Evaluating Open-Source ASR Systems: Performance Across Diverse Audio Conditions and Error Correction Methods

Saki Imai*
Northeastern University
imai.s@northeastern.edu

Tahiya Chowdhury
Colby College
tchowdhu@colby.edu

Amanda Stent
Colby College
amanda.stent@gmail.com

Abstract

Despite significant advances in automatic speech recognition (ASR) accuracy, challenges remain. Naturally occurring conversation often involves multiple overlapping speakers, of different ages, accents, and genders, as well as noisy environments and suboptimal audio recording equipment, all of which reduce ASR accuracy. In this study, we evaluate the accuracy of state of the art open source ASR systems across diverse conversational speech datasets, examining the impact of audio and speaker characteristics on WER. We then explore the potential of ASR ensembling and post-ASR correction methods to improve transcription accuracy. Our findings emphasize the need for robust error correction techniques and for continuing to address demographic biases to enhance ASR performance and inclusivity.

1 Introduction

Automatic Speech Recognition (ASR) technology has witnessed significant advancements in recent years, primarily due to the introduction of powerful transformer models trained on large datasets. These advancements have brought ASR systems close to human accuracy for conversational telephone speech (Stolcke and Droppo, 2017). However, ASR systems still face challenges in transcribing spontaneous human conversations (Szymański et al., 2020). In particular, naturally occurring conversation tends to involve multiple overlapping speakers, of different ages, accents, and genders, as well as noisy environments and suboptimal audio collection equipment. Transcription accuracy of state-of-the-art ASR systems varies significantly with audio characteristics such as accent, gender, overlapping speech, and background noise (Tatman and Kasten, 2017; Goldwater et al., 2008; Liesenfeld et al., 2023; Chen et al., 2022). Gaining a deeper understanding of how these factors influence ASR performance is essential for improving both accuracy and inclusivity.

To address these challenges, various ASR error correction methods have been explored. Traditional approaches like ROVER reflect the classic ASR pipeline comprising acoustic and language models (Fiscus, 1997). With the shift towards end-to-end (E2E) ASR systems, new transcription error correction methods have emerged (Chan et al., 2016). These methods often leverage large-scale pre-trained language models like BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) to enhance correction capabilities, especially in contexts with limited labeled speech data. However, previous studies have not extensively investigated error correction across ensembles of ASR systems.

In this study, we: (1) evaluate the accuracy of state-of-the-art open-source ASR systems across diverse conversational speech conditions, exploring the impact of speaker age, gender, accentedness, and background noise level; (2) examine the potential of ASR ensembling plus post-ASR correction methods to increase transcription accuracy; and (3) we release a dataset comprising the results from six different ASR systems across six datasets with varying audio characteristics, supporting reproducible research and encouraging further exploration of ASR ensembling strategies.[1]

2 Related Work

2.1 Disparities in ASR Performance

Previous research has highlighted disparities in ASR performance based on speaker attributes. Studies have demonstrated that gender, age, and accent can significantly influence WER. For instance, certain studies found female speech to be more accurately recognized than male speech, suggesting gender disparities in ASR performance (Adda-

*Work done while at Colby College

[1] https://github.com/sakimai/asr-evaluations

Proceedings of the 31st International Conference on Computational Linguistics, pages 5027–5039
January 19–24, 2025.

Decker and Lamel, 2005; Goldwater et al., 2008; Feng et al., 2024). Conversely, (Garnerin et al., 2019; Tatman, 2017) observed a lower WER for male speech in specific contexts. Some studies reported no significant differences based on gender, highlighting the variability of ASR performance across different systems and datasets (Tatman and Kasten, 2017). Age-related differences in WER have also been documented, with findings indicating that child speech poses more challenges for ASR systems compared to adult speech due to physiological and behavioral differences (Jain et al., 2023). Accent presents another significant challenge, with systems often mis-recognizing the speech of speakers with accents divergent from the "standard" variant of the language used in training datasets (Tatman and Kasten, 2017; Koenecke et al., 2020; Dorn, 2019; DiChristofano et al., 2023; Tadimeti et al., 2022; del Río et al., 2023).

While the impact of individual audio characteristics on ASR performance has been extensively studied, less is known about how open-source ASRs with different architectures, such as Hidden Markov Models (HMM), self-supervised learning (SSL) models, speech foundation models, and E2E models perform across diverse datasets featuring various audio characteristics. In this study, we aim to fill this gap by comparing the transcription capabilities of six different open-source ASR systems across six conversational speech datasets.

2.2 Transcription Error Correction

Error correction models focus on detecting and correcting errors within ASR hypotheses. Traditional methods like ROVER use multiple hypothesis alignment and voting mechanisms to generate the best hypothesis (Fiscus, 1997). Other early error correction methods included statistical correction and statistical machine translation systems, and ontology learning (Cucu et al., 2013; D'Haro and Banchs, 2016; Anantaram et al., 2018). More recently, non-autoregressive correction models have shown promise (Leng et al., 2021b,a, 2023). Following the trend of E2E ASR systems, recent methods combine these components in an E2E manner (Guo et al., 2023; Hrinchuk et al., 2020; Guo et al., 2019). Pre-trained language models such as T5, BERT, and BART have also shown promise in correcting ASR transcription errors (Ma et al., 2023a; Li et al., 2024, 2021; Zhao et al., 2021; Ma et al., 2023b). Finally, generative large language models (LLMs) like GPT-3.5 and LLaMA have also been explored for this purpose (Ma et al., 2023c; Chen et al., 2023). Moreover, using multiple hypotheses—such as an n-best list—from a single ASR system can enhance the contextual framework for error correction models, allowing for improved correction capabilities (Zhu et al., 2021; Leng et al., 2021a). In this study, we compare different approaches to ASR error correction while combining hypotheses from multiple ASR systems.

3 Data & Systems

Dataset selection. We chose datasets to ensure a comprehensive and accurate evaluation of conversational speech. Priority was given to datasets that are not commonly used in the training of ASR models. The datasets include child speech, different levels of noise, varying number of speakers, and multiple accents of English (Table 1).

On dataset complexity, we recognize that the multidimensional nature of complexity is difficult to capture in a 2-D table. For instance, child speech is challenging not only due to its pitch and acoustic features but also due to variability in linguistic and pronunciation characteristics (Lee et al., 1997, 1999). Two-speaker audio introduces multiple conversational dynamics and conditions, such as overlapped speech and acoustic variability related to changes in speaker characteristics like gender. In comparison, single-speaker accented speech lacks these dyadic conversational elements, making the former more complex (Takaaki et al., 2013). Our dataset selection reflects these nuances, ordered in decreasing order of complexity from top to bottom:

- *Sawyer* (Sawyer, 2013) – unstructured play sessions involving 20 children.
- *Cameron* (Starke, 2023) – structured interactions between an adult and a child in a controlled, quiet environment.
- *CHiME-5* (Barker et al., 2018) – multi-speaker conversations in noisy environments, such as dinner table recordings.
- *AMI* (Mccowan et al., 2005) – meetings involving multiple speakers.
- *Covid conversations* (Romero and Paxton, 2023) – two-speaker interactions from experimental psychology sessions.
- *SPGISpeech* (O'Neill et al., 2021) – earnings calls.

ASR system selection. We chose systems based on their performance, architecture diversity, and

Dataset	Data Split	Noisy audio?	# of speakers	Accented speech?	Child speech?	Total Length (Sample Length)
Sawyer	*All*		20		✓	24 h (23 min)
Cameron	*0wav*		2		✓	28 h (20 min)
CHiME-5	*dev*	✓	4	✓		4.5 h (22 min)
AMI	*IHM-test*	✓	4			100 h (16 min)
Covid Conversations	*All*		2			1.5 h (58 min)
SPGISpeech	*test*		1	✓		5,000 h (61 min)

Table 1: Characteristics of each dataset

representation of both traditional and modern approaches to speech recognition:

- *Kaldi (Librispeech)* (Povey et al., 2011) – A HMM system trained on 960 hours of Librispeech data.
- *Wav2Vec2 (Large-960h)* (Baevski et al., 2020) – A transformer-based SSL model pre-trained on unlabeled data and fine-tuned on 960 hours of labeled Librispeech data.
- *HuBERT (large-ls960-ft)* (Hsu et al., 2021) – Another transformer-based SSL model, also fine-tuned on 960 hours of Librispeech data.
- *Whisper (medium.en)* (Radford et al., 2023) – A transformer-based speech foundation model, multilingual, specifically fine-tuned for English language tasks.
- *Distil-Whisper (medium.en)* (Gandhi et al., 2023) [*DWhisper*] – A distilled version of the Whisper medium model.
- *NVIDIA (Conformer-Transducer X-Large)* (Gulati et al., 2020) – A conformer-based E2E model, trained on several thousand hours of English speech including Librispeech.

4 Metrics

4.1 Word Error Rate

Word error rate (WER) is a common metric used to evaluate the performance of ASRs. It quantifies the percentage of errors in the transcription generated by the ASR system compared to a reference transcription.

4.2 Stop Words Filtered WER (swf-WER)

Previous research has identified limitations in the utility of WER for accurately reflecting speech understanding (Wang et al., 2003). To address this, we also use Stop Words Filtered WER, *swf-WER*, which calculates WER excluding stop words (Garofolo et al., 1998). We used the part of speech tags provided by the spaCy NLP library to identify stop words in transcripts (Honnibal et al., 2020).

5 Methods

Figure 1 shows a visual presentation of our approach, which we describe below.

5.1 Dataset preprocessing

5.1.1 Generating turn-level audio files

ASR over long audio samples is more error-prone than ASR over short audio samples, so we constructed turn-level samples for our experiments. The SPGI and AMI datasets each contain a single utterance per audio file. However, for the other datasets, each recording contains an entire conversation. For the CHiME-5, Sawyer, and Cameron datasets, which include turn-level timestamps, we used ffmpeg (Tomar, 2006) to extract individual turns. The Psychology dataset came with non-timestamped turn-level transcripts, so we used the Montreal Forced Aligner (McAuliffe et al., 2017) to create time-aligned transcripts, from which we extracted turn-level timestamps.

5.1.2 Removing paralinguistic transcriptions

Transcriptions of the CHiME-5, Cameron, and Sawyer datasets included paralinguistic annotations such as "laugh" or "xxx" for unintelligible utterances. To focus on the linguistic content, we excluded these annotations from our analysis. However, the audio was retained as it provided more context information, and exposed the systems to real-world conditions and various speaker conditions (Shriberg, 2005; Schuller and Batliner, 2013). The Cameron dataset, in particular, offers both phonetic and semantic transcriptions. We chose to use the semantic transcriptions to align with modern ASR systems that use language models to generate hypotheses. For instance, phonetic transcriptions like "ah a gobritha [: gorilla]" were converted

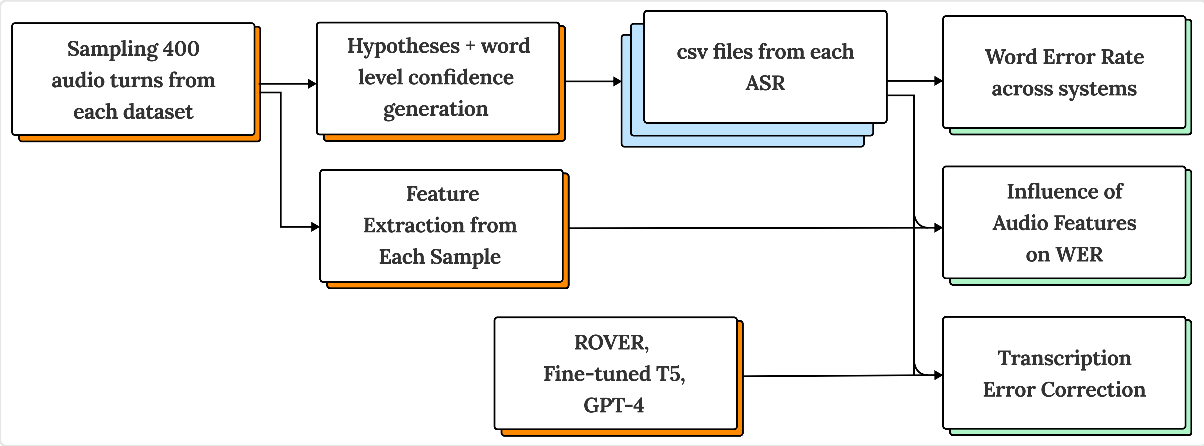

Figure 1: Method diagram

to their semantic equivalents, "ah a gorilla." It is important to note that this method might not be advantageous for ASR systems like Kaldi, which predominantly rely on acoustic models to produce transcriptions.

5.2 Random Sampling of Dataset

We randomly sampled 400 audio turns from each dataset. All audio files were then resampled to a mono channel at 16 kHz to meet the sample rate requirements of the ASR systems. The total audio duration for these 400 turns from each dataset is summarized in Table 1.

5.3 Feature Extraction from Each Sample

With limited speaker metadata available for our datasets, we used machine learning based models to estimate gender, age, and accent (Zuluaga-Gomez et al., 2023; Burkhardt et al., 2023; Ferreira, 2023). We measured noise levels using Power Spectral Density (PSD) estimates via Welch's method from scipy, and computed both audio duration and syllables per second (PyPI, 2024).

6 Analysis of ASR Performance

In this section, we analyze the effects of various speaker and audio characteristics on WER. In subsection 6.1, we look at the overall accuracy of each recognizer. In subsection 6.2, we analyze differences in WER due to continuous features (duration, speaking rate, noise level, age). In subsection 6.3, we analyze differences in WER due to categorical features (accentedness and speaker gender).

6.1 Comparative WER across different ASR systems

A comparative analysis of WER across different ASR systems is shown in Figure 2. We plot WER on a logarithmic scale due to its wide range and order datasets from the most to the least complex from left to right (see Table 1 for the features we consider as contributing to dataset complexity), with the rightmost column representing the results for all six datasets combined.

Performance Trends. As expected, we observe a decrease in WER as we move from the more complex datasets to the simpler ones. Speech Foundation models, such as DWhisper and Whisper, generally exhibit lower WER, suggesting a higher robustness to diverse audio characteristics. The E2E NVIDIA ASR has a slightly higher WER on simpler datasets, but a lower WER on the most complex dataset.

Variability in Performance. The error bars, representing standard deviations, indicate variability in WER for each system across the datasets. Systems exhibit higher variability (larger error bars) for more complex datasets like Sawyer and CHiME-5, implying fluctuating performance in challenging conditions. Whisper, in particular, shows extended error bars in these datasets, which is indicative of issues including hallucination.

ASR Recommendation. The NVIDIA ASR consistently achieves the lowest mean WER across all datasets, followed by DWhisper. For highly complex and noisy conditions, the E2E NVIDIA ASR offers reliable performance with the least variability. In contrast, DWhisper may be more suitable for environments with less complex settings.

In Table 2, we analyze swf-WER:

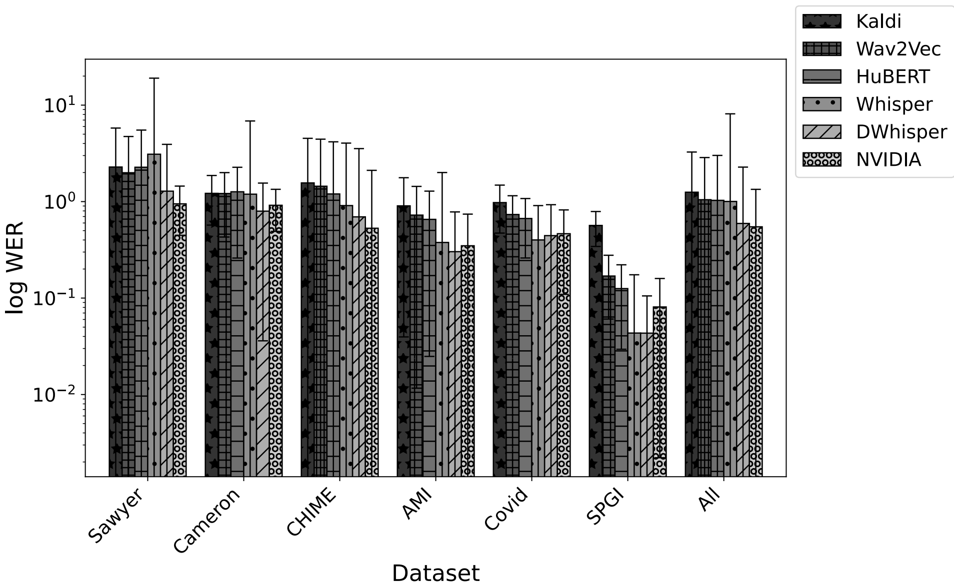

Figure 2: Mean and SD of WER by System and Dataset

Dataset	Mean		Standard Deviation		Minimum		Maximum	
	WER	swf-WER	WER	swf-WER	WER	swf-WER	WER	swf-WER
Sawyer	1.98	1.83	1.50	1.36	0.65	0.64	4.44	4.06
Cameron	1.10	1.15	0.56	0.60	0.59	0.61	1.94	2.07
CHiME-5	1.06	0.96	0.64	0.57	0.39	0.39	1.96	1.77
AMI	0.55	0.57	0.39	0.36	0.16	0.21	1.08	1.05
Covid	0.61	0.66	0.30	0.31	0.31	0.35	1.05	1.10
SPGI	0.17	0.21	0.21	0.24	0.02	0.05	0.57	0.66

Table 2: Comparison of WER by dataset

Lower swf-WER for Complex Datasets. In complex environments represented by Sawyer, Cameron, and CHiME-5, swf-WER is consistently lower than WER. This suggests that errors involving stop words, which are typically less acoustically distinct, are more prevalent in noisy settings, thus disproportionately affecting the perceived accuracy of ASR systems. Removing these words from the error calculation refocuses the metric on content that significantly impacts comprehension.

Consistent Performance for Simpler Datasets. Datasets like AMI, Covid, and SPGI, which exhibit lower and more stable WERs, reflect consistent ASR performance. For these acoustically simpler datasets, swf-WER is slightly higher than WER, possibly indicating greater linguistic complexity such as a greater variety of named entities in these datasets.

6.2 Impact of Continuous Features on WER

We analyzed the impact of continuous audio features including speaking rate (syllables per second), speaker age, noise level, and audio duration by fitting regression models. After visualizing the data, we used cubic regression functions to model these relationships. However, the results for noise level and duration were not statistically significant, likely due to data clustering within a narrow value range and limited variability. Despite attempts to rescale noise level using normalization, log transformation, and standardization, the regression remained statistically insignificant. This is likely due to data clustering within a narrow value range, a similar issue was observed with audio duration, where limited variability was present due to the uniform sampling of audio lengths. Therefore, this section focuses on the impact of speaking rate and speaker age on WER. Full regression equations, R^2 values, and p-values are provided in the appendix (Section A.1).

6.2.1 Effect of Speaking Rate on WER

Table 3 shows the local minimizer, R^2 value, and Bonferroni corrected p-value for the cubic coefficient for each ASR system. The local minimizer represents the number of syllables per second that achieves the local minimum WER according to the

System	Local Minimizer	R^2	Corrected p-value
Kaldi	6.851	0.128	<0.01
Wav2Vec	6.832	0.176	<0.01
HuBERT	6.668	0.178	<0.01
Whisper	6.810	0.013	0.085
DWhisper	6.276	0.077	<0.01
NVIDIA	6.430	0.103	<0.01

Table 3: Impact of syllables per second on WER

System	Local Minimizer	R^2	Corrected p-value
Kaldi	24.208	0.031	0.011
Wav2Vec	22.810	0.052	<0.01
HuBERT	28.148	0.070	<0.01
Whisper	32.274	0.012	1.000
DWhisper	31.898	0.036	0.035
NVIDIA	24.428	0.122	<0.01

Table 4: Effect of age on WER across systems

System	Mean Difference	Corrected p-value
Kaldi	-0.632	<0.01
Wav2Vec	-0.620	<0.01
HuBERT	-0.707	<0.01
Whisper	-0.334	1.00
DWhisper	-0.240	1.00
NVIDIA	-0.065	1.00

Table 5: Effect of speaker overlap on WER

System	Mean Difference	Corrected p-value
Kaldi	0.636	<0.01
Wav2Vec	0.691	<0.01
HuBERT	0.725	<0.01
Whisper	1.040	0.04
DWhisper	0.417	<0.01
NVIDIA	0.291	<0.01

Table 6: Gender-related differences in WER

regression equation. We observe that most systems achieve the minimum WER when the speaking rate is between 6.4 and 6.8 syllables per second, indicating optimal ASR performance around this range.

6.2.2 Effect of Speaker Age on WER

Speaker age significantly impacts ASR system performance (Table 4). The local minimizer represents the age at which the ASR system achieves the lowest WER according to the regression equation. Most systems achieve optimal performance with speakers in their 20s, aligning with previous studies indicating lower WER for young to middle-aged adults (Werner et al., 2019). However, we did not obtain statistically significant results for DWhisper, Whisper, or Kaldi, suggesting that these systems are less sensitive to speaker age.

6.3 Impact of Discrete Features on WER

We analyzed the impact of discrete audio features using the student's t-test. To control for any false discovery rate arising from multiple tests, we applied Holm's sequential Bonferroni procedure (Benjamini and Hochberg, 1995).

6.3.1 Speaker Overlap and WER Variability

The data in Table 5 highlights the impact of speaker overlap on ASR accuracy. The recent transformer-based ASR systems, such as Whisper and DWhisper, as well as the conformer-based NVIDIA system, did not show a significant mean difference under speaker overlap, indicating their robustness in challenging conditions. In contrast, Kaldi, Wav2Vec, and HuBERT exhibited significant mean differences, suggesting lower performance when dealing with overlapping speech.

6.3.2 Gender-Related Differences in WER

The analysis of gender-related differences in WER demonstrates a bias in ASR system performance against male speakers (coded as 1), with all ASR systems showing higher WER for male voices compared to female voices (coded as 0). There are positive mean differences across the board in Table 6, which are statistically significant, except for the Whisper model, suggesting that Whisper is more robust to both male and female speakers.

6.3.3 Accent-Related WER Discrepancies

One would expect non-general American accents (coded as 1) to be associated with higher WER compared to the general American accent (coded as 0). However, our results did not show this trend. None of the ASRs showed statistical significance in the mean WER between non-general American and general American accents. This lack of significance could be due to the subtle nature of accent differences or to the machine learning methods used to classify these accents.

System	Mean Difference	Corrected p-value
Kaldi	0.0165	1.00
Wav2Vec	-0.088	1.00
HuBERT	-0.075	1.00
Whisper	0.279	1.00
DWhisper	0.012	1.00
NVIDIA	-0.014	1.00

Table 7: Effect of accent on WER

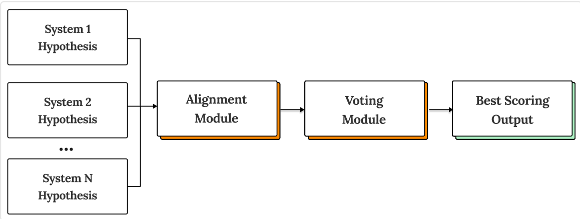

Figure 3: ROVER

7 Transcription Error Correction

In this section, we evaluate three ASR error correction methods described in the literature (see Section 2). For each method, we provide the one-best hypothesis from each of the ASR systems described in Section 3.

7.1 Recognizer Output Voting Error Reduction (ROVER)

ROVER combines multiple ASR outputs into a single word transition network (Fiscus, 1997). This network is then processed through a voting mechanism that identifies the most reliable output sequence based on the lowest aggregate score (Figure 3). The "voting" or rescoring process reconciles differences in ASR system outputs.

7.2 Fine Tuned T5 (Ft-T5)

We finetuned the T5 transformer model to perform ASR error correction over transcripts from an ensemble of ASR models, inspired by the approach described in (Ma et al., 2023a). As described in Section 5, we processed 400 audio turns from each of the 6 datasets through each of the 6 ASR systems, leading to 6 ASR transcripts for a total of 2400 audio files. We split these ASR transcripts into ten folds. In round-robin fashion, we then selected each fold for testing, finetuning T5 on the rest. T5 was finetuned using concatenated ASR transcripts and their corresponding word-level confidence scores, with the accurate transcription serving as the target output, as shown in Figure 4.

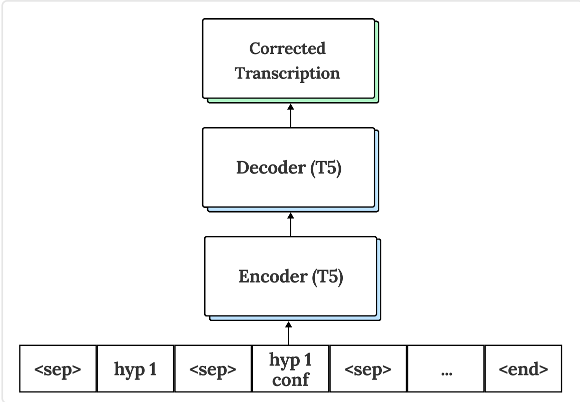

Figure 4: T5 fine-tuning

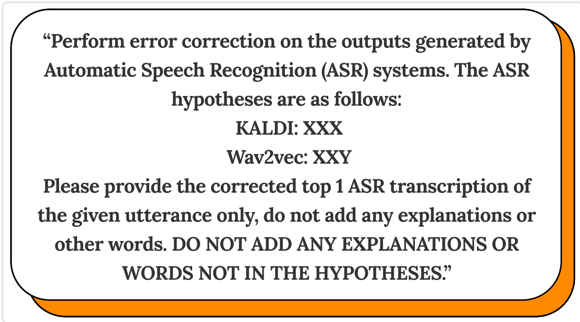

Figure 5: Zero-shot error correction using GPT-4

7.3 GPT-4

Following the work by (Ma et al., 2023c), we used GPT-4 zero-shot to perform ASR error correction over transcripts from an ensemble of ASR models. Following limited experimentation, we used the prompt shown in Figure 5. The capitalization in the prompt is important for best results.

7.4 Error Correction Results

The results of the three transcription error correction methods are shown in Tables 8 and 9. We present the results in two separate tables because subsets of the datasets were used during the fine-tuning process for the T5 model, as described in Section 7.2. Table 8 compares WER across different datasets, illustrating the effectiveness of each method per dataset. Table 9 provides an aggregated view of WER across all methods, highlighting the overall performance trends. These results include: (1) Oracle performance: the lowest possible WER achievable using all tokens from transcripts output by the six ASRs; (2) Baseline performance: the WER of the top-performing ASR for each audio file; and (3) Mean: the average WER of the six ASR systems. The baseline and mean values are equivalent to the minimum and mean values in Ta-

	Oracle		Baseline		Mean		ROVER		GPT-4	
Dataset	WER	swf-WER	WER	swf-WER	WER	swf-WER	WER	swf-WER	WER	swf-WER
Sawyer	0.49	0.53	0.65	0.64	1.98	1.83	0.99	0.97	6.60	4.79
Cameron	0.48	0.53	0.59	0.61	1.10	1.15	0.81	0.81	0.97	1.01
CHiME-5	0.24	0.28	0.39	0.39	1.06	0.96	0.66	0.64	0.97	0.80
AMI	0.13	0.19	0.16	0.21	0.55	0.57	0.33	0.34	0.48	0.48
Covid	0.22	0.27	0.31	0.35	0.61	0.66	0.40	0.41	0.81	0.55
SPGI	0.01	0.05	0.02	0.05	0.17	0.21	0.05	0.07	0.04	0.07

Table 8: Comparison of WER across error correction methods by dataset

Metric	Oracle	Baseline	Mean	ROVER	GPT-4	Ft-T5
WER	0.26	0.36	0.91	0.54	1.64	7.21
swf-WER	0.31	0.37	0.90	0.54	1.28	7.22

Table 9: Comparison of WER across error correction methods

ble 2.

ROVER Outperforms. Table 8 shows that ROVER outperforms the mean WER across different datasets. This indicates ROVER's effectiveness in leveraging multiple ASR outputs to improve transcription accuracy. Conversely, GPT-4 exhibits a higher WER, particularly for more complex datasets like Sawyer, where it even underperforms compared to the mean WER. This suggests that while GPT-4 shows promise for transcription error correction of clean speech (as the literature suggests with clean speech such as LibriSpeech), it struggles with more complex audio scenarios.

Hallucinations in Fine-Tuned T5. Despite its potential, the performance of Ft-T5 is significantly worse than that of the other methods. Also, we observed hallucinations in the output from the Ft-T5 model, where the same word was repeated multiple times. This issue may be due to the limited amount of data available for fine-tuning.

8 Conclusions and Future Work

In this study, we have provided a comprehensive evaluation of the accuracy of state of the art open-source ASR systems on naturally occurring human conversations covering a range of acoustic and speaker characteristics. In contrast to its benchmark performances on various datasets, Whisper's performance on challenging audio conditions was ineffective and exhibited hallucination issues.

Our analysis indicates a gender bias in ASR performance for most systems, with male speakers experiencing higher WER compared to female speakers. Additionally, speaker age significantly impacts ASR performance, with younger adult speakers generally achieving lower WER. These findings emphasize the need for ASR systems to account for demographic factors to improve accuracy.

Among the error correction methods evaluated, ROVER consistently outperformed others, demonstrating its effectiveness in leveraging ensembles of ASR systems to enhance transcription accuracy. Another advantage of ROVER is that the graph it generates can be used to identify regions of speech that likely need human editing, and potentially even to prioritize which regions to focus on when one has a limited budget or time for human editing. We leave this idea for future work.

Furthermore, while our study focused on the computational efficiency of a 1-best approach, we acknowledge that an n-best approach could provide richer context. Future work could investigate the benefits of incorporating n-best hypotheses to further improve error correction strategies.

Our findings provide insights into the performance of various ASR systems under complex audio conditions and the challenges of error correction compared to ideal scenarios. Moreover, by releasing 14,400 audio-hypothesis-transcription pairs (2400 pairs per ASR, totaling 200 minutes of audio) on less commonly used datasets, we aim to foster open collaboration in the field.

9 Ethical Discussion and Limitations

We used ML models to detect age, accent, and gender when demographic information was unavailable. For example, age was provided in the Sawyer and Cameron datasets, and gender was specified in the AMI dataset. This raises ethical questions regarding the potential for misclassification. Relying solely on automated methods for sensitive attribute detection can perpetuate stereotypes, further embedding bias within ASR systems. Ethical practices in this context necessitate transparency about the limitations of these models and the incorporation of human oversight to correct misclassifications.

In this paper, we focused on English speech primarily from Western countries; our evaluation data and most of the models we evaluated were English-only. However, there are more than 7000 languages spoken worldwide and many of these languages have no language technology at all (Hou et al., 2020; Conneau et al., 2023). There are compelling social justice reasons to do much more research on non-English ASR. Furthermore, the datasets we used do not encompass all possible characteristics of speakers of English language. The collection of more varied datasets, especially including underrepresented dialects of English, would be necessary to fully understand the scope of English ASR performance disparities.

Another limitation is the focus on open-source ASR systems may not fully represent the capabilities of proprietary ASR technologies used in commercial applications. The performance and biases of these open-source systems may differ from those of their commercial counterparts, potentially limiting the generalizability of our findings.

Acknowledgments

We thank the Davis Institute for Artificial Intelligence for supporting this research with computational resources. We also thank the anonymous reviewers for their constructive comments, which helped improve this paper.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Proceedings of the Ninth European Conference on Speech Communication and Technology*.

Chandrasekhar Anantaram, Amit Sangroya, Mrinal Rawat, and Aishwarya Chhabra. 2018. Repairing ASR output by artificial development and ontology based learning. In *International Joint Conference on Artificial Intelligence*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. pages 1561–1565.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. 2023. Speech-based age and gender prediction with transformers. In *Proceedings of the ITG conference on Speech Communication*.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Chen Chen, Nana Hou, Yuchen Hu, Shashank Shirol, and Eng Siong Chng. 2022. Noise-robust speech recognition with 10 minutes unparalleled in-domain data. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4298–4302.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Ensiong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Horia Cucu, Andi Buzo, Laurent Besacier, and Corneliu Burileanu. 2013. Statistical error correction methods for domain-specific ASR systems. In *International Conference on Statistical Language and Speech Processing*.

Miguel del Río, Corey Miller, Ján Profant, Jennifer Drexler-Fox, Quinn Mcnamara, Nishchal Bhandari, Natalie Delworth, Ilya Pirkin, Migüel Jetté, Shipra Chandra, Peter Ha, and Ryan Westerman. 2023. Accents in speech recognition through the lens of a world englishes evaluation set. *Research in Language*, 21:225–244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Performance disparities between accents in automatic speech recognition (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16200–16201.

Rachel Dorn. 2019. Dialect-specific models for automatic speech recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20, Varna, Bulgaria. INCOMA Ltd.

L. F. D'Haro and Rafael E. Banchs. 2016. Automatic correction of ASR outputs by using machine translation. In *Proceedings of Interspeech*.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Alef Iury Siqueira Ferreira. 2023. wav2vec2-large-xlsr-53-gender-recognition-librispeech. https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender representation in french broadcast corpora and its impact on ASR performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, page 3–9, New York, NY, USA. Association for Computing Machinery.

John Garofolo, Ellen Voorhees, C Auzanne, Vincent Stanford, and B Lund. 1998. 1998 trec-7 spoken document retrieval track overview and results.

Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2008. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase ASR error rates. In *Proceedings of ACL-08: HLT*, pages 380–388, Columbus, Ohio. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. pages 5036–5040.

Jiaxin Guo, Minghan Wang, Xiaosong Qiao, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Yinglu Li, Chang Su, Min Zhang, Shimin Tao, and Hao Yang. 2023. Ucorrect: An unsupervised framework for automatic speech recognition error correction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jinxi Guo, Tara N. Sainath, and Ron J. Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo*.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *Proceedings of Interspeech*, pages 1037–1041.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. 2020. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7074–7078.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008*.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Science*, 117(14):7684–7689.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1997. Analysis of children's speech: Duration, pitch and formants. In *Fifth European Conference on Speech Communication and Technology*.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Yichong Leng, Xu Tan, Wenjie Liu, Kaitao Song, Rui Wang, Xiang-Yang Li, Tao Qin, Ed Lin, and Tie-Yan Liu. 2023. Softcorrect: Error correction with soft detection for automatic speech recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:13034–13042.

Yichong Leng, Xu Tan, Rui Wang, Linchen Zhu, Jin Xu, Wenjie Liu, Linquan Liu, Xiang-Yang Li, Tao Qin, Edward Lin, and Tie-Yan Liu. 2021a. FastCorrect 2: Fast error correction on multiple candidates for automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4328–4337, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiang-Yang Li, Ed Lin, and Tie-Yan Liu. 2021b. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *ArXiv*, abs/2105.03842.

Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024. Investigating asr error correction with large language model and multilingual 1-best hypotheses. In *Proc. Interspeech 2024*, pages 1315–1319.

Wenkun Li, Hui Di, Lina Wang, Kazushige Ouchi, and Jing Lu. 2021. Boost transformer with bert and copying mechanism for asr error correction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 482–495, Prague, Czechia. Association for Computational Linguistics.

Rao Ma, Mark John Francis Gales, Kate Knill, and Mengjie Qian. 2023a. N-best t5: Robust ASR error correction using multiple input hypotheses and constrained decoding space. *ArXiv*, abs/2303.00456.

Rao Ma, Mengjie Qian, Mark JF Gales, and Kate M Knill. 2023b. Adapting an unadaptable ASR system. *arXiv preprint arXiv:2306.01208*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark John Francis Gales, and Kate Knill. 2023c. Can generative large language models perform ASR error correction? *ArXiv*, abs/2307.04172.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech*.

Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, V Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. 2005. The ami meeting corpus. *5th International Conference on Methods and Techniques in Behavioral Research*.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv e-prints*, arXiv:2104.02014.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The kaldi speech recognition toolkit.

PyPI. 2024. Syllables: A python package for counting syllables. `https://pypi.org/project/syllables/`. Accessed: 2024-05-16.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Veronica Romero and Alexandra Paxton. 2023. Stage 2: Visual information and communication context as modulators of interpersonal coordination in face-to-face and videoconference-based interactions. *Acta Psychologica*, 239:103992.

R Keith Sawyer. 2013. *Pretend play as improvisation: Conversation in the preschool classroom*. Psychology Press.

Björn Schuller and Anton Batliner. 2013. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Elizabeth Shriberg. 2005. Spontaneous speech: How people really talk and why engineers should care. pages 1781–1784.

Krystal Starke. 2023. *"Can I Play with the Toys Now?": An Exploration of Preschool Storytelling Structure in Wordless Picturebook and Guided Play Contexts to Inform Teacher Training and Toy Development*. Ph.D. thesis, State University of New York at Buffalo.

Andreas Stolcke and Jasha Droppo. 2017. Comparing Human and Machine Errors in Conversational Speech Transcription. In *Proceedings of Interspeech*, pages 137–141.

Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Online. Association for Computational Linguistics.

Divya Tadimeti, Kallirroi Georgila, and David Traum. 2022. Evaluation of off-the-shelf speech recognizers on different accents in a dialogue domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6001–6008, Marseille, France. European Language Resources Association.

Hori Takaaki, Araki Shoko, Nakatani Tomohiro, and Nakamura Atsushi. 2013. Advances in multi-speaker conversational speech recognition and understanding. *NTT Technical Review*, 11(12):10–18.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proceedings of Interspeech*, pages 934–938.

Suramya Tomar. 2006. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582.

Lauren Werner, Gaojian Huang, and Brandon J. Pitts. 2019. Automated speech recognition systems and older adults: A literature review and synthesis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):42–46.

Yun Zhao, Xuerui Yang, Jinchao Wang, Yongyu Gao, Chao Yan, and Yuanfu Zhou. 2021. BART Based Semantic Correction for Mandarin Automatic Speech Recognition System. In *Proceedings of Interspeech*, pages 2017–2021.

Linchen Zhu, Wenjie Liu, Linquan Liu, and Edward Lin. 2021. Improving ASR error correction using n-best hypotheses. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 83–89.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. CommonAccent: Exploring large acoustic pretrained models for accent classification based on common voice. In *Proceedings of Interspeech*.

A Regression Plots

A.1 Regression Table

Table 10 on the next page presents the regression table, including the regression equation, R^2 value, and Bonferroni-corrected p-values for the cubic coefficients analyzed in Section 6.2. This includes each continuous feature: duration, syllables per second, noise level and age.

A.2 Effect of speech rate on WER: syllables per second

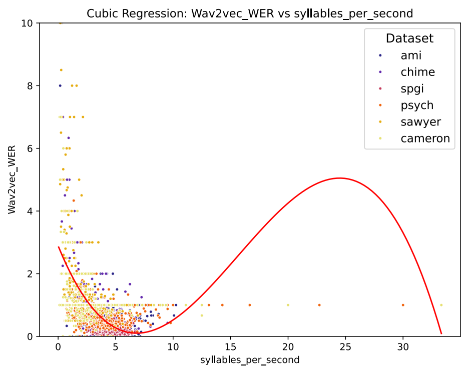

Figure 6: Syllables per second vs. Wav2Vec WER

A.3 Effect of speaker age on WER

Figure 7: Age vs. NVIDIA WER

Feature	System	Equation	R-squared	Cubic p-value corrected
duration	Hubert_WER	$1.143847 - 0.050390x + 0.003325x^2 - 0.00004419553x^3$	0.002007	1.000
duration	Nemo_WER	$0.742504 - 0.06105213x + 0.002466x^2 - 0.00002652036x^3$	0.029774	0.054
duration	Whisper_WER	$1.254405 - 0.1138149x + 0.007652x^2 - 0.000102221x^3$	0.000810	1.000
duration	DWhisper_WER	$0.696095 - 0.05223893x + 0.003812x^2 - 0.00005063922x^3$	0.003690	0.162
duration	Wav2vec_WER	$1.280726 - 0.09236164x + 0.005366x^2 - 0.00006851686x^3$	0.007801	0.016
duration	Kaldi_WER	$1.316008 - 0.04138082x + 0.003508x^2 - 0.00004977125x^3$	0.002428	0.623
syllables_per_second	Hubert_WER	$3.062961 - 1.003102x + 0.095750x^2 - 0.002054537x^3$	0.177675	<0.01
syllables_per_second	Nemo_WER	$1.164577 - 0.310183x + 0.030336x^2 - 0.0006422186x^3$	0.103088	<0.01
syllables_per_second	Whisper_WER	$3.005466 - 0.9785362x + 0.091793x^2 - 0.001950684x^3$	0.013453	0.085
syllables_per_second	DWhisper_WER	$1.717833 - 0.5726267x + 0.056937x^2 - 0.001197455x^3$	0.077001	<0.01
syllables_per_second	Wav2vec_WER	$2.890523 - 0.9004172x + 0.084326x^2 - 0.001794931x^3$	0.175698	<0.01
syllables_per_second	Kaldi_WER	$2.996386 - 0.850825x + 0.079613x^2 - 0.001700960x^3$	0.128006	<0.01
noise_level	Hubert_WER	$1.006434 + 8.550411 \times 10^6x + 0.941321x^2 + 0.0000001434107x^3$	0.002240	0.490
noise_level	Nemo_WER	$0.538853 + 3.193943 \times 10^6x + 0.351624x^2 + 0.00000005357002x^3$	0.001956	0.726
noise_level	Whisper_WER	$0.988826 + 5.316402 \times 10^6x + 0.585287x^2 + 0.0000000891687x^3$	0.000067	1.000
noise_level	DWhisper_WER	$0.576200 + 6.258136 \times 10^6x + 0.688963x^2 + 0.0000001049638x^3$	0.001660	1.000
noise_level	Wav2vec_WER	$1.033201 + 4.608494 \times 10^6x + 0.507353x^2 + 0.00000007729541x^3$	0.000774	1.000
noise_level	Kaldi_WER	$1.229745 + 7.710911 \times 10^6x + 0.848900x^2 + 0.0000001293303x^3$	0.001764	0.951
age	Hubert_WER	$2.278215 - 0.1313002x + 0.003599x^2 - 0.00003150352x^3$	0.069813	<0.01
age	Nemo_WER	$1.189794 - 0.06417423x + 0.001680x^2 - 0.00001438768x^3$	0.122077	<0.01
age	Whisper_WER	$2.919885 - 0.1956161x + 0.004967x^2 - 0.00004003803x^3$	0.012023	1.000
age	DWhisper_WER	$1.340553 - 0.07641958x + 0.002053x^2 - 0.00001787210x^3$	0.035992	0.035
age	Wav2vec_WER	$1.985948 - 0.1013680x + 0.002950x^2 - 0.00002734097x^3$	0.052151	<0.01
age	Kaldi_WER	$2.093572 - 0.09149935x + 0.002616x^2 - 0.00002364298x^3$	0.031291	0.011

Table 10: Regression analysis of continuous features on WER with cubic coefficients

B Violin Plots from Variance Analysis

B.1 Speaker overlap and WER variability

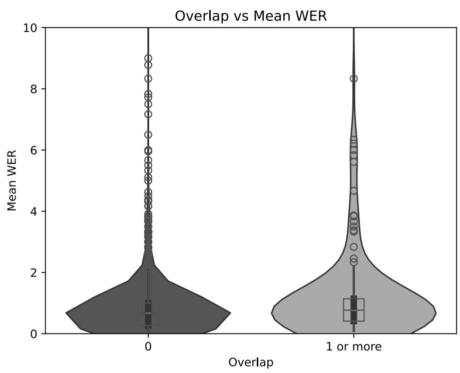

Figure 8: Overlapping speech

B.2 Gender-related WER differences

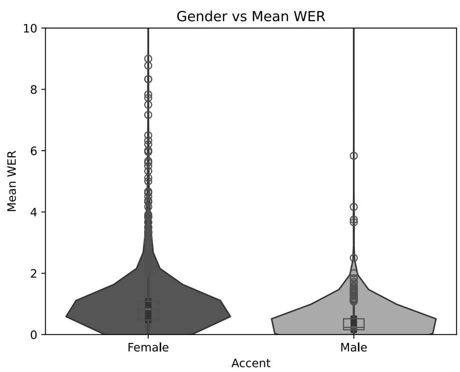

Figure 9: Gender

B.3 Accent-related WER discrepancies

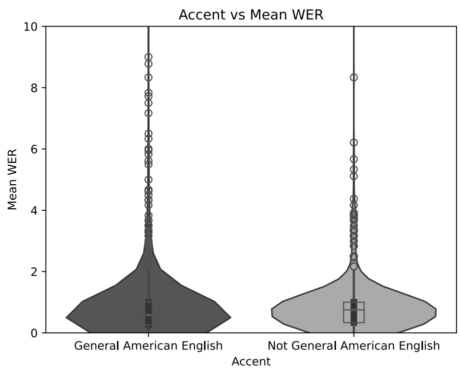

Figure 10: Accent