 2023.

We run speech-to-text transcription for each manually transcribed audio clip using OpenAI’s API for Whisper large-v2¹ during August, 2024 and October, 2024. The same clips are also transcribed using Whisper large-v3 through Hugging Face in February 2025. To benchmark Whisper’s performance, we compare its outputs against manually generated verbatim and semantic transcriptions. Evaluating against both types of transcriptions allows us to assess its accuracy for stuttered speech and its ability to preserve stuttering in transcriptions.²

3.4. Metrics

To evaluate Whisper’s transcription accuracy, we use Word Error Rate (WER) [30] to quantify syntax differences between Whisper output and manual ground truth, and BERT (Bidirectional Encoder Representations from Transformers) [31] F1 score to measure semantic differences. To examine how different stuttering subtypes affect Whisper’s performance, we also calculate average WER and BERT F1 scores separately for each stuttering type. Since WER accounts for word substitutions, deletions, and insertions relative to the reference transcription, we also analyze these individual error types to better understand Whisper’s behavior on stuttered speech.

As disfluent speech is reportedly more likely to trigger hallucination [11], we also analyzed Whisper’s hallucination frequency across different stuttering subtypes. To automate hallucination detection, we leverage the non-deterministic nature of hallucinations and follow a similar approach proposed by Koenecke *et al* [11], programmatically identifying hallucinations by comparing outputs from two separate runs of Whisper of the same audio clip using WER, BERT F1, and insertion rate. Specifically, for each audio clip, we treated the first run’s output as the reference and the second run’s output as the inference, calculating WER and BERT F1 between the two. We also count the number of words inserted by Whisper into the semantic ground truth in the first run. A transcription from the first run is automatically flagged as a potential hallucination if: (1) WER between the two runs is greater than 0.6; or (2) BERT F1 score between two runs is less than 0.6; or (3) number of words inserted into the semantic ground truth is greater than 4. The first author then manually examine all flagged transcriptions to correct false positives, and the second author review the final hallucination labels for quality control.

4. Results

4.1. Overall performance disparity

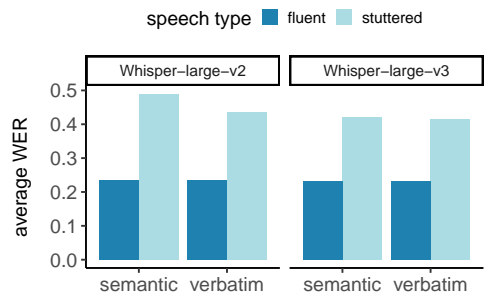
Consistent with findings on other ASR systems [3], our results show that Whisper consistently perform worse on stuttered speech than on fluent speech: producing more word-level mistakes and greater semantic divergence from the reference.

As illustrated in Figure 1a, while Whisper achieves relatively low WERs on fluent speech (0.234 for Whisper v2, 0.230 for v3)³, the error rates nearly double for stuttered speech: Whisper v2’s WERs are 0.489 (semantic) and 0.435 (verbatim); and Whisper v3’s WERs are 0.420 (semantic) and 0.414 (verbatim). A smaller yet persistent gap is also observed with BERT

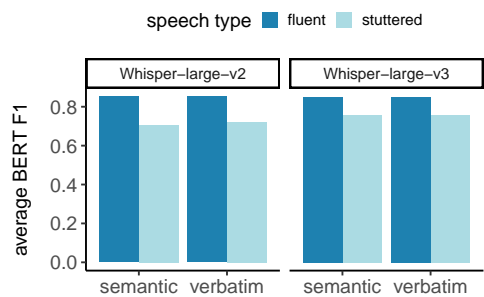
¹<https://platform.openai.com/docs/guides/speech-to-text>

²The code for this research is available at: <https://github.com/aimpowered/stuttered-speech-benchmark>

³Whisper’s performance on fluent clips is lower than previously reported [29], likely due to the short duration (3 seconds) of SEP-28K audio clips, which limits available context for the 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*

F1 scores (see Figure 1b).

When comparing semantic and verbatim transcriptions on stuttered clips, we find that Whisper are better at generating verbatim transcriptions than semantic ones (v2 semantic WER = 0.489, verbatim WER = 0.435), though this difference narrows in the newer version (v3 semantic WER = 0.420, verbatim WER = 0.414). Our review of Whisper’s outputs confirms Whisper v2’s ability to transcribe stuttering as they are – a capability that appears diminished in v3. For example, for a clip with verbatim transcription “so just” where the “j” sound is repeated multiple times, Whisper v2 is able to transcribed the repeated sound as “So, j-j-j-j-just”, while v3 simply transcribes it as “so just”.

Designed to measure semantic similarity, BERT F1 scores show minimal differences between semantic and verbatim transcription tasks (see Figure 1b). This is expected, as the two types of transcriptions of the same clip should differ only at the syntax level while preserving similar meaning – resulting in locational proximity in the BERT embedding space [31].

Overall, Whisper v3 has made progress in closing its performance gap between stuttered and fluent speech, improving both WER and BERT F1 scores. It is encouraging to see that advancements in Whisper also benefit people with speech diversity, contributing to greater equity in speech technology.

4.2. Performance disparity by stutter subtypes

Grouping stuttered clips by stutter subtypes, we find Whisper struggles most with **sound repetitions** (see Figure 2 *sound rep*) while performing 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 for stuttered clips, and more than double the WER for fluent clips.

Comparing to sound repetitions, Whisper handles interjec-

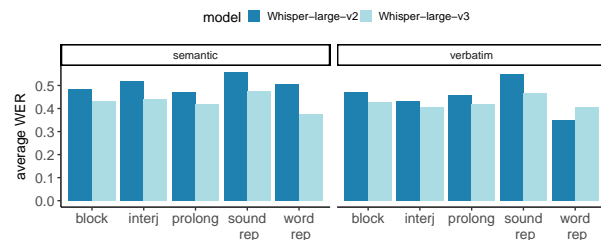


Figure 2: *Average WERs for different types of stuttered speech*

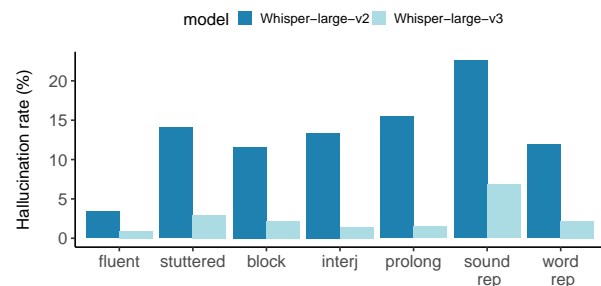


Figure 3: *Whisper hallucination frequency by speech type*

tions and word repetitions more effectively. As seen in Figure 2, Whisper achieves relatively low WERs for clips with interjections and word repetitions in verbatim transcriptions.

However, we also observe a regression in Whisper’s ability to transcribe repeated words in the newer version. As illustrated in Figure 2, Whisper v3 outperforms v2 across all stuttering subtypes and tasks except in transcribing clips with verbatim word repetitions. For example, for a clip with the verbatim transcription “which is which is terrible which is”, Whisper v2 outputs “it, which is, which is terrible, which is terrible.”, whereas v3 generates “which is terrible which is terrible”. Word Repetition is one of the most common forms of stuttering and a drop in performance in v3 can significantly impact stutterers.

4.3. Hallucination

Consistent with previous findings on aphasia speech [11], we find that Whisper is significantly more likely to hallucinate when transcribing stuttered speech than fluent speech. As shown in Figure 3, Whisper v2 hallucinates with 293 out of 2,086 (14%) stuttered clips, compared to just 18 out of 534 (3.3%) fluent clips. Whisper v3 has made significant progress in reducing hallucination rates, with only 2.9% (61) hallucinations for stuttered clips and 0.9% (5) for fluent clips.

Whisper v3 not only hallucinates less frequently than v2, but also hallucinate in qualitatively different ways. Our manual inspection of hallucinations reveals taht Whisper v2 tends to hallucinate with a set of typical phrases (e.g. “thank you”, “bye bye”), in a foreign language, or by adding large blocks of unrelated content. In contrast, Whisper v3 often adds only one or two words at the end of a sentence to complete it. Table 1 provides examples of typical hallucinations from both models.

In Figure 3, we can see Whisper’s hallucination frequency varies across different stutter subtypes. Consistent with the trends observed in Figure 2, clips with sound repetitions remain the most challenging: Whisper v2 hallucinates in 22.6% of clips with sound repetitions, while v3 reduces this rate to 6.9%.

7. References

- [1] O. Bloodstein, N. Ratner, and S. Brundage, *A Handbook on Stuttering, Seventh Edition*. Plural Publishing, 2021.
- [2] J. S. Yaruss and R. W. Quesal, “Overall assessment of the speaker’s experience of stuttering (oases): Documenting multiple outcomes in stuttering treatment,” *Journal of fluency disorders*, vol. 31, no. 2, pp. 90–115, 2006.
- [3] 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.
- [4] 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.
- [5] 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>
- [6] 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>
- [7] 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/>
- [8] 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.
- [9] D. Mujtaba, N. R. Mahapatra, M. Arney, J. S. Yaruss, C. Her-ring, 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.
- [10] 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.
- [11] 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.
- [12] 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>
- [13] 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.
- [14] 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>
- [15] 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. Association for Computing Machinery, 2023. [Online]. Available: <https://doi.org/10.1145/3544548.3581357>
- [16] 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>
- [17] 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>
- [18] 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.
- [19] 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>
- [20] 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.
- [21] 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>
- [22] 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: www.sciencedirect.com/science/article/pii/S0094730X17300931
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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. Association for Computing Machinery, 2024, p. 1926–1941. [Online]. Available: <https://doi.org/10.1145/3643834.3661533>
- [28] M. Zusage, L. Wagner, and B. Thallinger, “Crisperwhisper: Accurate timestamps on verbatim speech transcriptions,” in *Proc. Interspeech 2024*, 2024, pp. 1265–1269.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [30] 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 City University and Association for Computational Linguistics, 2014, pp. 1813–1823.
- [31] 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>