10th Workshop on Speech and Language Technology in Education (SLaTE)
22-24 August 2025, Nijmegen, Netherlands

Scaling *Conform*ation Bias in Automatic Speech Recognition

Nathan Roll[1], *Calbert Graham*[2]

[1]Department of Linguistics, Stanford University, USA
[2]Phonetics Laboratory, University of Cambridge, UK
nroll@stanford.edu, crg29@cam.ac.uk

Abstract

Modern ASR systems achieve remarkable accuracy on dominant English varieties, but they systematically fail when exposed to even subtle variation—such as the speech of English language learners. We present a comprehensive evaluation of ten ASR models on 300+ distinct English varieties. We find that as models increase in size and L1 English performance, the relative disparity for L2 speakers worsens. These effects are most notable in state-of-the-art conformer-based systems, which exhibit the lowest error rates overall. We also find a significant bias against speakers whose native language is tonal (e.g., Mandarin, Vietnamese) compared to speakers of languages with other prosodic typologies. These findings expose a need to develop and integrate the speech of English languages learners into training and evaluation pipelines.

Index Terms: Speech Recognition Bias, Non-native Speaker Recognition, Conformer Architectures, Linguistic Typology Effects, Accent Variability, Cross-Corpus Evaluation

1. Introduction

Automatic Speech Recognition (ASR) systems have become integral to human-computer interaction, powering applications from voice assistants to clinical documentation. However, despite remarkable technological advances driven by sophisticated architectures like the Conformer [1] and large-scale training datasets, ASR performance remains uneven across diverse speaker populations. This disparity is particularly pronounced for non-native (L2) English speakers, who constitute a significant portion of global users.

1.1. Documented Performance Disparities

Systematic biases in ASR systems are well-established across multiple dimensions. Koenecke et al. [2] demonstrated that Black speakers experience nearly double the word error rate (WER) of White speakers in commercial ASR systems. Gender disparities persist, with studies showing better performance for male voices [3], though these effects can be complex and intersectional [4].

Most relevant to this work are disparities related to accent and linguistic variety. DiChristofano et al. [5] showed that ASR performance correlates with speakers' first language, while studies on Dutch ASR confirmed significant biases against non-native speakers even in state-of-the-art models [6]. When these ASR systems used in downstream tasks [7] they can carry these uneven performance profiles into new domains. These disparities have substantial real-world implications, from healthcare quality [8] to accessibility for vulnerable populations [9, 10].

1.2. Recent Architectural Advances

Concurrent with documentation of these biases, ASR technology has undergone transformative changes. End-to-end neural architectures have largely replaced traditional HMM-based systems, with Conformer-based models setting new benchmarks by combining convolution with self-attention mechanisms. Large pre-trained have demonstrated improved robustness on "standard" English benchmarks through leveraging vast amounts of speech data. However, the extent to which these powerful architectures ameliorate or exacerbate these issues remains unclear.

1.3. The Conformation Bias Hypothesis

This paper investigates what we term "conformation bias" – the phenomenon where modern ASR systems, in their optimization for majority data, not only underperform on minority varieties but can also widen the performance gap as their overall capabilities increase. We conduct an extensive evaluation of ten models from the NVIDIA Canary, OpenAI Whisper, and Microsoft Phi-4 families across over 300 English varieties to test two core hypotheses:

1. As ASR models increase in size and their performance on L1 English improves, the *relative* performance gap for L2 speakers will systematically worsen. 2. This performance degradation is not uniform across L2 speakers, but is disproportionately severe for speakers whose native language is tonal, due to systematic phonetic variations for which current models are not optimized.

By testing these hypotheses, we aim to demonstrate that progress in ASR accuracy does not automatically equate to progress in fairness. This work builds on our previous investigations into ASR adaptation [11] and accent-based performance disparities [12] to provide a comprehensive analysis of systematic biases in state-of-the-art systems.

Our key findings reveal that while modern systems achieve lower overall WERs, L2 english speakers face elevated word error rates compared to L1 speakers. Moreover, speakers of tonal languages (e.g., Sino-Tibetan, Austroasiatic) consistently exhibit markedly higher WERs compared to Indo-European backgrounds, suggesting that typological distance from English significantly impacts recognition accuracy. These results underscore the urgent need for targeted interventions to ensure equitable ASR technology for all users.

2. Methods

This study comprehensively evaluates the performance of ten state-of-the-art ASR systems across a diverse range of English speech varieties. The methodology encompasses model selection, dataset curation, a systematic evaluation pipeline, and sub-

10.21437/SLaTE.2025-27

Table 1: *Word Error Rate (WER %) performance for various ASR models across seven speech corpora. Models include the NVIDIA Canary family (C, with parameter count in millions (M) or billions (B)), Microsoft's Phi-4, and OpenAI's Whisper series (W, with size denoted by T-tiny, S-small, B-base, M-medium, L-large, L-T-large-tuned). Corpora are grouped by speaker status: L1 (native) and L2 (non-native) English. The data shows that L2 English speakers experience systematically higher WERs across all models, and that the magnitude of this performance gap varies heavily by corpus.*

Group	Corpus	C180M-F	C1B-F	C1B	Phi-4	W-T	W-S	W-B	W-M	W-L	W-L-T
	SAA (L1 English)	1.2	2.3	0.8	1.1	3.6	1.3	2.5	1.0	0.9	0.8
L1	CMU-Arctic	2.9	2.3	2.4	1.7	7.4	3.8	5.8	3.3	2.8	3.0
	CommonVoice	5.7	4.2	4.7	4.7	-	-	-	-	-	-
	SAA (L2 English)	5.0	4.0	4.0	5.0	11.9	5.6	9.4	4.6	4.2	3.7
	Cambridge Phonetic	38.8	38.4	15.9	14.7	22.8	16.5	30.7	20.7	13.9	14.3
L2	L2-Arctic	8.9	6.8	7.1	7.8	-	-	-	-	-	-
	ALLSSTAR	10.7	9.3	9.4	24.8	-	-	-	-	-	-
	SANDI	24.0	16.6	16.7	23.0	22.2	14.3	18.5	13.1	12.8	12.9
Overall Mean WER (%)		12.1	10.5	7.6	10.3	13.6	8.3	13.4	8.5	6.9	7.0

sequent data analysis for generating insights and visualizations.

2.1. ASR Systems

Ten models from three prominent state-of-the-art ASR families were selected for evaluation, all based on or incorporating Transformer/Conformer architectures, representing the current state of the art:

- **NVIDIA Canary-1B (C1B):** A large encoder-decoder model with FastConformer encoder and Transformer decoder, comprising 1 billion parameters across 24 encoder and 24 decoder layers. The model supports multilingual ASR and translation capabilities.
- **NVIDIA Canary-1B-Flash (C1B-F):** An optimized variant with 883 million parameters, featuring an asymmetric architecture of 32 encoder and 4 decoder layers. This model achieves inference speeds exceeding 1000 RTFx while maintaining competitive accuracy.
- **NVIDIA Canary-180M-Flash (C180M-F):** A compact model from the same family with 182 million parameters (17 encoder and 4 decoder layers), demonstrating inference speeds over 1200 RTFx, suitable for resource-constrained deployments.
- **Microsoft Phi-4-Multimodal (Phi-4-MM):** A 5.6 billion parameter multimodal transformer built upon the Phi-4-Mini-Instruct backbone, capable of processing text, image, and audio inputs with a 128K token context length.
- **OpenAI Whisper:** A family of encoder-decoder Transformer models pre-trained on 680,000 hours of weakly supervised data. We evaluated six models from this family to analyze the effects of model scaling: **Whisper-tiny (W-T)**, **Whisper-small (W-S)**, **Whisper-base (W-B)**, **Whisper-medium (W-M)**, **Whisper-large (W-L)**, and a tuned version, **Whisper-large-tuned (W-L-T)**.

The Canary models utilize a concatenated SentencePiece tokenizer supporting multiple languages and were accessed via pre-trained checkpoints through the NVIDIA NeMo toolkit. All Canary models were trained on 85K hours of multilingual speech data. The Phi-4-MM model was loaded using the Hugging Face transformers library with necessary modifications to ensure compatibility for ASR tasks. All models were run on CUDA-enabled GPUs when available, with FlashAttention-2 utilized where applicable to optimize inference speed and memory usage.

2.2. Evaluation Datasets

A diverse set of seven publicly available English speech corpora was utilized to ensure broad representation of speech varieties, accents, and recording conditions, collectively representing over 300 English varieties and accents:

- **CMU-Arctic** [13]: Read speech from native and non-native English speakers in controlled recording conditions.
- **CommonVoice Accent Stratified** [14]: An accent-upsampled subset of the crowdsourced CommonVoice corpus, stratified by self-reported accent information to ensure balanced representation of diverse English varieties.
- **L2-Arctic** [15]: Read speech from L2 English speakers with diverse L1 backgrounds, recorded in controlled environments.
- **ALLSSTAR** [16]: Speech from L2 English speakers in academic contexts, featuring spontaneous and semi-spontaneous speech.
- **Cambridge Phonetics Corpus** A private corpus containing speech from English language learners during formal assessment tasks.
- **Speech Accent Archive (SAA)** [17]: Recordings of a standardized elicitation paragraph read by speakers representing over 250 native languages.
- **Speak & Improve Corpus (SANDI)** [18]: Speech samples from English language learners using an online practice platform, representing diverse proficiency levels and L1 backgrounds.

For each corpus, relevant metadata columns (audio, reference text, speaker ID, and variety/L1 information) were standardized. Audio data from all corpora was resampled to 16kHz and normalized to float32 format prior to processing.

2.3. Evaluation Pipeline

The evaluation pipeline was implemented in Python, processing each corpus systematically. For each audio sample, the pipeline performed the following steps:

1. **Audio Processing:** Audio arrays were normalized and resampled to 16kHz using librosa. For Canary models, audio was processed through the NeMo transcription interface. For

Phi-4-MM, a zero-shot prompt (“Transcribe the audio clip into text”) was employed with the audio data.

2. **Text Normalization:** Both reference and hypothesis texts underwent standardized normalization: conversion to lowercase, replacement of hyphens with spaces, removal of all non-alphabetic characters except apostrophes and spaces, and collapsing of multiple spaces.
3. **WER Calculation:** Word Error Rate was calculated using the jiwer library, comparing normalized reference and hypothesis texts. WER = (substitutions + insertions + deletions) / reference_length.
4. **Aggregation:** Results were aggregated at multiple levels: overall per corpus, per unique speech variety within each corpus, and by speaker type (L1/L2). Both micro-averaged WER (total errors / total words) and macro-averaged WER (mean of individual sample WERs) were computed.

2.4. Data Analysis and Visualization

The aggregated results underwent comprehensive statistical analysis:

- **Speaker Type Categorization:** Varieties were programmatically categorized as L1 (native English) or L2 (non-native English) based on variety identifiers. Additionally, L2 varieties were mapped to language families based on speakers’ L1 backgrounds.
- **Statistical Testing:** Multiple statistical tests were performed including Shapiro-Wilk tests for normality, Levene’s test for homogeneity of variance, and correlation analyses. Effect sizes (Cohen’s d) were calculated for L1 versus L2 comparisons.
- **Weighted Averaging:** Performance metrics were weighted by sample size when computing aggregate statistics for speaker groups and language families.

All analyses were designed to investigate the core research questions concerning ASR performance disparities across different English varieties and model configurations.

3. Results

3.1. Overall Performance Patterns

Table 1 presents the comprehensive WER performance of all ten ASR models across the seven evaluation corpora. The results reveal substantial variation in performance both across models and corpora, with overall corpus WERs ranging from 0.023 (CMU-Arctic) to 0.269 (cam phonetic).

Among the models evaluated, Canary-1B demonstrated the lowest average WER across all corpora (0.076), followed by Canary-1B-Flash (0.105) and Phi-4-MM (0.103), and Canary-180M-Flash (0.121). Notably, the performance advantage of the larger Canary-1B model was not uniform across all test conditions, with certain corpora showing minimal differences between model variants.

3.2. The Scaling Law of Disparity: Worsening Gaps with Better Models

A central finding of this study is the persistent and substantial performance gap between L1 (native) and L2 (non-native) English speakers across all models tested, and more critically, that this gap systematically worsens as models improve for L1 speakers. Table 1 establishes that L2 speakers consistently experience significantly higher WERs than L1 speakers across all

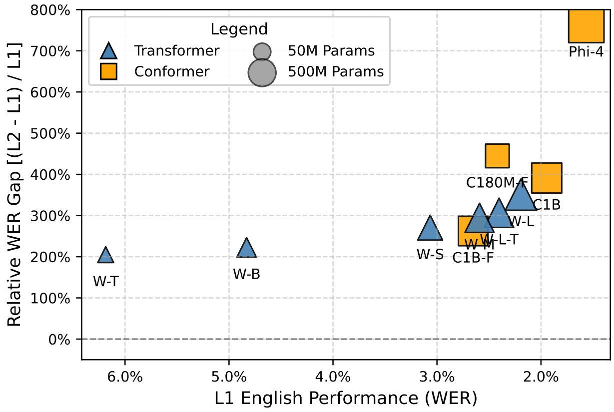

Figure 1: *Relationship between model performance on native English (L1) and the relative error gap for non-native speakers (L2). Each point represents a model, color-coded by architecture and sized by parameter count*

ten evaluated models.

The key insight emerges when examining the relationship between L1 performance and the relative L2 performance gap. For example, while the smallest Whisper model (W-T) shows a relative disparity of +243.6%, this gap widens significantly for the larger models, peaking at +504.0% for Whisper-medium (W-M). Figure 1 visualizes this inverse relationship, demonstrating a clear downward trend: as models achieve better performance on L1 English (moving left on the x-axis), the relative performance gap for L2 speakers systematically increases (moving up on the y-axis). We analyzed whether improvements in ASR performance for native (L1) English speakers translate to more equitable outcomes for non-native (L2) speakers. Figure 1 plots the average Word Error Rate (WER) on L1 corpora (x-axis) against the relative WER gap for L2 speakers (y-axis), defined as:

$$\frac{L_2\ \text{WER} - L_1\ \text{WER}}{L_1\ \text{WER}}$$

Each point represents an individual model, categorized by architecture (Transformer or Conformer) and parameter count. The plot reveals a concerning trend: as overall performance improves for L1 users (lower WER), the relative gap for L2 users often widens

Statistical analysis confirmed these differences were significant, with Cohen’s d effect sizes ranging from 0.11 to 1.42, indicating small to large practical significance depending on the model. Notably, Phi-4-MM, despite achieving competitive overall performance, exhibited the largest L1-L2 performance gap, suggesting that multimodal capabilities do not inherently address speech recognition biases.

3.3. Bias Against Tonal Language Backgrounds

The performance penalty for L2 speakers is not evenly distributed; it is strongly correlated with the speaker’s native language typology. Figure 2 presents average Word Error Rates (WER) for the Canary-1B model, grouped by the prosodic typology of speakers’ first language: Tonal, Pitch-Accent, and Stress-Accent. Native English speakers are included as a baseline. The number of speakers in each group is indicated on the plot. The results show a clear pattern: speakers from tonal language backgrounds are consistently misrecognized at substan-

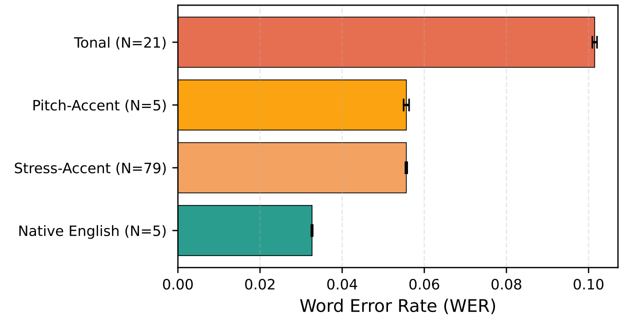

Figure 2: *ASR performance by speakers' prosodic L1 typology. Average WER is shown for Tonal, Pitch-Accent, and Stress-Accent language backgrounds, with native English as baseline.*

tially higher rates than other groups, suggesting a typological bias in ASR performance.

As shown, speakers from tonal language backgrounds experience a mean WER of approximately 10%, which is more than double the error rate for speakers of stress-accent languages (around 5.5%) and nearly three times that of native English speakers. This pattern was consistent across all Canary models, suggesting that acoustic-phonetic properties associated with tonal L1s may present particular challenges for current ASR architectures, potentially due to differences in prosodic patterns, pitch contours, and suprasegmental features that transfer to L2 English production.

3.4. Corpus-Specific Performance

Performance varied dramatically across corpora, reflecting differences in speech style, recording conditions, and speaker populations. The Cambridge Phonetic corpus, featuring English language learners in assessment contexts, showed the highest error rates across all models (average WER: 0.269), while CMU-Arctic, with controlled laboratory recordings, demonstrated the lowest (average WER: 0.023).

Interestingly, within corpora containing both L1 and L2 speakers, the typical L1 advantage was not always observed. In CommonVoice, for instance, L1 and L2 speakers showed nearly identical performance for most models, suggesting that crowd-sourced data collection methods may introduce variability that masks typical L1-L2 differences.

4. Discussion and Conclusion

Our comprehensive evaluation of ten state-of-the-art ASR systems reveals compelling evidence for "conformation bias." We have demonstrated two distinct manifestations of this bias. First, we identified a *scaling law of disparity*: as ASR models become more larger and accurate for native English speakers, the relative performance gap for non-native speakers systematically widens (Figure 1). Second, this bias is not evenly distributed, with speakers of tonal languages experiencing a disproportionately severe performance penalty (Figure 2). Together, these findings show that progress in ASR accuracy does not guarantee progress in fairness and can, in fact, exacerbate inequities.

This bias manifests as a complex interaction between speakers' linguistic backgrounds and ASR architectures, resulting in persistent 2.6-5.9× performance gaps between L1 and L2 speakers across all evaluated models. Most concerning is that the newest model tested (Phi-4-MM) exhibited the largest L1-L2 performance gap, suggesting that advancing model capabilities may inadvertently widen rather than narrow these disparities without deliberate intervention. This aligns with recent benchmarking efforts like ASR-FAIRBENCH [19], which provide systematic frameworks for evaluating fairness alongside accuracy.

Our analysis demonstrates a strong correlation between typological properties of speakers' L1s and ASR performance on their L2 English. Speakers of tonal languages (Sino-Tibetan, Austroasiatic, Tai-Kadai families) consistently experience 2-3× higher word error rates than Indo-European L1 speakers, indicating that current models struggle with systematic phonetic variations associated with tonal L1 backgrounds. This finding builds upon research showing that rhythmic and prosodic variations contribute to ASR bias [20] and underscores the need to consider typological distance in ASR evaluation. These disparities have critical implications for global ASR deployment, as these language families represent over 1.5 billion speakers worldwide who often use English as a second language in professional and technological contexts, potentially limiting equitable access to speech-enabled technologies.

The heterogeneous response to model scaling within the Canary family provides additional evidence for the complexity of conformation bias. While some L2 varieties benefited from increased model size, others experienced degraded performance, suggesting that current scaling approaches may amplify beneficial patterns while simultaneously reinforcing biases against specific acoustic-phonetic characteristics. Additionally, the superior performance of Canary-1B relative to its Flash-optimized variants for certain L2 populations indicates that architectural optimizations for inference efficiency may have unintended consequences for recognition accuracy on diverse speech, revealing important trade-offs between efficiency and equity in production deployments.

Several limitations warrant consideration when interpreting these findings. Our evaluation focused exclusively on English ASR, and the patterns observed may not generalize to multilingual or code-switching scenarios. While we evaluated contemporary models, the rapidly evolving ASR landscape means newer architectures may exhibit different bias patterns. Our analysis of language family effects, though revealing clear patterns, necessarily involved broad categorizations that may obscure important within-family variation. Additionally, as Li et al. [21] demonstrated, recording quality differences can create confounding effects in fairness studies, though our use of multiple corpora helps mitigate this concern.

These findings underscore that achieving equitable ASR requires moving beyond optimizing aggregate performance metrics. Merely scaling up models with existing data paradigms is insufficient and, as we have shown, can be counterproductive to fairness. The path forward demands a fundamental shift: we must actively develop and integrate diverse corpora of L2 speech, especially from underrepresented linguistic backgrounds like tonal languages, into both the training and evaluation pipelines. Future work must focus not only on lowering overall WER but on creating fairness-aware training objectives and evaluation metrics that explicitly measure and minimize performance disparities across user groups. Without these deliberate interventions, ASR technology risks becoming a tool that works best for those who need it least, failing the billions of language learners it should be designed to empower.

5. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, R. Rickford, John D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020. [Online]. Available: https://doi.org/10.1073/pnas.1915768117

[3] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 53–59. [Online]. Available: https://aclanthology.org/W17-1606/

[4] C. Harris, C. Mgbahurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 15 166–15 184. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.890.pdf

[5] A. DiChristofano, H. Shuster, S. Chandra, and N. Patwari, "Global performance disparities between english-language accents in automatic speech recognition," *arXiv preprint arXiv:2208.01157*, 2022. [Online]. Available: https://arxiv.org/abs/2208.01157

[6] M. Fückner, S. Horsman, I. Janssen, and P. Wiggers, "Uncovering bias in asr systems: Evaluating the performance of wav2vec2 and whisper for dutch speakers," in *2nd Dutch Speech Tech Day*, Hilversum, Netherlands, February 2024. [Online]. Available: https://research.hva.nl/en/publications/uncovering-bias-in-asr-systems-evaluating-the-performance-of-wav2

[7] N. Roll, C. Graham, and S. Todd, "Psst! prosodic speech segmentation with transformers," *arXiv preprint arXiv:2302.01984*, 2023.

[8] M. Zolnoori, S. Vergez, Z. Xu, E. Esmaeili *et al.*, "Decoding disparities: Evaluating automatic speech recognition system performance in transcribing black and white patient–nurse communication," *JAMIA Open*, vol. 7, no. 4, p. ooae130, 2024.

[9] D. Mujtaba, N. R. Mahapatra, M. Arney, J. S. Yaruss *et al.*, "Lost in transcription: Identifying and quantifying accuracy biases of automatic speech recognition systems against disfluent speech," *arXiv preprint arXiv:2405.06150*, 2024. [Online]. Available: https://arxiv.org/abs/2405.06150

[10] World Health Organization, "Ageism in artificial intelligence for health," Policy brief (no. 9789240040793), 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240040793

[11] N. Roll, C. Graham, Y. Tatsumi, K. T. Nguyen, M. Sumner, and D. Jurafsky, "In-context learning boosts speech recognition via human-like adaptation to speakers and language varieties," *arXiv preprint arXiv:2505.14887*, 2025.

[12] C. Graham and N. Roll, "Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[13] J. Kominek and A. W. Black, "The cmu arctic speech databases." in *SSW*, 2004, pp. 223–224.

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

[15] G. Zhao, E. Chukharev-Hudilainen, S. Sonsaat, A. Silpachai, I. Lucic, R. Gutierrez-Osuna, and J. Levis, "L2-arctic: A non-native english speech corpus," 2018.

[16] A. R. Bradlow, "Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts & recordings," https://speechbox.linguistics.northwestern.edu/allsstar, n.d., dataset repository, retrieved May 28, 2025.

[17] S. H. Weinberger and S. A. Kunath, "The speech accent archive: towards a typology of english accents." *Language & Computers*, vol. 73, no. 1, 2011.

[18] K. Knill, D. Nicholls, M. J. Gales, M. Qian, and P. Stroinski, "The Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback," 2025. [Online]. Available: https://doi.org/10.17863/CAM.114333

[19] A. Rai, S. Rahangdale, U. Anand, and A. Mukherjee, "Asr-fairbench: Measuring and benchmarking equity across speech recognition systems," *arXiv preprint arXiv:2505.11572*, 2025. [Online]. Available: https://arxiv.org/abs/2505.11572

[20] L. Lai and N. Holliday, "Exploring sources of racial bias in automatic speech recognition through the lens of rhythmic variation," in *INTERSPEECH 2023*, 2023, pp. 1284–1288. [Online]. Available: https://www.isca-archive.org/interspeech_2023/lai23_interspeech.html

[21] C. Li, T. Cohen, and S. Pakhomov, "Re-examining racial disparities in automatic speech recognition performance: The role of confounding by provenance," *arXiv preprint arXiv:2407.13982*, 2024. [Online]. Available: https://arxiv.org/abs/2407.13982