Integrating Base Performance and Performance Differences in Automatic Speech Recognition Metrics

Bram Vincent van Vliet

Introduction

Bias Metrics

In **Automatic Speech Recognition (ASR)**, there are very little defined ways of measuring bias, including the following [1]:

Group-to-min Absolute Difference ($G2_{m,a}$):

$$\text{Bias}_{\text{abs},i} = \text{Base}_i - \text{Base}_{\text{min}} \quad (1)$$

Group-to-norm Absolute Difference ($G2_{n,a}$):

$$\text{Bias}_{\text{abs},i} = \text{Base}_i - \text{Base}_{\text{norm}} \quad (2)$$

Group-to-min Relative Difference ($G2_{m,r}$):

$$\text{Bias}_{\text{rel},i} = \frac{\text{Base}_i - \text{Base}_{\text{min}}}{\text{Base}_{\text{min}}} \quad (3)$$

Group-to-norm Relative Difference ($G2_{n,r}$):

$$\text{Bias}_{\text{rel},i} = \frac{\text{Base}_i - \text{Base}_{\text{norm}}}{\text{Base}_{\text{norm}}} \quad (4)$$

where *Base_i*, *Base_min* and *Base_norm* are the base performances for group *i*, min and norm groups respectively.

Research Question

How to incorporate both **performance difference** and **actual performance** in a bias metric?

Experimental Setup

- Output of Patel et al. [1], tested on the **JASMIN** dataset
- For every speaker, data on the words spoken
- 5 types of ASR models, some including **speed augmentation** (SpAug), **speed + spectral augmentation** (SpSpecAug) or **fine-tuning** (FT-Wpr)

- **JASMIN dataset:**
  - Dutch Children (**DC**)
  - Dutch Teenagers (**DT**)
  - Dutch Seniors (**DOA**)
  - Non-native Teenagers (**NnT**)
  - Non-native Adults (**NnA**)
- **ASR Models:**
  - ***NoAug***
  - ***SpAug***
  - ***SpSpecAug***
  - ***FT-Wpr***
  - ***Whisper***

Methodology

Weighted Bias Metrics

In order to answer the main research question, the following bias metrics were created:

Weighted Performance Bias:

$$PD_i = \text{Base}_i - BP \quad (6)$$

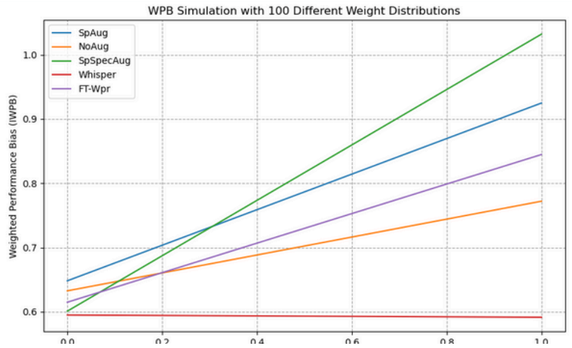
$$\text{WPB} = \frac{1}{n}\sum_{i=1}^{n}\left(w_1 \cdot \frac{PD_i}{BP} + w_2 \cdot \text{Base}_i\right) \quad (7)$$

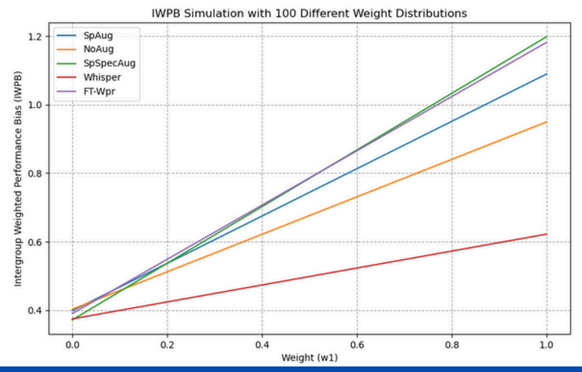
Intergroup Weighted Performance Bias:

$$PD_{ij} = \text{Base}_i - \text{Base}_j \quad (8)$$

$$\text{IWPB} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\left(w_1 \cdot \frac{PD_{ij}}{BP} + w_2 \cdot \text{Base}_i\right) \quad (9)$$

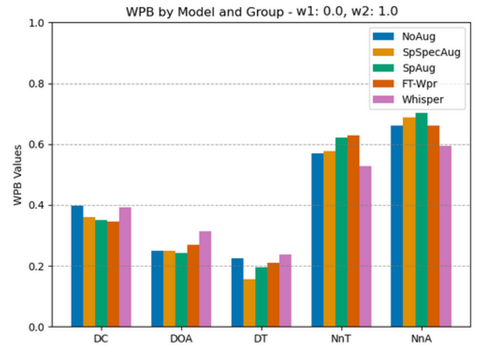
Results: Weight selection

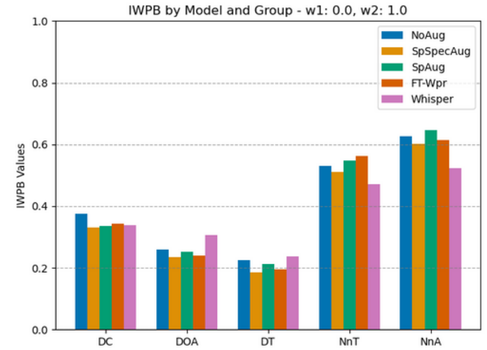




Results: Measured Bias

Figures 1 and 2: WPB and IWPB values per Model and Speaker Group





Limitations

- Absence of ground truth makes results hard to verify
- Current optimal weights don't leverage idea of metric
- Future research:
  - weight selection
  - comparison of weights among each other
  - Other methods of combining without weighted average

Conclusion

- New metrics show similar trends to existing methods
- Non-native speech still shows the most bias, Dutch teenagers show the least
- Further optimizations are possible

References

[1] 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