

Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech

Dena Mujtaba¹, Nihar R. Mahapatra¹, Megan Arney¹, J. Scott Yaruss¹,
Hope Gerlach-Houck², Caryn Herring³, and Jia Bin¹

¹Michigan State University, ²Western Michigan University,

³FRIENDS: The National Association of Young People Who Stutter

¹{mujtabad, nrm, arneymeg, jsy, binjia}@msu.edu

²hope.gerlach@wmich.edu

³caryn@friendswhostutter.org

Abstract

Automatic speech recognition (ASR) systems, increasingly prevalent in education, healthcare, employment, and mobile technology, face significant challenges in inclusivity, particularly for the 80 million-strong global community of people who stutter. These systems often fail to accurately interpret speech patterns deviating from typical fluency, leading to critical usability issues and misinterpretations. This study evaluates six leading ASRs, analyzing their performance on both a real-world dataset of speech samples from individuals who stutter and a synthetic dataset derived from the widely-used LibriSpeech benchmark. The synthetic dataset, uniquely designed to incorporate various stuttering events, enables an in-depth analysis of each ASR's handling of disfluent speech. Our comprehensive assessment includes metrics such as word error rate (WER), character error rate (CER), and semantic accuracy of the transcripts. The results reveal a consistent and statistically significant accuracy bias across all ASRs against disfluent speech, manifesting in significant syntactical and semantic inaccuracies in transcriptions. These findings highlight a critical gap in current ASR technologies, underscoring the need for effective bias mitigation strategies. Addressing this bias is imperative not only to improve the technology's usability for people who stutter but also to ensure their equitable and inclusive participation in the rapidly evolving digital landscape.

1 Introduction

1.1 Background and Motivation

Stuttering is a complex neurodevelopmental condition affecting over 80 million people worldwide (Wu, 2023; Yairi and Ambrose, 2013). This condition is characterized by various involuntary speech disruptions or disfluencies (Johnson, 1959; Ambrose and Yairi, 1999), including repetitions of parts of words (“li-li-like this”), prolongations of

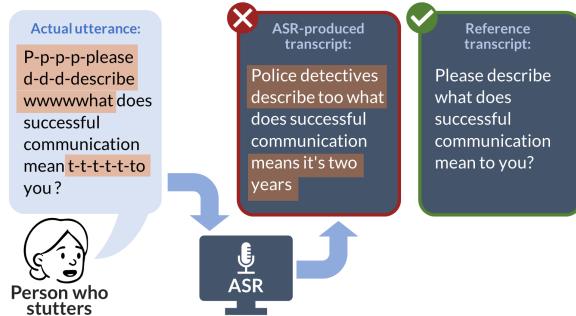


Figure 1: Transcript produced by an automatic speech recognition (ASR) model from a speech sample of an individual who stutters, sourced from FluencyBank (Ratner and MacWhinney, 2018). This figure illustrates a real-world scenario where the individual is conducting an interview. Highlighted are both the disfluencies characteristic of stuttered speech and the corresponding transcription inaccuracies produced by the ASR.

sounds (“lllllike this”), and blocks (silent periods, “l—ike this”). While everyone may experience disfluencies, in people who stutter, these are more pronounced and are accompanied by a loss of control over speech (Tichenor and Yaruss, 2019). Society’s negative perceptions of stuttered speech often leads to discrimination and lower job satisfaction, as fluent speakers tend to be preferred in employment settings (Gerlach et al., 2018; Gonzales, 2022; Plexico et al., 2019).

In the post-COVID-19 era, characterized by the widespread adoption of voice-activated artificial intelligence (*voice AI*), individuals who stutter are confronted with heightened challenges, especially in employment contexts. Automatic speech recognition (ASR) systems, the fundamental technology underpinning *voice AI* and designed to transcribe spoken language into text (Prabhavalkar et al., 2023), have emerged as notable barriers in various aspects of daily and professional life for this demographic. The broad integration of ASR technology across sectors such as education, healthcare, and employment, as well as in mobile devices and

other applications (Prabhavalkar et al., 2023; Malik et al., 2021), has led to the inadvertent marginalization of people who stutter. These systems frequently struggle to accurately decode speech that differs from typical patterns (Wheeler; Wu, 2023; Lea et al., 2023), a fact exemplified in Fig. 1. In stark contrast, ASRs demonstrate up to 95% word accuracy when processing standard speech (Tobin and Tomanek, 2022). The ramifications of this disparity are profound, ranging from difficulties in executing simple tasks, such as using automated phone systems for prescription refills, to experiencing bias in AI-driven job interviews (Junior et al., 2020; Center, 2019; Meyer, 2018).

As the global voice recognition market is projected to nearly quintuple by 2030, reaching a valuation of US\$59.62 billion (Insights, 2022), the prevalence of AI-powered voice technologies, bolstered by the recent success of large language models, is set to increase dramatically (Horvitz, 2016). This growth suggests an impending escalation in bias against individuals who stutter. Leading technology companies like Apple and Google have begun addressing variations in speech, introducing alternative solutions such as typing commands to voice assistants (Wheeler). However, these measures, while helpful, fall short of addressing the inherent bias in ASR accuracy. In particular, initiatives like Google’s Project Euphonia focus on collecting data for differing speech patterns, aiming to enhance ASR capabilities for understanding people with speech impairments like dysarthria (Tobin and Tomanek, 2022). However, these efforts often neglect the unique aspects of stuttered speech. Given the significant variability of stuttering across different individuals and situations (Tichenor and Yaruss, 2021), this oversight reveals a critical gap. Unless ASRs are specifically tailored and refined to accommodate the diverse expressions of stuttering, they are likely to continue underperforming for individuals who stutter, exacerbating existing societal biases.

1.2 Our Contributions

This paper examines the response of contemporary ASRs to disfluent speech characteristic of stuttering, with a focus on identifying and quantifying biases. While previous research has explored speech variations such as dysarthria (Shor et al., 2019; Moore et al., 2018) and ASR fairness in the context of gender, age, or accent differences (Feng et al., 2021; Liu et al., 2022), biases specific to disflu-

ent speech have remained largely unexplored. The work of Lea et al. (Lea et al., 2023) investigates interactions between people who stutter and AI-based voice assistants, yet it stops short of providing a quantitative analysis of biases.

Overall, the main contributions of this paper relative to previous work are threefold: (1) Quantitative analysis of bias: We conduct a comprehensive analysis of bias in both the word error rate (WER) and character error rate (CER) of six popular ASRs, utilizing real speech data from people who stutter. This approach offers a deeper understanding of how these systems perform with speech patterns that differ from the norm. (2) Semantic accuracy evaluation: We extend our analysis to the semantic accuracy of these ASRs, exploring how well they preserve the meaning of disfluent speech, which is crucial for effective communication. (3) Synthetic disfluent speech dataset: To address the scarcity of stuttered speech datasets, we have developed a synthetic disfluent speech dataset using novel text-to-speech and audio manipulation techniques. This synthetic dataset, designed to complement real stuttered speech samples, allows us to finely control the types and frequencies of disfluencies in speech samples. This controlled environment enables a granular examination of how various disfluencies and the inherent variability in stuttering affect ASR performance. Our findings aim to inform bias mitigation strategies in ASR development, ultimately aiding people who stutter in their daily interactions with voice-driven technologies.

The remainder of the paper is organized as follows. In Section 2, we provide a literature review of state-of-the-art ASRs, ASRs for people with speech variations, and define fairness/bias measurement in the context of AI. Section 3 details our methodology, datasets, models, and metrics for studying bias. Section 4 presents our findings and their analysis. Finally, Section 5 concludes with remarks and suggestions for future research directions.

2 Related Work

This section provides a brief literature review of three key areas to set the stage for subsequent sections examining bias in ASRs: the evolution and current state of ASR models, their adaptation for people with speech differences, and the measurement of bias in AI models.

2.1 Automatic Speech Recognition

Recent advancements in ASR have been largely driven by the integration of deep neural networks, leading to a significant reduction in WER compared to classical statistical multi-stage models (Prabhavalkar et al., 2023). These neural network models adopt an end-to-end (E2E) approach, directly mapping input acoustic features to their corresponding transcripts (Prabhavalkar et al., 2023). Key datasets like LibriSpeech (Panayotov et al., 2015), encompassing 1,000 hours of diverse audiobook recordings, and TIMIT (Garofolo, 1983), featuring recordings of 630 speakers across American English dialects, have been pivotal in training and benchmarking ASRs. Among the notable ASR architectures are the Conformer (Gulati et al., 2020), wav2vec 2.0 (Baevski et al., 2020), and Whisper (Radford et al., 2023), with Whisper demonstrating remarkable zero-shot performance and the best WER at the time of this writing (Gulati et al., 2020; Baevski et al., 2020; Radford et al., 2023). However, these models' effectiveness on stuttered speech remains underexplored. Comprehensive surveys on ASR systems are provided by Malik et al. (Malik et al., 2021) and on E2E ASRs by Prabhavalkar et al. (Prabhavalkar et al., 2023).

2.2 Automatic Speech Recognition for People with Speech Differences

Efforts to enhance ASR performance for people with different speech patterns, such as stuttering, have primarily focused on disfluency identification. This involves classifying disfluency events to enhance ASR accuracy, either by adjusting, ignoring, or removing these elements before they are processed by an ASR, or by providing their location to the ASR (Zayats et al., 2016; Shonibare et al., 2022). Notable datasets supporting stuttering identification include SEP-28K, KSoF, and LibriStutter, although they lack properly coded transcripts for ASR training and evaluation (Sheikh et al., 2022; Lea et al., 2021; Kourkounakis et al., 2020). While there is limited research on directly enhancing ASRs for stuttering, some progress has been made. Lea et al. fine-tuned Apple's Speech framework models with data from people who stutter, observing a reduction in WER. They also applied disfluency refinement and endpoint truncation techniques to improve accuracy (Lea et al., 2023). Mitra et al.'s work on tuning decoding parameters of ASRs also contributes to this effort (Mitra et al.,

2021). Additionally, other research, such as those by Tobin et al., Shor et al., and MacDonald et al., has focused on adapting ASRs for various speech differences (Tobin and Tomanek, 2022; Shor et al., 2019; MacDonald et al., 2021; Kim et al., 2008; Moore et al., 2020).

2.3 Fairness in Artificial Intelligence

Fairness in AI involves mitigating human and systemic biases that may arise from factors like unrepresentative training data, inappropriate features used during training, or biased target labels (Mujtaba and Mahapatra, 2019). The concept of fairness in AI has evolved, with historical foundations in anti-discrimination legislation and contemporary frameworks like the AI Risk Management Framework by NIST aiming to enhance AI trustworthiness (Commision; NIST). A fundamental fairness criterion in AI literature is “fairness through unawareness” (Kusner et al., 2017):

Definition 1. Fairness Through Unawareness: An AI model C is deemed fair under the principle of ‘Fairness Through Unawareness’ if it does not explicitly use protected attributes A in its decision-making process. This principle is operationalized through a mapping $\hat{Y}: X \rightarrow Y$, where X represents the set of observable features excluding any protected attributes A , \hat{Y} represents a model’s prediction, and Y the intended outcome.

In this paper, A represents the presence of disfluencies, C is the ASR model under evaluation, X is the user’s speech input, and Y is the transcript produced by an ASR. Our study investigates the correlation between the presence of disfluencies (A) and model error rates to evaluate ASR fairness. This can be captured through the notion of “counterfactual fairness” (Kusner et al., 2017; Wachter et al., 2017):

Definition 2. Counterfactual Fairness: An AI model C is deemed counterfactually fair if its predictions \hat{Y} in the real world are the same as those in the counterfactual world where an individuals’ protected attributes A belonged to a different group.

3 Methodology

To investigate accuracy biases in contemporary ASR systems against disfluent speech, we have designed two experiments, each focusing on a different speech dataset. The first dataset is well-known in the speech-language pathology (SLP) community, and the second is widely used in ASR mod-

eling and evaluation. Our objective is to compare ASR performance on two versions of these datasets: one with disfluencies and another identical in content but without disfluencies.

Experiment 1 utilizes the FluencyBank dataset ([Ratner and MacWhinney, 2018](#)), known for its collection of real stuttered speech samples with accompanying transcripts (detailed in Sec. 3.2). We refer to this original dataset as FluencyBank-Y (FB-Y), where “Y” indicates the presence of disfluencies. These samples are transcribed using six different ASR models (Sec. 3.1), and the accuracy of these transcriptions is measured using WER, CER, and BERTScore, a metric for semantic similarity analysis (Sec. 3.4). For comparative analysis, we generate a version of this dataset with apparent instances of stuttering or disfluencies removed, named FluencyBank-N (FB-N), where “N” signifies the absence of disfluencies, using text-to-speech (TTS) techniques (Sec. 3.3). We then employ the Spearman correlation coefficient test to systematically compare the performance of the ASRs on both the original FB-Y and the modified FB-N datasets.

To complement Experiment 1, Experiment 2 utilizes the LibriSpeech dataset ([Panayotov et al., 2015](#)), frequently used in training and evaluating ASRs and characterized by standard, non-disfluent speech samples. We refer to this dataset as LibriSpeech-N or LS-N to convey the absence of disfluencies. To assess the impact of disfluencies, we create a synthetic dataset by introducing controlled disfluency events into the LibriSpeech dataset. This is achieved through a combination of audio processing and TTS methods. This newly modified dataset, denoted LibriSpeech-Y or LS-Y, where “Y” indicates the presence of disfluencies, is then subjected to the same testing process as the FluencyBank dataset in Experiment 1, using the same set of six ASR models. This approach ensures a systematic analysis of bias within these ASR systems across diverse speaking scenarios.

3.1 ASR Models Studied

We investigate a mix of commercially-available market ASRs and research-oriented, open-source models, to uncover biases across diverse ASR systems. The market ASRs include: (a) *Google Cloud’s Speech-to-Text V1* ([Cloud](#)), (b) *IBM Watson’s Speech-to-Text* ([IBM](#)), (c) *Microsoft Azure* ([Azure](#)), and (d) *RevAI* ([Rev.AI](#)). It is important to note that the specific details regarding the training

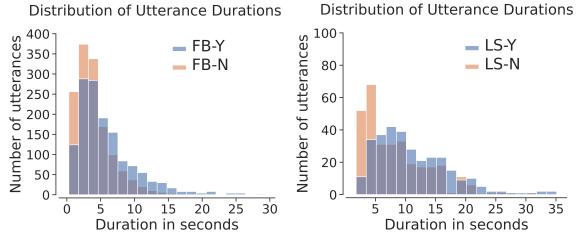


Figure 2: Histogram showing the frequency distribution of utterance durations in seconds for each of the four speech datasets.

processes of these commercial models are proprietary and not publicly disclosed.

In addition to these market ASRs, our investigation extends to two open-source, research-based ASR models, which are:

- *wav2vec 2.0* ([Baevski et al., 2020](#)): This is an end-to-end ASR model, utilizing connectionist temporal classification loss for training. For our analysis, we employ the ‘wav2vec2-large-960h’ configuration, available from Meta on the HuggingFace Library ([Facebook](#)). This specific model was trained on a substantial corpus of 960 hours of the LibriSpeech dataset.
- *Whisper* ([Radford et al., 2023](#)): Representing the current state-of-the-art in ASR technology, Whisper is a transformer-based encoder-decoder model. It has been pretrained on over 680k hours of multilingual data, leveraging semi-supervised learning and multitasking approaches—including tasks like language identification and voice activity detection. The primary aim of Whisper is to achieve generalization across a variety of benchmarks. For our purposes, we utilize the ‘whisper-1’ model configuration provided by OpenAI API, which is the most up-to-date version at the time of this writing.

3.2 FluencyBank: Stuttered Speech Dataset

Our study utilizes the FluencyBank dataset ([Ratner and MacWhinney, 2018](#)) to examine biases in real stuttered speech. FluencyBank comprises a series of video recordings featuring individuals who stutter, captured while reading passages and during interviews. Each video in this dataset is accompanied by a transcript, which has been segmented into individual utterances—a sentence or a fragment of a sentence (containing whole phrases) spoken by a single individual.

Due to temporal misalignment and incorrect transcript errors, which were identified by us and others (Sheikh et al., 2022; Lea et al., 2021), we had the dataset relabeled by trained SLPs with expertise in stuttering. They used the standard Codes for the Human Analysis of Transcripts format (MacWhinney, 2021), and added start and end timestamps for each utterance within a given audio file, while ensuring inter-annotator agreement. The dataset utilized in our analysis consists of 1,373 utterances, derived from 7 reading and 12 interview videos. The distribution of these utterances in terms of duration is shown in Fig. 2, with the cumulative audio duration amounting to approximately 2.21 hours.

In addition to the original dataset (FB-Y), we produced a version with apparent instances of stuttering or disfluencies removed, referred to as FluencyBank-N or FB-N. This variant was created using TTS, which synthesized speech from reference transcripts devoid of disfluencies. This approach, detailed in Sec. 3.3, serves as a crucial component in our comparative analysis of bias.

3.3 Synthetic Disfluent Speech Generation

To complement our analysis using the FluencyBank dataset, we produce a synthetic dataset (LibriSpeech-Y or LS-Y) derived from LibriSpeech, an ASR dataset featuring public domain audiobook recordings. We utilized the ‘clean’ sample from LibriSpeech’s test set, available through the HuggingFace library, to ensure that the wav2vec 2.0 model was not previously exposed to this data during training. However, for commercial models (i.e., those available from Google and Microsoft), we do not have specific training data details and hence can not exclude the possibility that LibriSpeech has been used in their training, thus potentially boosting their performance in our analysis.

Our approach improves upon the LibriStutter dataset by FluentNet (Kourkounakis et al., 2020), which lacks diversity in disfluency events. LibriStutter’s limitations include inadequate representation of the variability of stuttering, such as inserting only one to two word repetitions. Moreover, LibriStutter only uses audio processing methods for disfluency insertion. We introduce a more diverse range of disfluency events (Einarsdóttir and Ingáham, 2005): *interword blocks*, *intraword blocks*, *prolongations*, *word repetitions*, *phrase repetitions*, and *interjections*. While these do not cover all types of disfluencies associated with stuttering,

they serve to demonstrate key variations for our bias study.

We introduce disfluencies into the speech samples using two methods: *textual disfluency insertion* and *audio disfluency insertion*. For textual disfluency insertion, we modify the transcript and then use a TTS model, specifically SpeechT5 TTS (Ao et al., 2022, 2021), to generate speech that naturally includes the disfluency events. This process, supported by a set of pre-trained speaker vectors (Hollemans), allows us to simulate a diverse range of voices. This technique is primarily used for word repetitions, phrase repetitions, and interjections. The frequency and location of these disfluencies within each sentence are selected randomly within predetermined ranges: (1) for word repetition, we select between 1 to 3 words in each utterance to repeat 1 to 4 additional times; (2) for phrase repetition, we choose a phrase of 2 to 4 words and repeat it 1 to 3 times; (3) for interjections, we randomly insert either “uh” or “um” in 1 to 4 locations and repeat them 1 to 4 times.

For audio disfluency insertion, we directly modify the speech waveform to add disfluencies that TTS cannot create, such as blocks and prolongations. We employ a pre-trained wav2vec 2.0 model and PyTorch CTC forced alignment (Zhang and Hira) for precise alignment of words with their timestamps. This method allows for the insertion of interword blocks, intraword blocks, and additional word and phrase repetitions with parameters similar to those used in textual disfluency insertion. The parameters for disfluency events applied in audio processing are also randomly selected: (1) for interword blocks, we insert a pause after 1 to 4 words with a duration of 1 to 3 seconds; (2) for intraword blocks, we insert a pause into 1 to 3 words, each lasting between 1 to 3 seconds; (3) for prolongations, we extend the sound of 1 to 3 words at their beginning, stretching it to 5 to 25 times its original length. Additionally, word and phrase repetitions are applied following the same parameters as in the textual disfluency insertion process.

Each disfluency is randomly applied, without replacement, to an equal number of samples, amounting to 10% of the LibriSpeech dataset, ensuring a balanced representation. The resulting dataset comprises approximately one hour of speech samples, with 308 records in total. The duration of these samples, in seconds, is depicted in Figure 2.

This approach allows for a comprehensive and nuanced analysis of ASR performance across a

Table 1: Examples of disfluency events including word repetition (WR), phrase repetition (PR), prolongation (Prolong.), interword block (InterWB), intraword block (IntraWB), and interjection (Interject.). Words in $\langle \rangle$ brackets are added through textual disfluency insertion, while (pause) signifies the audio insertion of silence.

Disfluency Event	Example
WR	<i>How <are> are you?</i>
PR	<i>How <are you> are you?</i>
Interject.	<i>How are <um> you?</i>
InterWB	<i>How are (pause) you?</i>
IntraWB	<i>How a(pause)re you?</i>
Prolong.	<i>How aaare you?</i>

range of disfluency types, thereby enriching our understanding of bias in these systems.

3.4 ASR Evaluation Metrics

To effectively assess biases in ASR systems, we employ three key metrics, each offering a unique lens on transcription accuracy and semantic fidelity.

The first metric is the *word error rate (WER)*, a widely recognized standard in ASR evaluations, defined as: $WER = \frac{S+D+I}{S+D+C}$, where S , D , I , and C represent the number of word substitutions, word deletions, word insertions, and correct words, respectively, in the ASR-produced transcript relative to the reference transcript. WER provides a direct measure of transcription accuracy at the word level. A lower WER, approaching zero, is desirable, indicating higher word transcription accuracy.

In conjunction with WER, we analyze the *character error rate (CER)*, which evaluates transcription errors at the character level. Similar to WER but more granular, CER offers insight into the accuracy of finer transcription details. As with WER, a lower CER is indicative of better performance.

Our third metric, *BERTScore* (Zhang et al., 2019), examines semantic similarity between predicted and target sentences. This metric is crucial for cases where a transcription error might involve just a single word, yet lead to a significant change in meaning. For instance, incorrectly transcribing "not" in a sentence could completely reverse its intended meaning, despite being a single-word error. BERTScore assesses this aspect, with a score closer to 1 indicating greater semantic similarity between the predicted and target sentences.

With BERTScore, we can observe the meaning behind each word. Given an ASR-produced sentence $x = \langle x_1, \dots, x_k \rangle$ and a reference sentence $\hat{x} = \langle \hat{x}_1, \dots, \hat{x}_l \rangle$, where x_i and \hat{x}_i repre-

sent the i^{th} token embeddings derived directly from a pre-trained BERT model's contextual embeddings, BERTScore computes the cosine similarity of each token in x with every token in \hat{x} to obtain an F1 score. We utilize embeddings from the 'microsoft/deberta-large-mnli' model (He et al., 2021) alongside the re-scaled version of BERTScore, wherein the range is within $[-1, 1]$, where a 1 would be semantically identical transcripts, and -1 opposite. We refer to this as F_{BERT} (Tobin et al., 2022; Zhang et al., 2019) and it is defined as the following:

$$F_{BERT} = 2 \frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}}, \text{ where } \quad (1)$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} (x_i^T \cdot \hat{x}_j), \text{ and } \quad (2)$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} (x_j^T \cdot \hat{x}_i). \quad (3)$$

We use the Spearman correlation coefficient to analyze the relationship between ASR performance metrics across disfluent and non-disfluent speech datasets. We opt for this non-parametric test because the performance distributions are non-normal (as detailed in Appendix B.1), necessitating an analysis approach that does not assume normality. The Spearman correlation coefficient, defined as:

$$r_{WER} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (4)$$

facilitates our understanding of variations in ASR model performance across the two datasets. Here, d_i is the difference in ranks between each observation's WER in disfluent and non-disfluent speech datasets, and n represents the total number of observations or utterances. This approach helps us quantify the impact of disfluent speech patterns, like stuttering, on ASR accuracy. A value closer to 1 or -1 indicates a strong positive or negative correlation, respectively, between the presence of disfluencies and the difference in WER, suggesting a significant impact of disfluencies on ASR performance. Conversely, a correlation closer to 0 implies a weaker relationship. We perform similar computations for CER (r_{CER}) and BERTScore (r_{BERT}), offering a multidimensional view of each ASR system's bias and accuracy.

Additionally, we incorporate the calculation of the *p-value* to determine the statistical significance of the observed biases. The *p-value* quantifies the

Table 2: Comparative analysis of ASR model performance metrics (WER, CER, and F_{BERT}) on FluencyBank datasets: ‘Y’ denotes the presence of stuttering (FB-Y), and ‘N’ denotes its absence (FB-N). The mean of scores across ASRs is given by μ , and the standard deviation by σ .

	WER		CER		F_{BERT}	
	Y	N	Y	N	Y	N
Azure	.168	.073	.117	.052	.805	.896
GCP	.275	.069	.182	.050	.623	.920
IBM	.476	.167	.407	.150	.463	.838
RevAI	.263	.159	.187	.133	.586	.508
Whisp	.121	.063	.082	.047	.870	.935
W2V2	.380	.101	.212	.061	.549	.865
μ	.281	.105	.198	.082	.649	.827
σ	.120	.042	.103	.042	.143	.146

probability of observing the results under the null hypothesis, which in this context is the absence of bias. A p -value lower than a typical threshold (e.g., 0.05) suggests that the observed bias is statistically significant and not due to random chance. This metric is essential for substantiating the reliability of our findings.

4 Results and Discussion

Our analysis, divided into two parts corresponding to each experiment in Sec. 3, reveals significant insights into ASR performance with real and synthetic disfluent speech. The evaluation was conducted using ASR services as of March. 14, 2024, utilizing default API settings. For wav2vec 2.0, the analysis was performed locally using an NVIDIA T4 GPU. Transcripts returned blank by models due to audio processing challenges (e.g., due to noise or blocks) were retained in our dataset. Prior to evaluation, all transcripts underwent normalization using Whisper’s “BasicTextNormalizer” and “EnglishNumberNormalizer” functions from the Transformers library ([HuggingFace](#)), ensuring consistent casing and punctuation removal. Any speech atmospherics in output transcripts, such as laughing or indications of silence, were also removed.

4.1 Experiment 1: Real Stuttered Speech

The outcomes of ASR model performance on the FluencyBank dataset, with and without disfluencies (denoted as FB-Y and FB-N, respectively), are presented in Table 2. Our findings indicate a consistent pattern across all models: lower WER, CER,

Table 3: Spearman correlation coefficients (r), all of which had $p < 0.05$, indicating the statistical significance of the impact of disfluencies on Δ WER, Δ CER, and ΔF_{BERT} . FB refers to comparisons between FB-Y and FB-N datasets, and LS to comparisons between synthetic (LS-Y) and original (LS-N) LibriSpeech datasets. “All” refers to the combination of all ASR predictions.

	Δ WER		Δ CER		ΔF_{BERT}	
	FB	LS	FB	LS	FB	LS
Azure	.363	.656	.385	.700	-.33	-.62
GCP	.552	.547	.559	.609	-.35	-.47
IBM	.531	.616	.539	.690	-.52	-.49
RevAI	.245	.557	.246	.611	-.21	-.50
Whisp	.285	.501	.298	.528	-.24	-.48
W2V2	.563	.712	.562	.781	-.54	-.67
All	.413	.566	.423	.623	-.38	-.50

and higher F_{BERT} on the FB-N dataset, implying a systemic bias against disfluent speech. Among the evaluated models, Whisper exhibited the lowest WER and highest F_{BERT} . However, its performance on FB-Y was markedly worse (about two times) compared to that on FB-N, where it achieved a WER of just 6.3%. We also note that the RevAI model showed a higher F_{BERT} for FB-Y than for FB-N, primarily due to blank transcript predictions. Upon removing these from the analysis, the reported F_{BERT} on FB-Y is 0.809 and on FB-N is 0.933. This disparity in performance highlights a significant bias in ASRs against disfluent speech.

Moreover, Fig. 3 illustrates the WER distribution for the wav2vec 2.0 model across all datasets. The mean WER is noticeably higher for datasets with disfluencies, highlighting a broader range and more outliers with elevated WER.

4.2 Experiment 2: Synthetic Speech

In our second experiment, we evaluated the ASR models against our synthetic disfluency dataset compared to the original LibriSpeech. The results (Table 4) mirror those from Experiment 1, with higher WER and CER for the synthetic disfluency dataset. This finding confirms the bias observed in Experiment 1, not just for real disfluent speech but also for synthesized versions. The Whisper model again showed superior performance, but with a significant drop (WER of 3.1% vs. 19.8%) when disfluencies were introduced.

Further analysis on disfluency-event-type specific WER (Table 5) reveals that certain disfluencies, such as word repetitions, prolongations, and

Table 4: Evaluation of ASR models on synthetic stuttering and original LibriSpeech datasets. Metrics include WER, CER, and F_{BERT} , with ‘Y’ indicating synthetic stuttering (LS-Y) and ‘N’ indicating the original dataset (LS-N). The mean of scores across ASRs is given by μ , and the standard deviation by σ .

	WER		CER		F_{BERT}	
	Y	N	Y	N	Y	N
Azure	.228	.037	.197	.018	.772	.957
GCP	.268	.070	.217	.034	.744	.904
IBM	.560	.191	.518	.127	.448	.757
RevAI	.254	.062	.209	.031	.742	.901
Whisp	.198	.031	.157	.015	.811	.969
W2V2	.259	.033	.219	.009	.718	.952
μ	.295	.071	.253	.039	.706	.907
σ	.121	.056	.120	.040	.119	.072

interjections, significantly increase WER. In contrast, intraword block and interword block events showed a less pronounced impact. This suggests that different disfluency types vary in their effect on ASR performance, underscoring the need for more inclusive and diverse training data in ASRs.

The Spearman correlation coefficients and corresponding p -values (Table 3) across all models and for different disfluencies (Appendix B.2) underscore this finding. We find all correlation values in Table 3 are statistically significant, with $p < 0.05$, indicating that the presence of disfluencies substantially impacts ASR accuracy. For example, prolongations and interjections showed a high correlation with increased WER, highlighting specific areas where ASRs struggle in handling disfluent speech. We also note the correlation coefficients for F_{BERT} present an inverse correlation, where ASRs consistently had lower F_{BERT} scores on FB-Y and LS-Y, indicating poor semantic similarity.

Overall, our results indicate a clear bias in current ASR technology against disfluent speech, both real and synthetic. This bias is evident across various disfluency event types and is statistically significant, emphasizing the need for targeted improvements in ASRs to accommodate speech variations.

5 Conclusion

This study represents a significant step in identifying and quantifying the biases in ASR systems against individuals who stutter. By analyzing the performance of six popular ASR models using two distinct datasets—FluencyBank and a custom-

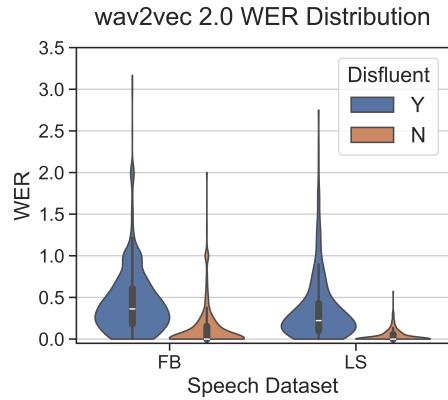


Figure 3: Violin plot illustrating the range and distribution of WER for the wav2vec 2.0 model on FluencyBank (FB) and synthetic (LS) datasets. The plot highlights mean values and variability in WER scores.

generated synthetic dataset incorporating disfluency events from LibriSpeech—we have uncovered substantial biases. Our findings reveal that all six models demonstrate notably worse WER, CER, and BERTScore when processing disfluent speech. This bias is consistent across both real and synthetic datasets and is statistically significant.

Our research also delves into disfluency-event-specific biases, revealing that the majority of disfluencies—six out of eight types—exhibit a statistically significant WER bias across all models. This discovery is crucial in understanding the nuances of ASR performance and the challenges faced by people who stutter in real-world scenarios.

To address these challenges, there are a few approaches that can be considered. First, incorporating speech from people who stutter in ASR training; this is challenging due to the limited availability of relevant data. Additionally challenging is the task of acquiring sufficient data that encompasses samples from a diverse population to capture the variability in speech associated with stuttering. However, with a strategic data collection approach, this can address a primary limitation in ASR accessibility. Secondly, obtaining feedback from people who stutter and benchmarking ASRs before deployment are crucial steps towards accessibility. Rigorous testing on disfluent speech can identify ASR deficiencies, allowing for targeted training or post-processing to address them.

Looking forward, we also identify two promising directions for future research. Firstly, there is a need to explore the impact of longer-form audio on ASR biases. Current models often struggle with correctly segmenting and transcribing longer audio samples containing disfluencies, leading to signifi-

Table 5: WER results for each ASR on the synthetic dataset, broken down by specific disfluency event types. These include audio-based word repetition (Aud-WR), audio-based phrase repetition (Aud-PR), prolongation (Prolong.), interword block (InterWB), intraword block (IntraWB), interjection (Interject.), text-TTS-based word repetition (Txt-WR), and text-TTS-based phrase repetition (Txt-PR). ‘Y’ denotes the presence of each disfluency event type, and ‘N’ indicates its absence. The mean of scores across ASRs is given by μ , and the standard deviation by σ .

	Aud-WR		Aud-PR		Prolong.		InterWB		IntraWB		Interject.		Txt-WR		Txt-PR	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Azure	.27	.02	.22	.03	.35	.04	.12	.05	.08	.03	.32	.05	.32	.05	.26	.03
GCP	.33	.05	.24	.06	.53	.08	.11	.08	.08	.05	.38	.07	.36	.09	.28	.10
IBM	.67	.42	.35	.22	.65	.20	.72	.21	.68	.11	.52	.15	.42	.20	.49	.12
RevAI	.34	.05	.23	.04	.44	.08	.10	.08	.07	.05	.36	.08	.36	.09	.28	.04
Whisp	.23	.03	.12	.03	.32	.03	.07	.04	.04	.02	.41	.04	.33	.05	.23	.02
W2V2	.33	.02	.21	.03	.55	.03	.11	.03	.09	.03	.34	.04	.32	.04	.26	.05
μ	.36	.10	.23	.07	.47	.08	.20	.08	.17	.05	.39	.07	.35	.09	.30	.06
σ	.14	.14	.07	.07	.12	.06	.23	.06	.23	.03	.07	.04	.04	.05	.08	.04

cant transcription inaccuracies or even cessation of live transcription. This aspect is particularly critical as it can severely limit the practical usability of ASR technologies for individuals who stutter.

Secondly, incorporating human feedback from people who stutter represents an important area of exploration. Current metrics like WER and semantic accuracy might not fully capture the transcription preferences of individuals who stutter. While our study assumes that the transcription of FluencyBank, with stuttering indicators removed, accurately represents the intended speech, this may not align with the preferences of those who stutter. They might desire their speech patterns, such as interjections, prolongations, and other disfluencies, to be reflected in the transcripts. Hence, engaging directly with individuals who stutter to establish ground-truth labels could offer a more authentic and respectful approach to ASR development.

In conclusion, our research is among the first to systematically examine and highlight the biases in ASR systems towards disfluent speech. The implications of these biases are far-reaching, potentially exacerbating disparities in various domains, from healthcare to recruitment, where ASR technologies are increasingly integrated. To pave the way for more inclusive and equitable ASR technologies, bias mitigation strategies, including diversifying training data with disfluent speech and incorporating feedback from the stuttering community, are essential. Our work lays the foundation for this crucial endeavor, opening avenues for future research and development in the field of speech recognition.

Limitations

We acknowledge three limitations of our work. First, our study focuses solely on accuracy bias by ASR models transcribing speech in English, and people who stutter speaking English. This does not cover the many multilingual ASR models and people who stutter speaking other languages. Second, our synthetic dataset does not encompass all types of disfluencies, and not all disfluencies (e.g., interjections) can be considered stuttering in all circumstances since they depend upon the presence of a loss of control by the speaker while speaking (Tichenor and Yaruss, 2019). Our intended use for this dataset is for studying bias at a fine-grained disfluency level, with control over the frequency and location of occurrence of disfluencies. Therefore, this dataset should not be used for ASR model training as a bias mitigation measure, and should also not be used to study bias unaccompanied by other datasets. Likewise, third, our study method should not be used as the sole assessment of bias in ASRs. We define bias by observing the transcript accuracy (i.e., WER, CER, and semantic accuracy) difference with and without disfluencies. However, other metrics for bias and other facets, such as age or gender, can affect ASR bias. Overall, we seek to develop more inclusive ASRs, and have these limitations addressed in future work.

Acknowledgment

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. 2235916 and 2345086.

References

- Nicoline Grinager Ambrose and Ehud Yairi. 1999. Normative disfluency data for early childhood stuttering. *Journal of Speech, Language, and Hearing Research*, 42(4):895–909.
- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.
- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.
- Microsoft Azure. Azure Speech to Text. <https://azure.microsoft.com/en-us/products/ai-services/speech-to-text>.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Electronic Privacy Information Center. 2019. EPIC - EPIC Files Complaint with FTC about Employment Screening Firm HireVue.
- Google Cloud. Google Cloud Text-to-Speech AI. <https://cloud.google.com/text-to-speech>.
- U.S. Equal Employment Opportunity Commission. Federal law prohibiting job discrimination questions and answers. <https://www.eeoc.gov/fact-sheet/federal-laws-prohibiting-job-discrimination-questions-and-answers>.
- Jóhanna Einarsdóttir and Roger J Ingham. 2005. Have disfluency-type measures contributed to the understanding and treatment of developmental stuttering?
- Facebook. wav2vec2-base-960h. <https://huggingface.co/facebook/wav2vec2-base-960h>.
- Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.
- John S Garofolo. 1983. TIMIT acoustic-phonetic continuous speech corpus.
- Hope Gerlach, Evan Totty, Anu Subramanian, and Patricia Zebrowski. 2018. Stuttering and labor market outcomes in the united states. *Journal of Speech, Language, and Hearing Research*, 61(7):1649–1663.
- Matt Gonzales. 2022. Stuttering discrimination and the workplace. <https://www.shrm.org/resourcesandtools/hr-topics/behavioral-competencies/global-and-cultural-effectiveness/pages/stuttering-discrimination-and-the-workplace.aspx>.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.
- Mathijs Hollemans. Mattijs/cmu-arctic-xvectors. <https://huggingface.co/datasets/Mattijs/cmu-arctic-xvectors>.
- Eric Horvitz. 2016. One hundred year study on artificial intelligence.
- HuggingFace. Normalization and pre-tokenization. <https://huggingface.co/learn/nlp-course/chapter6/4>.
- IBM. IBM Watson Speech to Text. <https://www.ibm.com/products/speech-to-text>.
- Fortune Business Insights. 2022. The global speech and voice recognition market size. <https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382>.
- Wendell Johnson. 1959. *The onset of stuttering: Research findings and implications*. U of Minnesota Press.
- Julio Junior, CS Jacques, Agata Lapedriza, Cristina Palmero, Xavier Baró, and Sergio Escalera. 2020. Person perception biases exposed: Revisiting the first impressions dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 13–21.
- Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*.
- Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. FluentNet: End-to-end detection of speech disfluency with deep learning. *arXiv preprint arXiv:2009.11394*.
- Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

- Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P Bigham. 2021. SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6798–6802. IEEE.
- Chunxi Liu, Michael Picheny, Leda Sari, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166. IEEE.
- Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, et al. 2021. Disordered speech data collection: lessons learned at 1 million utterances from project euphonia.
- Brian MacWhinney. 2021. Tools for analyzing talk, part 1: The CHAT transcription format (2019). *DOI*, 10:66.
- Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.
- David Meyer. 2018. Amazon reportedly killed an AI recruitment system because it couldn't stop the tool from discriminating against women. *Fortune. Tillgänglig online: https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/* (2019-09-27).
- Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Sarah Wu, Darren Botten, Ashwini Palekar, Shrinath Thelapurath, Panayiotis Georgiou, Sachin Kajarekar, et al. 2021. Analysis and tuning of a voice assistant system for dysfluent speech. *arXiv preprint arXiv:2106.11759*.
- Meredith Moore, Piyush Papreja, Michael Saxon, Visar Berisha, and Sethuraman Panchanathan. 2020. Uncommonvoice: A crowdsourced dataset of dysphonic speech. In *Interspeech*, pages 2532–2536.
- Meredith Moore, Hemanth Venkateswara, and Sethuraman Panchanathan. 2018. Whistle-blowing asrs: Evaluating the need for more inclusive speech recognition systems. *Interspeech 2018*.
- Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in AI-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE.
- NIST. 2023. AI risk management framework. <https://www.nist.gov/itl/ai-risk-management-framework>.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.
- Laura W Plexico, Megan-Brette Hamilton, Haley Hawkins, and Stephen Erath. 2019. The influence of workplace discrimination and vigilance on job satisfaction with people who stutter. *Journal of Fluency Disorders*, 62:105725.
- Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *arXiv preprint arXiv:2303.03329*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Nan Bernstein Ratner and Brian MacWhinney. 2018. Fluency Bank: A new resource for fluency research and practice. *Journal of fluency disorders*, 56:69–80.
- Rev.AI. The world's most accurate API for AI- and human-generated transcripts. <https://www.rev.ai/>.
- Shakeel A Sheikh, Md Sahidullah, Fabrice Hirsch, and Slim Ouni. 2022. Machine learning for stuttering identification: Review, challenges and future directions. *Neurocomputing*.
- Olabanji Shonibare, Xiaosu Tong, and Venkatesh Ravichandran. 2022. Enhancing ASR for stuttered speech with limited data using detect and pass. *arXiv preprint arXiv:2202.05396*.
- Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, et al. 2019. Personalizing ASR for dysarthric and accented speech with limited data. *arXiv preprint arXiv:1907.13511*.
- Seth E Tichenor and J Scott Yaruss. 2019. Stuttering as defined by adults who stutter. *Journal of Speech, Language, and Hearing Research*, 62(12):4356–4369.
- Seth E Tichenor and J Scott Yaruss. 2021. Variability of stuttering: Behavior and impact. *American Journal of Speech-Language Pathology*, 30(1):75–88.

Jimmy Tobin, Qisheng Li, Subhashini Venugopalan, Katie Seaver, Richard Cave, and Katrin Tomanek. 2022. Assessing ASR model quality on disordered speech using BERTScore. *arXiv preprint arXiv:2209.10591*.

Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641. IEEE.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841.

Kevin Wheeler. For people who stutter, the convenience of voice assistant technology remains out of reach. <https://techxplore.com/news/2020-01-people-stutter-convenience-voice-technology.html>.

Shaomei Wu. 2023. “The World is Designed for Fluent People”: Benefits and challenges of videoconferencing technologies for people who stutter. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Ehud Yairi and Nicoline Ambrose. 2013. Epidemiology of stuttering: 21st century advances. *Journal of fluency disorders*, 38(2):66–87.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional LSTM. *arXiv preprint arXiv:1604.03209*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Xiaohui Zhang and Moto Hira. CTC forced alignment API tutorial. https://pytorch.org/audio/main/tutorials/ctc_forced_alignment_api_tutorial.html.

A Evaluation of Synthetic Disfluent Speech

To ensure that our synthetic disfluent speech dataset (LS-Y) presented in Section 3.3 is comparable to real stuttered speech, we assess its similarity to the FluencyBank dataset (FB-Y). It is important to note that our dataset’s objective is not to mimic stuttering precisely; instead, it aims to provide a comparable framework for evaluating ASR biases toward stuttered speech.

Table 6: Comparative analysis of similarity scores between our synthetic dataset LS-Y with disfluencies, and both FluencyBank datasets where ‘Y’ denotes the presence of disfluencies (FB-Y), and ‘N’ denotes its absence (FB-N). This is broken down by specific disfluency event types. These include audio-based word repetition (Aud-WR), audio-based phrase repetition (Aud-PR), prolongation (Prolong.), interword block (InterWB), intraword block (IntraWB), interjection (Interject.), text-TTS-based word repetition (Txt-WR), and text-TTS-based phrase repetition (Txt-PR). The mean of similarity scores is given by μ , and the standard deviation by σ . ‘All’ is the average distance across all disfluency types. We observe across all disfluency types, our synthetic speech dataset is closer to resembling stuttered speech than non-stuttered speech.

	FB-Y		FB-N	
	μ	σ	μ	σ
Aud-WR	.659	.178	.614	.190
Aud-PR	.589	.192	.524	.192
Prolong.	.634	.178	.590	.190
InterWB	.637	.187	.569	.192
IntraWB	.546	.190	.472	.176
Interject.	.667	.174	.635	.189
Txt-WR	.648	.184	.597	.197
Txt-PR	.598	.193	.531	.194
All	.623	.189	.567	.197

To compare, our evaluation procedure starts by computing the cosine similarity between all spectrograms of audio samples in LS-Y and FB-Y. Following this, we calculate the cosine similarity between all spectrograms of audio samples in LS-Y and FB-N, our corresponding FluencyBank dataset without disfluencies. This approach aims to determine whether our dataset exhibits greater similarity to FB-Y, which contains stuttered speech, than to FB-N, which lacks stuttering. Such a comparison helps gauge the resemblance of our dataset to stuttered speech versus non-stuttered speech.

The effectiveness of this approach and our dataset is demonstrated through the results presented in Table 6, which display the average and standard deviation of similarities on a disfluency-specific level between LS-Y and both FB-Y and FB-N datasets. We observe that our synthetic speech dataset exhibits a greater similarity to real stuttered speech (FB-Y) than to non-stuttered speech (FB-N) across all disfluency types. This finding is further supported by the fact that our FB-N dataset was generated using voices through TTS (Sec. 3.3)

Table 7: Spearman correlation coefficient (r_{WER} denoted by r below) and corresponding p -value results for each ASR model on the synthetic dataset, broken down by specific disfluency event types. These include audio-based word repetition (Aud-WR), audio-based phrase repetition (Aud-PR), prolongation (Prolong.), interword block (InterWB), intraword block (IntraWB), interjection (Interject.), text-TTS-based word repetition (Txt-WR), and text-TTS-based phrase repetition (Txt-PR). ‘Y’ denotes the presence of each disfluency event type, and ‘N’ indicates its absence. “All” refers to the combination of all ASR predictions.

	Aud-WR		Aud-PR		Prolong.		InterWB		IntraWB		Interject.		Txt-WR		Txt-PR	
	r	p	r	p	r	p	r	p	r	p	r	p	r	p	r	p
Azure	.78	.00	.83	.00	.40	.00	.56	.00	.58	.00	.70	.00	.76	.00	.79	.00
GCP	.72	.00	.81	.00	.45	.00	.18	.11	.27	.01	.71	.00	.69	.00	.62	.00
IBM	.54	.00	.51	.00	.48	.00	.76	.00	.84	.00	.61	.00	.48	.00	.66	.00
RevAI	.79	.00	.80	.00	.46	.00	.11	.33	.29	.01	.65	.00	.68	.00	.74	.00
Whisp	.64	.00	.40	.00	.36	.00	.17	.13	.10	.40	.77	.00	.74	.00	.78	.00
W2V2	.82	.00	.83	.00	.71	.00	.48	.00	.66	.00	.78	.00	.78	.00	.72	.00
All	.67	.00	.67	.00	.46	.00	.33	.00	.40	.00	.69	.00	.67	.00	.70	.00

overlapping with our LS-Y dataset, possibly increasing the similarity between LS-Y and FB-N. Therefore, overall, our synthetic stuttered speech generation method proves to be an effective means of introducing disfluencies in non-stuttered speech.

B Additional Experiment Results

B.1 ASR performance distributions

The distribution of WER, CER, and F_{BERT} scores for all ASR models on real and stuttered speech datasets are presented in Figure 4, Figure 5, and Figure 6, respectively. We observe that the mean is lower for each ASR when disfluencies are present, specifically in LS-Y and FB-Y, in comparison to LS-N and FB-N without perceived stuttering.

B.2 Correlation Coefficient Test Results

We test the correlation between degraded ASR accuracy and the presence of disfluent speech using the Spearman rank correlation coefficient, as presented in Table 3. Similarly, we also compute the individual correlation coefficients and associated p -values for each disfluency type in our synthetic dataset (LS-Y). These results are presented in Table 7. We find all correlation coefficients, excluding those for interword blocks, statistically significant, with $p < 0.05$. In the case of interword blocks, we find a weak correlation, $r_{WER} < 0.20$, in the GCP, RevAI, and Whisper models without statistical significance, whereas other ASR models have a strong correlation with statistical significance. This is due to the difference in WER being quite small when an interword block is present, as observed in

Table 1, with only a 2% change for RevAI, and 3% change for GCP and Whisper. This is also seen for the Whisper ASR on the intraword block, where a low correlation $r_{WER} = 0.10$ is observed, as well as only a small decrease, of 2%, in WER. In comparison, other disfluency types such as word repetitions, interjections, and phrase repetitions show a high correlation with their presence in speech and a decrease in WER. When combining predictions across ASR models, we also observe a statistically significant correlation between all disfluency types, with phrase repetitions having the strongest, and interword blocks having the weakest.

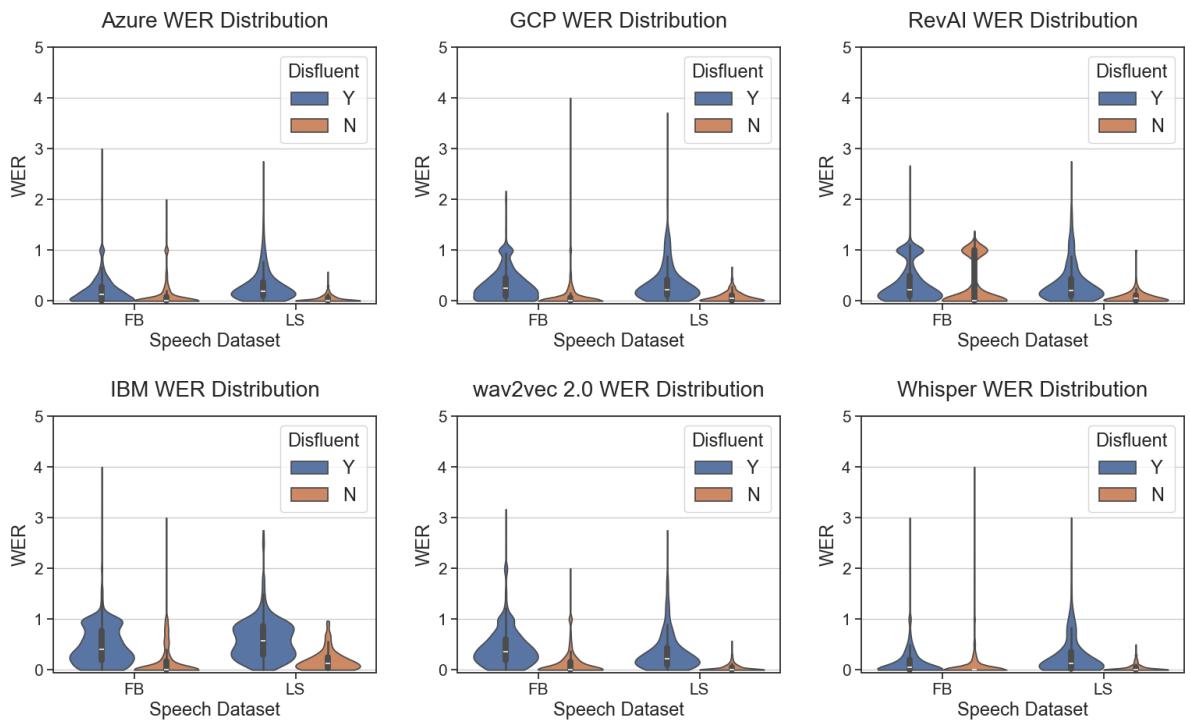


Figure 4: Violin plots illustrating the range and distribution of WER for each ASR model on FluencyBank (FB) and LibriSpeech (LS), both with disfluencies (Y) and without disfluencies (N).

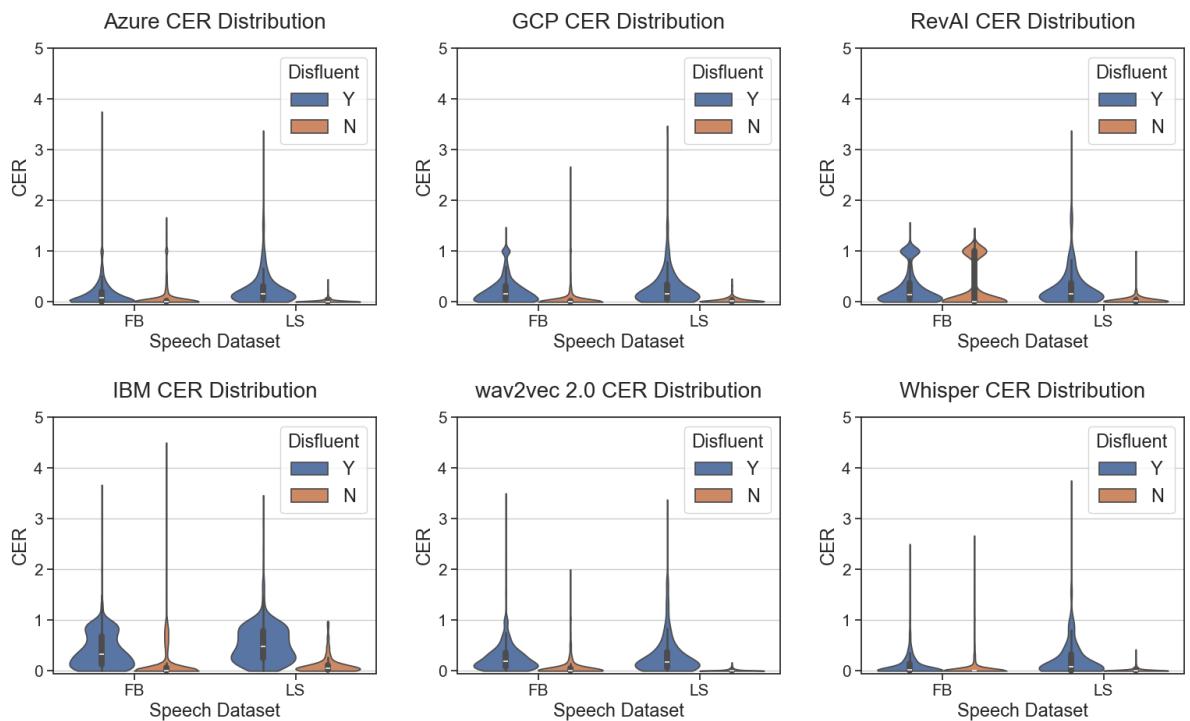


Figure 5: Violin plots illustrating the range and distribution of CER for each ASR model on FluencyBank (FB) and LibriSpeech (LS), both with disfluencies (Y) and without disfluencies (N).

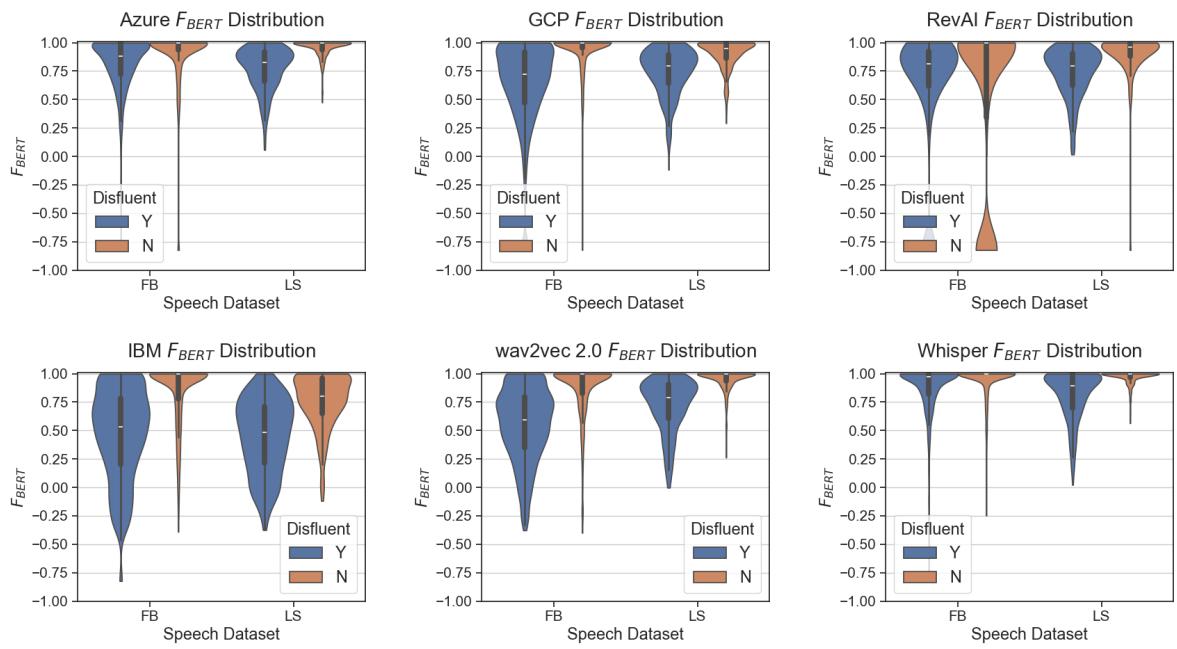


Figure 6: Violin plots illustrating the range and distribution of F_{BERT} for each ASR model on FluencyBank (FB) and LibriSpeech (LS), both with disfluencies (Y) and without disfluencies (N).