Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems

Anand Rai, Satyam Rahangdale, Utkarsh Anand, Animesh Mukherjee

[1]Indian Institute of Technology, Kharagpur, India

raianand.1991@gmail.com, satyamrahangdale196@kgpian.iitkgp.ac.in, ua28012006@kgpian.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Automatic Speech Recognition (ASR) systems have become ubiquitous in everyday applications, yet significant disparities in performance across diverse demographic groups persist. In this work, we introduce the ASR-FAIRBENCH leaderboard which is designed to assess both the accuracy and equity of ASR models in real-time.

Leveraging the Meta's Fair-Speech dataset, which captures diverse demographic characteristics, we employ a mixed-effects Poisson regression model to derive an overall fairness score. This score is integrated with traditional metrics like Word Error Rate (WER) to compute the Fairness Adjusted ASR Score (FAAS), providing a comprehensive evaluation framework. Our approach reveals significant performance disparities in SOTA ASR models across demographic groups and offers a benchmark to drive the development of more inclusive ASR technologies.

Index Terms: speech recognition, fairness benchmark, leaderboard

1. Introduction

Automatic Speech Recognition (ASR) systems have revolutionized human-computer interaction, powering applications from virtual assistants to real-time transcription services. However, despite impressive strides in overall accuracy, these systems often exhibit significant performance disparities across diverse demographic groups. Variations in accent, gender, age, and linguistic background can lead to unequal ASR performance [1, 2, 3], raising concerns about bias and fairness in deployed technologies.

While traditional ASR leaderboards [4, 5] focus solely on overall accuracy—typically measured by Word Error Rate (WER)—this approach overlooks critical fairness concerns. Many commercial and open-source ASR models are benchmarked on public datasets that lack diverse speech characteristics, perpetuating biases against underrepresented groups [6]. As a result, models optimized solely for low WER may perform inadequately for speakers with varied accents, dialects, and other demographic features. Establishing a dedicated fairness leaderboard is essential to ensure that ASR systems are evaluated not only for accuracy but also for equity, promoting more inclusive and responsible technology development.

Our proposed **ASR-FAIRBENCH** leaderboard utilizes the Fair-Speech dataset [7], a public corpus containing 26,500 utterances from 593 compensated U.S. participants with self-reported demographics including age, gender, ethnicity, location, and native language status. The dataset spans seven voice assistant domains: music, capture, utilities, notification control, messaging, calling, and dictation.

ASR-FAIRBENCH is the first real-time benchmarking platform evaluating both fairness and accuracy, using 10% stratified samples from the Fair-Speech dataset. It enhances mixed-effects Poisson regression by integrating it with WER to produce a Fairness-Adjusted ASR Score (FAAS), benchmarks five models, and introduces a five-tier classification scheme from "severely biased" to "exemplarily fair" for overall and attribute-specific assessments.

2. Methodology

2.1. Dataset

A stratified 10% sample was extracted from the fair speech dataset to benchmark ASR models on the **ASR-FAIRBENCH** leaderboard. As shown in Table 1, the entropy values across all demographic attributes remain nearly identical between the original and sampled datasets, ensuring that the distribution is preserved while reducing the evaluation data by 90% and significantly decreasing inference time.

Attributes	Original	Sampled
Total samples	26,471	2,648
Total Duration (hrs)	54.56	5.46
Entropy (Gender)	0.99	0.99
Entropy (First Lang.)	1.40	1.39
Entropy (Socioec. Bkg.)	1.27	1.27
Entropy (Ethnicity)	2.50	2.50

Table 1: *Original vs. sampled dataset comparison*

2.2. Evaluation metric

We present a framework for quantifying fairness in ASR systems through the Fairness-Adjusted ASR Score (FAAS).

2.2.1. Word error rate and fairness model

The WER metric evaluates ASR performance: $WER = \frac{S+D+I}{N}$, where S, D, and I represent substitutions, deletions, and insertions. To assess fairness across demographic groups, we employ mixed-effects Poisson regression:

$$\log(WER_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \qquad (1)$$

where X_i represents demographic attributes, Z_i represents covariates, and β_1 quantifies disparity through disparity ratio $= e^{\beta_1}$.

2.2.2. Fairness score calculation

For each demographic group, we compute predicted WER:

$$\widehat{WER}_g = \frac{e^{(\beta_0 + \beta_g + \beta_{logRef} \cdot \bar{X})}}{e^{\bar{X}} - 1} \qquad (2)$$

Raw fairness scores are scaled to 0-100:

$$\text{Raw fairness score}_g = 100 \times \left(1 - \frac{\widehat{WER}_g - \min(\widehat{WER})}{\max(\widehat{WER}) - \min(\widehat{WER})}\right) \tag{3}$$

The category score is calculated as: Category score $= \sum_g p_g \times$ Raw fairness score$_g$

2.2.3. Statistical adjustment and FAAS

Statistical significance of fairness disparities is assessed via the Likelihood Ratio Test comparing full and reduced models:

$$LRT = 2 \times (\log L_{full} - \log L_{reduced}) \tag{4}$$

where L_{full} includes the demographic attribute and $L_{reduced}$ excludes it. The resulting p-value indicates whether disparities are statistically significant. When disparities are significant ($p < 0.05$), we apply a proportional penalty to the category score:

$$\text{Adjusted score} = \text{Category score} \times \left(\frac{p}{0.05}\right) \tag{5}$$

For non-significant disparities ($p \geq 0.05$), the category score remains unchanged. The overall fairness score is calculated as a weighted average across all demographic categories:

$$\text{Overall score} = \frac{\sum_c w_c \times \text{Adjusted score}_c}{\sum_c w_c} \tag{6}$$

where w_c represents optional weighting factors for each category c. Finally, the **Fairness-Adjusted ASR Score (FAAS)** integrates recognition accuracy with fairness in a single metric:

$$FAAS = 10 \times \log_{10}\left(\frac{\text{Overall Score}}{WER}\right) \tag{7}$$

This logarithmic formulation ensures that improvements in both fairness and accuracy contribute positively to the final score, providing a comprehensive evaluation of ASR system performance across diverse demographic groups. The detailed formulation of the **FAAS** metric can be found in the HuggingFace Space[1].

3. Application overview

The **ASR-FAIRBENCH** leaderboard is a web-based platform designed for evaluating and ranking ASR models. Built with React.js, leveraging an NVIDIA T4 GPU for inference, it offers an interactive platform for model submissions, performance analysis, and real-time leaderboard tracking. The platform is fully reproducible from its repository[2]. The live leaderboard[2] can be accessed by users to submit their ASR models and receive an audit within minutes. The platform offers intuitive graphical representations of results, including box plots and histograms as shown in Figure 1, making it easier to interpret model performance.

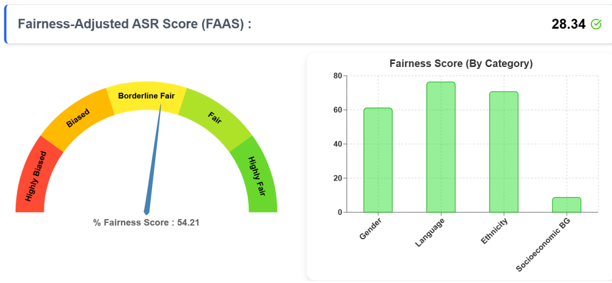
4. Conclusion

Our **ASR-FAIRBENCH** leaderboard (Figure 1a) highlights Whisper models as leading in fairness-adjusted ASR scores, with Whisper-medium scoring highest at 29.41. Despite higher WER, Whisper-tiny demonstrates better overall fairness. Fine-tuned Wav2Vec and Hubert models show efficiency but lag in fairness. While **FAAS** generally follows WER trends, the comparison between fine-tuned Wav2Vec-large and Hubert-large reveals that fairness issues can outweigh slight WER improvements—emphasizing the leaderboard's focus on multidimensional fairness evaluation.

[1] https://huggingface.co/spaces/satyamr196/ASR-FairBench
[2] https://github.com/SatyamR196/ASR-FairBench

Model Name	Fairness Adjusted ASR Score	Average WER	Average RTFX	Overall Fairness Score
openai/whisper-medium	29.41	0.06	0.13	56.57
openai/whisper-tiny	24.25	0.24	0.13	64.78
facebook/wav2vec2-large-960h	20.51	0.43	0.03	48.56
facebook/hubert-large-ls960-ft	20.31	0.4	0.03	42.94
facebook/wav2vec2-base-960h	19.82	0.48	0.66	45.86

(a) ***ASR-FAIRBENCH** leaderboard*



(b) *Summarized results view.*

Figure 1: *Key UI features: **ASR-FAIRBENCH** leaderboard, result summary, and performance analysis for whisper-small.*

5. References

[1] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Interspeech*, 2017.

[2] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[3] A. K. Rai, S. D. Jaiswal, and A. Mukherjee, "A deep dive into the disparity of word error rates across thousands of nptel mooc videos," in *AAAI ICWSM*, 2024.

[4] Y. Wang, A. Alhmoud, and M. Alqurishi, "Open universal arabic asr leaderboard," *arXiv preprint arXiv:2412.13788*, 2024.

[5] V. Srivastav, S. Majumdar, N. Koluguri, A. Moumen, S. Gandhi, H. F. Team, N. N. Team, and S. Team, "Open automatic speech recognition leaderboard," urlhttps://huggingface.co/spaces/huggingface.co/spaces/open-asr-leaderboard/leaderboard, 2023.

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

[7] I.-E. Veliche, Z. Huang, V. A. Kochaniyan, F. Peng, O. Kalinli, and M. L. Seltzer, "Towards measuring fairness in speech recognition: Fair-speech dataset," *arXiv preprint arXiv:2408.12734*, 2024.