How to Measure Bias in Automatic Speech Recognition Systems?

A Bias Metric Without a Reference Group

Tereza Ležovičová

EEMCS, Delft University of Technology, The Netherlands

25-06-2024

1 Introduction

Automatic speech recognition (ASR) is concerned with converting spoken language into text. However, as human speech is naturally highly variable, depending on context, region, or age, producing an objective transcription of the speech [2] is challenging. Research shows that ASR models show bias, e.g., between genders and races [4].

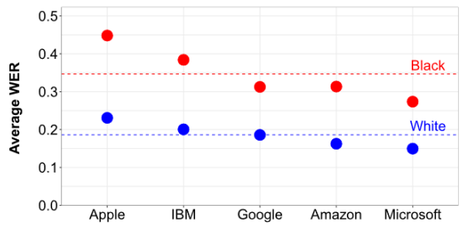

Figure: Average WER across multiple models [4]

Group fairness

Accepting similar outcomes across different speaker (demographic) groups [7]

Bias

A performance difference, specifically an error rate difference, between various speaker groups [6]

1.1 Bias metric

$$G2_{\text{reference group}_{\text{diff}}} = b_{\text{spk}_g} - b_{\text{reference group}}$$

$G2_{\text{reference group}_{\text{diff}}}$ refers to Bias metric

b refers to Base metric

- in our case Word Error Rate (WER)

2 Problem description

Existing bias metrics typically use a “reference group". The reference group serves as a baseline, i.e., its word error rate is compared to that of other groups to observe potential bias in the system.

Limitations of this approach:

a. **Binary decision outcomes**
   - The privileged group and the non-privileged group [5]
   - No insight into inter-group biases

b. **Standardizing reference group characteristics**
   - Ethical concern regarding the standardization of what constitutes a “normal" speech characteristic

Aim

Approach to measuring bias in ASR systems that eliminates the use of a reference group

3 Methodology

I evaluate these models on my bias metric:

- the E2E **Conformer** model trained on CGN dataset (*NoAug*, *SpAug*, *SpSpecAug*)
- Open AI’s **Whisper** model (W_s, W_sFt_{cgn})

I use **Jasmin dataset** as my testing set with two speaking styles (Rd, HMI) and speaker groups:

- Dutch Children (DC)
- Dutch Teenagers (DT)
- Dutch Older Adults (DOA)
- Non-native Teenagers (NnT)
- Non-native Adults (NnA)

4 Final Design

$$b_{\text{avg}_{(MS)}} = \frac{1}{G}\sum_{g \in G} \text{WER}_{g_{(MS)}}$$

$$G2_{\text{avg log ratio}}(g_{(MS)}) = -\ln \frac{b_{g_{(MS)}}}{b_{\text{avg}_{(MS)}}}$$

$$\text{Sum}_M = \sum_{g \in G} \text{B}(g_{M\text{Rd}}) + \text{B}(g_{M\text{HMI}})$$

5 Limitations and Future Work

- Hard to estimate effectiveness of bias metric as there is no ground truth to bias
- Simulated environment where one or multiple groups are artificially biased [1]

6 Results

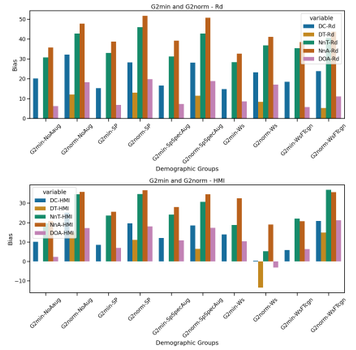

Figure: $G2_{min_{diff}}$ and $G2_{norm_{diff}}$ [6]

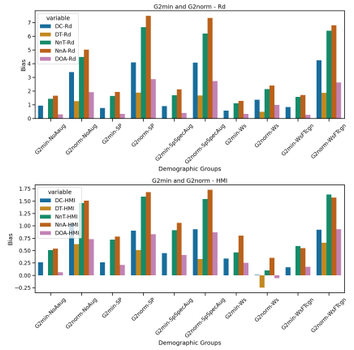

Figure: $G2_{min_{reldiff}}$ and $G2_{norm_{reldiff}}$ [6]

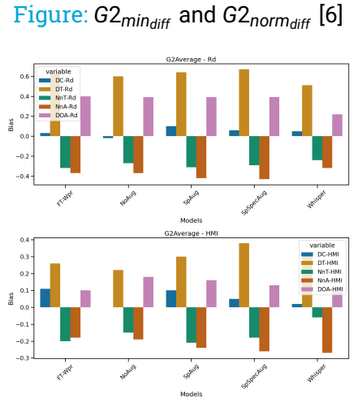

Figure: $G2_{avg\,\log ratio}$ [3]

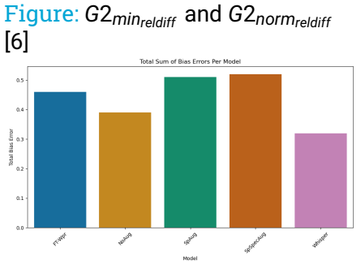

Figure: $\sum G2_{avg\,\log ratio}$ [3, 1]

References

[1] A. Elobaid, N. Ramoly, L. Younes, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris. “Sum of Group Error Differences: A Critical Examination of Bias Evaluation in Biometric Verification and a Dual-Metric Measure”. In: *18th International Conference on Automatic Face and Gesture Recognition (FG)* (Apr. 2024). DOI: 10.48550/arXiv.2404.15385.

[2] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg. “Quantifying Bias in Automatic Speech Recognition”. In: (Apr. 2021). DOI: 10.48550/arXiv.2103.15122.

[3] W. Hutiri, T. Patel, A. Ding, and O. Scharenborg. “As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research”. Submitted to Interspeech 2024. 2024.

[4] A. Koenecke et al. “Racial disparities in automated speech recognition”. In: *Proceedings of the National Academy of Sciences* 117.14 (Apr. 2020), pp. 7684–7689.

[5] K. Lum, Y. Zhang, and A. Bower. “De-biasing “bias” measurement”. In: June 2022, pp. 379–389. DOI: 10.1145/3531146.3533105.

[6] T. Patel, W. Hutiri, A. Ding, and O. Scharenborg. “How to Evaluate Automatic Speech Recognition: Comparing Different Performance and Bias Measures”. Work in progress. 2024.

[7] M.-H. Yeh, B. Metevier, A. Hoag, and P. Thomas. “Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics”. In: (June 2024), pp. 518–528. DOI: 10.1145/3630106.3658922.