Integrating Base Performance and Performance Differences in Automatic Speech Recognition Metrics

Bram Vincent van Vliet[1]

Supervisors: Odette Scharenborg[1]**, Jorge Martinez Castinada**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Bram Vincent van Vliet
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Jorge Martinez Castaneda and Merve Gürel

Abstract

Automatic Speech Recognition (ASR) systems are becoming increasingly popular in this day and age. Unfortunately, due to inherent biases within these systems, performance disparities exist among specific demographic groups. Bias metrics can be used to measure this bias. Within ASR they represent a niche area that has not yet been thoroughly explored. The few bias metrics that exist in literature mainly centre around the performance differences between speaker groups. This paper proposes two new bias metrics that focus not only on performance differences, but also take the base performance into account: Weighted Performance Bias (WPB) and Intergroup Weighted Performance Bias (IWPB). Although the lack of ground truth makes the results less easily interpretable, the results show similar trends within the new metrics as those defined in literature: bias is greatest among non-native Dutch speech.

Index terms: speech recognition, bias measurement, performance evaluation

1 Introduction

Performance evaluation is a critical aspect of various domains, such as natural language processing, machine learning and Automatic Speech Recognition (ASR). The performance of an ASR system across different groups or contexts can vary due to inherent biases [11]. In the context of automatic speech recognition, it has been shown that these biases also exist, with ASR systems particularly exhibiting gender, age, non-native speech and regional accent bias [3]. Differing performance due to these characteristics can lead to a situation where certain groups of people are not able to use the system as successfully as other groups. For a system to remain fair and unbiased, the aim is to build systems such that all users can use it just as effectively as one another. This makes it essential to have an effective way of quantifying bias in ASR systems.

ASR systems are designed to convert spoken language into text. These models are typically developed using machine learning techniques, especially deep learning. The training process of an ASR system involves feeding the model large datasets of paired audio recordings and their corresponding text transcriptions [6]. The model then learns to map audio signals to textual representations by identifying patterns and features within the data [12]. Key components of ASR models include acoustic models, which interpret speech signals, and language models, which in turn predict words. Once trained, ASR models can be deployed in various applications, such as virtual assistants, transcription services, and voice-controlled interfaces. These models process real-time or recorded audio, convert it to text, and deliver transcriptions.

Bias is a significant concern in machine learning as it can lead to unfair and incorrect outcomes. This often arises from training data where certain groups may be underrepresented [1]. This can cause machine learning models to produce inaccurate output that can disproportionately affect those underrepresented groups [7]. Addressing bias requires diverse and representative datasets, careful monitoring of model performance across demographics, and implementing fairness-aware algorithms [1]. Due to ASR models being trained on traditionally 'standard' (minimal accent, well articulated) data, bias poses a challenge for these types of models as well [16].

To address the issue of bias, bias metrics can be used to quantify bias in performance assessments. Feng et al. attempted to capture bias by using the difference in Word Error Rate (WER) between different demographic groups [4]. This performance metric evaluates ASR performance, which can then be used in bias metrics to quantify the bias. Other performance metrics exist, which will be further explained in appendix A, but for this study, the industry-standard WER was used.

Performance difference refers to the gap or variation in performance between different demographic groups. For instance, if one demographic group has a WER of 5% and another group has a WER of 10%, the performance difference would be 5%. This difference indicates disparities in the ASR system's performance for each group. Base performance refers to the absolute WER values, regardless of the demographic group. It represents the accuracy of the ASR system for each specific group.

Currently, there is no universally agreed upon definition for bias in ASR, but the first attempts at creating bias metrics utilised the difference in performance [4]. The main issue with a metric that only examines the WER performance difference between speaker groups without considering the base performance for that group, is that it may obscure significant details about the overall performance of the ASR system. While focusing solely on the differences between groups can highlight relative disparities, it does not provide insight into the absolute effectiveness of the system. For example, if an ASR system has uniformly high WER across all groups, the differences between groups might be small, indicating low bias. However, the system's overall performance would still be poor, affecting all users negatively. Thus, a metric that only considers the differences between speaker groups could fail to capture the absolute performance and could potentially misrepresent the bias of the ASR system.

Designing a measurement approach in the form of a bias metric that considers not only the disparities in performance between groups, but also the overall accuracy level of the ASR system could be a better way to measure bias. This approach ensures that an ASR system is not only equitable across different speaker groups but also meets a high standard of performance for all users, providing a more comprehensive assessment of bias.

1.1 Related work

This research is closely related to the work by Patel et al. in '*How to Evaluate Automatic Speech Recognition: Comparing Different Performance and Bias Measures*' [14]. The output of the trained models from this research was used as the input to this research. Therefore, this section will first clarify the experiment performed by T. Patel.

Related paper explanation

In the paper, Patel et al. conducted an in-depth analysis to evaluate bias in ASR systems by assessing bias across different speaker groups, based on WER.

The models were trained on data from the Corpus Gesproken Nederlands (CGN) [15], consisting mainly of native Dutch adult speech. The JASMIN Corpus[1] was used to test the ASR systems. This corpus includes speech from native Dutch children (DC), Dutch teenagers (DT), and Dutch seniors (DOA), as well as non-native teenagers (NnT) and adults (NnA). These speech inputs are then categorised into Read speech and Human Machine Interaction (HMI) speech [2].

Based on the methodology outlined in [14], performance and bias metrics were evaluated using the JASMIN dataset across five distinct ASR models. The first three models are conformer models trained without any data augmentation (NoAug), enhanced with additional training on speed-perturbed speech (SpAug) and a combination of speed-perturbated and spectral augmented speech (SpSpecAug). The final two models are OpenAI Whisper models, one trained on normal data (Whisper) and the other trained on fine-tuned data (FT-Wpr).

Multiple bias measures were used to compare different speaker groups relative to a reference group, including Group-to-min and Group-to-norm. Both compare each group based on WER difference to a specific baseline group: the best performing group (Group-to-min) or the norm speaker group (Group-to-norm). Both measures have absolute and relative variants. The equations for the bias metrics used by Patel et al. [14] are as follows:

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

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

where Base_i is the base performance for group i and Base_{min} is the group with the minimum error rate.

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

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

where $\text{Base}_{\text{norm}}$ is the base performance for the norm group.

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

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

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

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

By calculating the absolute and relative differences between the groups and a reference (either the minimum WER observed or a norm speaker group), base performance metrics across different speaker groups can be compared. Bias can then be measured to be employed as comparison material for a new bias metric for this research.

It is imperative that bias can be properly measured and as such this research aims to create a new method of calculating bias. Patel et al. concluded that while error rates are fundamental to performance evaluation in ASR, it 'does not reflect the performance and bias within and across speaker groups well' [14]. While data augmentation techniques can somewhat mitigate bias, substantial disparities in ASR performance across different speaker groups persist. This shows the necessity for ongoing advancements in bias metrics for ASR to ensure accurate recognition for all users.

This paper aims to solve this issue by creating a new bias metric that includes the overall accuracy of the system as well as the differences between groups. In particular, the output of the ASR models trained in [14] was utilised to implement the new metrics, with the shape of the data explained in 4.1. An explanation of how the data was used can be found in section 2: Methodology.

1.2 Research Question

The research conducted for this paper aims to provide insight into the following proposed research question:

"How to incorporate both the performance difference and base performance in a bias metric?"

An experiment was conducted to explore how to effectively integrate both the 'performance difference' and 'base performance' aspects of an ASR system into a single unified metric. By exploring different methods to integrate these components of an ASR system, the research could identify how capable the proposed bias metrics are in recognising bias.

2 Methodology

In this section, the rationale behind the new bias metrics will be introduced. Subsequently, the tools used in this research will be explained.

2.1 Bias Metrics

To combine performance difference and base performance, the Weighted Performance Bias (WPB) and Intergroup Weighted Performance Bias (IWPB) were created. WPB iterates over all groups and takes the weighted average of the relevant aspects: the performance difference and the base performance. Weights w_1 and w_2 are used to determine the

[1] https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus-c/

ratio of importance for the performance difference and the base performance, leading to an adjustable metric by varying the weight values. IWPB works similarly by computing the weighted average but differs in the performance difference calculation. Rather than making use of the predefined performance differences (min and norm), IWPB computes the performance difference between one group and every other group, then takes the average of that to be the total performance difference. This was done in an attempt to capture bias in a different way to the $G2$ variants. The equations for both WPB and IWPB can be found in section 3.1.

The output of the models from Patel et al. [14] was used to create these new bias metrics. Evaluation of the new bias metrics is possible by comparison with a defined reference bias metric. The found bias values for all metrics could be compared by applying the same input for the new metrics. While not ideal, this evaluation method is currently the most logical approach due to the absence of a defined ground truth for bias in this context. This will be further discussed in section 5.

2.2 Tools

To conduct the experiments of this research, a range of tools and libraries were used. The processing and visualisation of the data were handled using Python 3.10, which provides a robust and versatile environment for data analysis.

Data Processing

The data was processed using multiple libraries, including Pandas [18], NumPy [9], Seaborn [17] and Matplotlib.pyplot [10]. These libraries were used in conjunction and aided in data processing and visualisation, allowing the complex data to be read, stored and presented in a useful manner.

Since the data used in this study is licensed for research purposes only, it cannot simply be stored within the research repository, as that repository becomes public data on the TU Delft repository. Therefore, all processing was done on the DelftBlue supercomputer[2]. This ensured that processing was efficient as this computer was more than capable of running the program in a small amount of time. It also meant that the input data was kept within the university's secured environment, since none of the licensed data was stored in the Github repository and was only run from within the DelftBlue system.

Custom Scripts

Custom scripts were written to calculate the error rates and bias metrics. These scripts automated the process of extracting relevant data, performing the necessary calculations, and generating visualisations. By developing these scripts, it was ensured that the calculations were tailored to the specific requirements of this research and could be adapted as needed. Finally, the calculated error rates and bias metrics were analysed to see if the data showed a combination of the performance difference and base performance to be effective.

3 Weighted Bias Metrics

In this chapter, two approaches for measuring bias are introduced: the weighted bias metric and the extended weighted bias metric. First, the equations themselves are explained, with an explanation and justification of the metrics following after.

3.1 Equations

For readability, the bias metric equations from section 1.1 have been renamed such that they are easier to utilize in the new metrics:

Baseline Performance (BP):

$$BP = x \in [\text{Base}_{\text{min}}, \text{Base}_{\text{norm}}] \tag{5}$$

such that the baseline performance reference can be chosen from either of the known options: either the minimum WER observed or a norm speaker group.

Performance Difference (PD):

$$PD_i = \text{Base}_i - BP \tag{6}$$

The new metrics mentioned in section 2.1, WPB and IWBP, are defined as follows:

Weighted Performance Bias (WPB):

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

Performance Difference (PD) between two speaker groups:

$$PD_{ij} = \text{Base}_i - \text{Base}_j \tag{8}$$

where Base_i and Base_j are the base performance for group i and group j.

Intergroup Weighted Performance Bias (IWPB) between each speaker group:

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

where PD_{ij} is the performance difference between groups i and j.

3.2 Explanation and Justification

The proposed bias metrics were designed to incorporate both the performance differences between demographic groups and the baseline performance of a separate group. First, the base performance metric is calculated per speaker group. It serves as a benchmark to compare the performance of the ASR system across different demographic groups. For the new metrics, the relative difference is used from these metrics (as can be seen from equations 7 and 9), since Patel et al. recommend 'using a relative measure that considers all

[2] https://www.tudelft.nl/dhpc/system

speaker groups' [14], such that it represents the performance difference between the speaker groups relative to the baseline performance.

The base metrics per speaker group have already been previously defined, but now need to be included. A weighted average of the two components of equation 7 (base performance and performance difference) has been taken to combine their aspects so that different weights can be tested to find an optimal value for the data.

Two options for a new bias metric were experimented with: the weighted bias metric and the intergroup weighted bias metric. Both options require weights w_1 and w_2 to determine the ratio of which the base performance is deemed more important than the performance differences or vice versa. The main difference between the two metrics is that WPB uses a defined baseline performance to compare individual groups, while IWPB averages over comparing the current group with every other group. This distinction was used to test what type of comparison performed best.

4 Experimental Setup and Results

4.1 Experimental Setup

The objective of this experiment was to evaluate the proposed bias metrics, WPB and IWPB, in the context of ASR systems. Specifically, the experiment aims to determine how well the metrics measure bias across different demographic groups while combining performance difference and base performance. The data used in this experiment is derived from the JASMIN corpus, after being used in the models provided in [14].

The data is separated by model type, leading to the following categories: $NoAug$, $SpAug$, $SpSpecAug$, $Whisper$ and $FTWpr$. These directories are then split into both Read and HMI speech categories. Finally, for every speaker group, there is a separate output file containing the following required data for this research: the number of words, correct words, substitutions, deletions and insertions.

The performance differences are calculated after reading the data. Specifically for the new metrics, simulations are run with a uniform distribution of the weights with a size of 100, retrieving the optimal weights for the metrics.

To ensure reproducibility, all code and functions used for processing data, visualising data and calculating metrics have been provided in the project repository.

4.2 Results

In this section, the outcomes of evaluating the proposed WPB and IWPB metrics are presented across the various models and speaker groups. The values of WPB and IWPB were compared with the known bias values for $G2_{m,a}$, $G2_{n,a}$, $G2_{m,r}$ and $G2_{n,r}$. The simulations of WBP and IWBP metrics to find optimal weights are discussed, as well as the bias values for each model. This analysis demonstrates the potential effectiveness of the new metrics in capturing performance disparities.

Bias Metrics Results

Optimal weights had to be found in order to properly evaluate the bias metrics. Thus, a simulation was done by attempting 100 different weights distributed uniformly between [0,1]. Figures 1 and 2 show plots of these simulations for each model.

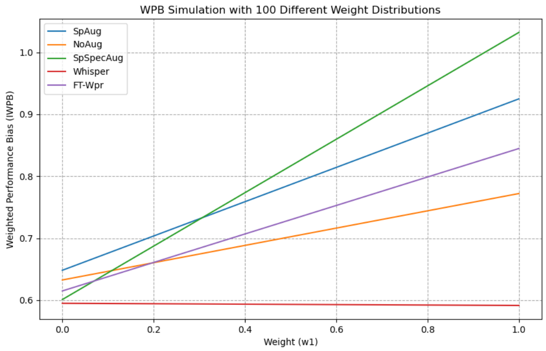

Figure 1: Weighted Performance Bias Metric simulation.

The graph in figure 1 shows WPB across five ASR models ($SpAug$, $NoAug$, $SpSpecAug$, $Whisper$, $FTWpr$) for 100 different weight distributions. It can be seen that WPB values increase with the weight (w_1) for all models. Whisper consistently shows the lowest WPB across all weight distributions, indicating minimal bias of the models. SpSpecAug exhibits the highest WPB, particularly as w_1 approaches 1, indicating significant bias. NoAug, SpAug, and FT-Wpr have intermediate WPB values, with NoAug being the lowest among them. WPB is lowest when w_1 is 0.0.

The graph in figure 2 presents the IWPB for the five ASR models, again across 100 different weight distributions. IWPB values increase with the weight (w_1) for all models, similarly to WPB. Whisper once again exhibits the lowest IWPB, indicating reduced bias. SpSpecAug shows the highest IWPB as w_1 increases, highlighting bias, with similar values for FT-Wpr. The models SpAug and NoAug fall between these two extremes, with NoAug and FT-Wpr displaying similar trends. At a w_1 value of 0.0, all models produce a comparably low amount of bias.

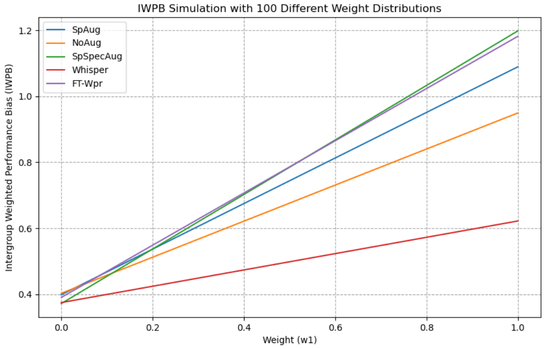

Figure 2: Intergroup Weighted Performance Bias Metric simulation.

Figure 3 shows the WPB metric values after applying the weights found during the simulation in figure 1: $w_1 = 0.0$ and $w_2 = 1.0$. It can be seen that the bias values remain relatively high for the non-native speakers in comparison to the native speakers, agreeing with the results of Patel et al [14]. Bias is lowest for the DT group. For WPB, $SpAug$ and $SpSpecAug$ generally produce a similar amount of bias in comparison to $NoAug$. When looking at the Whisper output in particular, $FTWpr$ produces slightly less bias than $Whisper$ for native speakers, with the opposite holding true for non-native speakers.

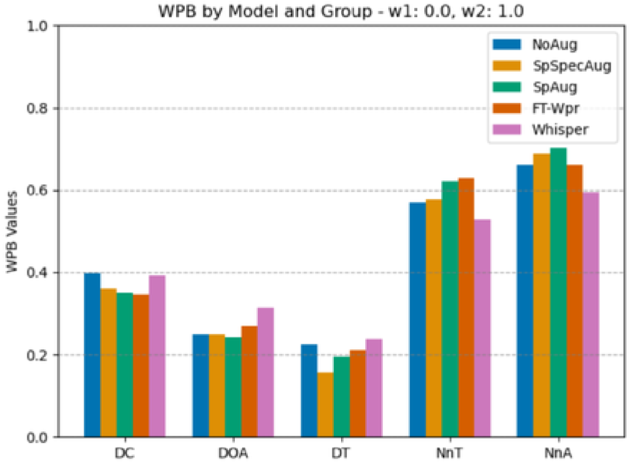

Figure 3: Weighted Performance Bias Metric.

Figure 4 shows the IWPB metric, also with weights from the simulation in figure 2: $w_1 = 0.0$ and $w_2 = 1.0$. Similarly to WPB, DC, DOA and DT have the lowest bias values, indicating that native speech might produce less bias.

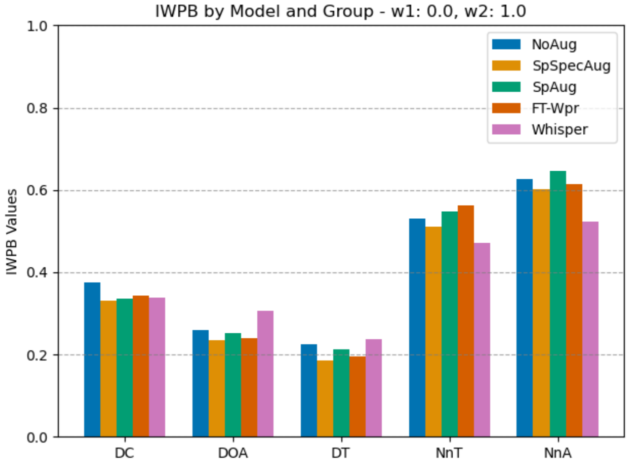

Figure 4: Intergroup Weighted Performance Bias Metric.

When comparing figures 3 and 4, it is evident that both metrics made use of a high w_2 value (1.0), with a low w_1 value (0.0), which will be further discussed in section 5. The shapes of both plots are very similar, with IWPB producing slightly less bias in general. The general performance of the metrics in identifying bias is alike.

Table 1: WPM and IWPM Bias Values per Speaker Group

Metric	Group	NoAug	SpAug	SpSpecAug	Whisper	Ft-Wpr
WPB	DC	37.6	33.7	33.0	**33.9**	34.3
	DT	22.6	21.3	18.5	23.7	19.5
	DOA	25.9	25.2	23.4	30.6	24.0
	NnA	62.7	64.5	60.1	52.3	61.3
	NnT	52.9	54.8	51.1	**47.1**	56.1
IWPB	DC	**24.4**	26.8	**33.9**	**21.1**	26.9
	DT	15.9	18.3	20.7	14.7	17.2
	DOA	17.5	20.9	26.0	17.9	20.4
	NnA	**77.2**	92.5	**103.0**	**59.1**	84.5
	NnT	57.4	73.5	77.4	45.4	76.1

For the WPB values in table 1, Whisper has a relatively lower IWPB compared to SpSpecAug and SpAug, for DC and NnT. NoAug and Ft-Wpr show similar IWPB values, with NoAug being slightly higher in some cases. SpAug has consistently higher values than Whisper but lower than SpSpecAug, indicating moderate bias levels.

For the IWPB values in table 1, SpSpecAug consistently shows the highest bias values across all groups, with notable values for DC and NnA. In contrast, Whisper has the lowest WPB values for all metrics, indicating the least bias. SpAug and Ft-Wpr show moderate bias values, with Ft-Wpr having slightly higher bias than SpAug in most cases. NoAug exhibits the least bias variability, suggesting it maintains a relatively consistent bias level across different metrics.

Table 2: Overall Difference Bias per model for different bias metrics.

ASR Model	**JASMIN Read Speech**						**JASMIN HMI**					
	$G2_{m,a}$	$G2_{n,a}$	$G2_{m,r}$	$G2_{n,r}$	***WPB***	***IWPB***	$G2_{m,a}$	$G2_{n,a}$	$G2_{m,r}$	$G2_{n,r}$	***WPB***	***IWPB***
NoAug	23.18	30.54	1.08	3.21	47.96	40.33	13.20	25.46	0.34	1.08	46.54	51.99
SpAug	23.18	31.70	1.17	4.59	49.38	39.90	16.13	24.00	0.49	1.10	47.57	50.40
SpSpecAug	23.55	30.34	1.28	4.40	49.98	37.22	18.73	21.48	0.71	1.08	46.04	45.38
Whisper	21.08	25.26	0.83	1.48	51.17	37.51	18.85	1.58	0.46	0.03	45.76	54.97
FT-Wpr	24.50	24.90	1.10	4.38	41.67	39.07	13.73	25.88	0.37	1.14	49.55	51.41

Table 2 shows the bias values for each metric used in the paper, divided among Read and HMI speech. The values for Read speech are larger than for HMI for the $G2_{m,a}$, $G2_{n,a}$, $G2_{m,r}$ and $G2_{n,r}$ metrics, while the opposite is true for IWBP. WBP shows higher bias values for HMI $FTWpr$ speech, while the rest of the models for this metric follow the same trend as the G2 metrics. The IWPB metric produces lower values for read speech than HMI speech, while WPB produces similar values regardless of the speech type.

In general, IWPB has slightly lower values most of the time, but the general distribution of the bias is similar between WPB and IWPB. This will be further discussed in section 5.

5 Discussion

In section 5.1, it will be explained why the weights might not be as optimal as they seem, but for the discussion of these results, it is assumed that they are optimal. As mentioned in section 4.2, simulations of weights for both WPB and IWPB led to optimal weights of w_1 = 0.0 and w_2 = 1.0. A high w_1 value means that the performance difference aspect of both WPB and IWPB is deemed more important, with a value of 1.0 indicating that the final value is 100% dependent on this part of the equation. The same holds for w_2, with the optimal value being 1.0 for w_2, showing that only the base performance component is fully used, while the performance difference is not used at all. Although the weight distribution is not uniform and the performance difference is not properly incorporated, the outputted values in figures 3 and 4 show a similar distribution of the bias in comparison to [14]. In particular, these weights lead to a high bias for non-natives, while having relatively lower bias values for the native Dutch speakers, with DT showing the least bias. This is logical since this group 'has the closest acoustic match to the training data' [14].

The results show that when comparing the tables and the figures, the WPB and IWPB values differ. The main reason for this is because the tables make use of a speech type split of Read vs HMI data. The figures, on the other hand, combine both speech types as one, since the figures are more focused on the overall picture of bias between the speaker groups. This likely led to disparities between these values, but both interpretations of the data can be used to draw different conclusions.

When comparing the WPB and IWPB values, the amount of recognised bias is alike. Both metrics measure consistently high bias for $SpSpecAug$, suggesting that the augmentation method may introduce bias, particularly for certain groups like NnA. Regardless of whether or not the amount of bias is correct (which will be further discussed in section 5.1), the results show that the ability to measure bias is similar for both metrics. Table 1 indicates that WBP and IWBP both measure more bias in relation to the non-native speakers than the native speakers, which is to be expected according to Feng et al. [3]. This table also reveals that at speaker group level, regardless of speech type, the $Whisper$ and $FTWpr$ models originating from OpenAI have around as much bias as the other models. Table 2 shows that when comparing the WPB and IWPB values with the $G2$ variants, although there is a resemblance between the distance between groups, the absolute values are much higher, in some cases double that of $G2$. This suggests that should the new metrics be correct, more bias is present in the speech models than previously shown by [14].

5.1 Limitations

Although the proposed metrics seem to perform relatively similarly to Patel et al.'s metrics [14], the fact that the value of w_2 is 1.0 means that the performance difference itself goes against the entire idea of combining both metrics. This likely means that the determination of best weights needs to be updated such that the parts are both properly contributing to the overall output. However, in the case that the weights indeed are optional, it is clear that the edge cases of one weight being 1.0 should be included. If the part of the equation being affected by w_2 also incorporated the performance difference, then the situation of $w_2 = 1.0$ would be less of a problem, as the performance difference would still be taken into account.

The main reason that it is difficult to determine which weights are optimal, is the lack of a ground truth table. This means there is no way of knowing whether an outputted bias value is correct. The current implementation chooses the weights for which the bias values are minimal, but arguments can be made for choosing other bias values. In truth, there is not currently a simple answer for determining *true* amounts of bias, since there is no predetermined known bias. Thus, this paper compared the metrics with the values from [14] and attempted to find a similar distribution of bias among the groups. With figures 3 and 4 showing similar bias for specific groups, this implementation can indeed recognise apparent bias. However, the true limitation of the system is that the apparent bias is not definite due to a lack of control values for bias.

5.2 Future Work

Recommendations for further research include searching for a better way to combine performance difference and base performance. This paper shows that the proposed metrics are not incapable of measuring bias, however, work still needs to be done to optimise the metrics. For this research, a weighted average was chosen to answer the main research question, yet more possible solutions exist. Future studies should look into weighted averages with the absolute performance difference, rather than the relative performance difference in WPB. Also, other combinations or aggregations without using weighted averages could exist, although that was out of the scope of this project.

As mentioned in 5.1, including the performance difference in the base performance affected by w_2 could lead to the incorporation of performance differences in edge cases.

The main suggestion would be to search for a new method of weight selection, such that the optimal weights truly are the best choices. Although the absence of a ground truth will pose a problem for future research, investing time into experimenting with which bias values work better than others could improve formulating proper base performance to performance difference ratios. Better argumentation for why specific weights perform better than others would greatly increase the effectiveness of the research.

6 Responsible Research

In this chapter, the various aspects of responsible research that underpin this study are discussed. The following subsections cover the key components of ensuring the reproducibility, accessibility, and ethical considerations of the research process.

6.1 Reproducibility

Ensuring reproducibility is a crucial part of responsible research. To this end, all code and scripts developed for this study have been documented and will be made publicly available via the TU Delft Repository. This transparency allows other researchers to replicate the experiments, verify results, and build upon the work presented. By sharing the data processing and analysis pipelines, the aim is to contribute to a robust and reproducible research culture. Given the correct input data, the code should be capable of producing the results provided in this paper. This commitment to reproducibility enhances the reliability and impact of the research.

6.2 User-Friendly Figures

Accessibility in research is critical, particularly when presenting data. In creating visualisations, special attention will be paid to the selection of colours to ensure they are user-friendly. This includes using colour palettes that are accessible to individuals with colour blindness, ensuring that all figures are easily interpretable by a broad audience.

6.3 Ethics in Speech Technology

The primary focus of this project is to address and measure bias in ASR systems. This research is rooted in the ethical imperative to improve technology for all users, especially those currently underrepresented. Bias in ASR can lead to significant disparities in how individuals are understood by these systems, affecting older individuals, non-native speakers, and others with diverse accents.

By developing and refining metrics to measure this bias, the project aimed to highlight and quantify disparities in ASR performance. The ultimate goal was to inform and influence the development of more equitable and inclusive ASR technologies. Ethically, this work is crucial as it strives to ensure that advancements in speech technology do not disproportionately benefit certain groups while disadvantaging others. By bringing these issues to the forefront, the research contributes to the broader effort of creating fair and just technology.

Thus, the ethical dimension of this research is not only about identifying bias but also about advocating for solutions that promote fairness and inclusivity in speech technology. Addressing these biases can help ensure that ASR systems serve all users effectively, regardless of their background or characteristics. This commitment to ethical research practices underpins the project and aligns with the broader goals of promoting equity and representation in technology.

7 Conclusions

This paper took the first steps towards integrating base performance and performance difference in an Automatic Speech Recognition (ASR) bias metric. It was shown that the proposed Weighted Performance Bias (WPB) and Intergroup Weighted Performance Bias (IWBP) metrics not only produced relatively similar values to each other, but the trends found by Patel et al. [14] also hold for these new metrics, which indicate that the metrics are indeed capable of measuring bias in an improved way. In general, the non-native speakers still give rise to the most bias, while the Dutch Teenager speaker group shows the least bias. This shows room for improvement in the ASR systems.

Although the absence of a ground truth for bias continues to pose a problem in bias measurement innovation for ASR, experiments like the one conducted in this paper are imperative to continuing the development of bias evaluation in this area. Future research should focus on optimizing the weight selection for WPB and IWPB and exploring alternative options for bias evaluation. This way, the development of novel methods to recognise bias continues to promote fairness and inclusivity in ASR technologies.

A Performance Metric Analysis

During this research, while conducting experiments it gradually became clear that the evaluation and comparison of different performance metrics deviated too much from the main topic of this paper, namely combining performance difference and base performance in a single bias metric. It was chosen to halt the experiment in order to focus on the main subject, yet there are results to be considered. In this section, the preliminary results of the performance metric experiment will be discussed.

A.1 Performance Metric Experiment

In order to quantify potential bias, first the performance of the ASR system must be measured. The method in which this is done can differ since the choice of performance metric can influence the measured performance. ASR research often relies on standard performance metrics like Word Error Rate (WER), Character Error Rate (CER) and Phoneme Error Rate (PER) [14]. These three metrics all measure the percentage of errors found, each at a different level: WER measures at word level, CER measures at character level and PER at phoneme level. Therefore, these different metrics look at different aspects of transcription, from single recognised characters to a set of specific speech sounds perceived (phonemes). However, these metrics may not capture the full picture of ASR performance and may overlook disparities [8]. Searching for a novel solution to evaluate these systems could lead to a more accurate bias calculation.

The used metrics are calculated based on the concept of Levenshtein distance, which indicates how different two given strings are from one another [19]. It involves three types of errors: substitutions, insertions, and deletions. Substitutions occur when one element (character, phoneme, or word) is replaced by another. Insertions happen when extra elements are added to the ASR output that do not exist in the reference text. Deletions occur when elements present in the reference text are missing in the ASR output. By measuring these errors, WER, CER and PER can be calculated with the following equation:

$$ErrorRate = \frac{S+I+D}{Total} \cdot 100\%$$

where S, I and D is respectively the total number of substitutions, insertions and deletions. $Total$ differs per error rate:

- For **WER**: $Total =$ Total words in reference
- For **CER**: $Total =$ Total characters in reference
- For **PER**: $Total =$ Total phonemes in reference

To comprehensively evaluate the performance of ASR systems, an experiment can be conducted to compare the aforementioned metrics. Since this experiment was not the main concern of the research, a smaller experiment was conducted by incorporating the most standard performance metric, WER, as well as another, less commonly used metric: Match Error Rate (MER).

Match Error Rate (MER): MER is the proportion of incorrect word matches, which is the same as the probability of a given match being incorrect [5]. It can be defined as:

$$MER = \frac{S+I+D}{S+I+D+H} \cdot 100\%$$

A.2 Preliminary Results

The comparison of WER and MER statistics across different models and speaker groups shows the behaviour of these metrics concerning the performance of the ASR systems. The figures 5 and 6 show a breakdown of the median, standard deviation, maximum, and minimum values for every group within each model.

The comparison of WER and MER values across models and speaker groups reveals significant trends in ASR performance. Both metrics show the highest median error rates for the NnA group across all models, indicating substantial difficulty in recognising non-native speech. The median MER values are generally lower than WER, suggesting better performance in exact match scenarios, yet the similarity between WER and MER values across groups and models implies that MER may not offer significant additional insights over WER in this context.

Standard deviation values for both WER and MER are low across models, reflecting consistent performance within each group. However, the NnA group exhibits slightly higher variability, once more highlighting the challenges in accurately recognising non-native accents. The range of values (difference between max and min) shows that the NnA group consistently has the highest maximum error rates, nearing 1.0 for WER in some cases, indicating poor recognition performance. While the maximum MER values are also high, they are generally lower than the WER maximum values, suggesting slightly better performance regarding character and phoneme recognition. The low minimum values for both WER and MER across all groups suggest that ASR systems can achieve low error rates under optimal conditions.

In conclusion, the comparison of WER and MER metrics showed very similar values between WER and MER, which suggests that although MER had slightly lower values in general, it might not provide substantial additional benefits over WER in this context. Both metrics appear to recognise similar performance characteristics, similar to the research by Hamed et al. [8]. In this study, while MER is a valuable metric, it may not necessarily offer enough beneficial insights compared to WER to warrant choosing MER over the more thoroughly researched standard performance metric, WER.

A.3 Future Research

In future research, a full comparison of all mentioned metrics might lead to more conclusive arguments for one metric over another. With more time and metrics to work with, a more extensive analysis could prove useful. It could also prove interesting to look into other performance metrics not previously mentioned in this research, such as Word Information Lost (WIL) [13].

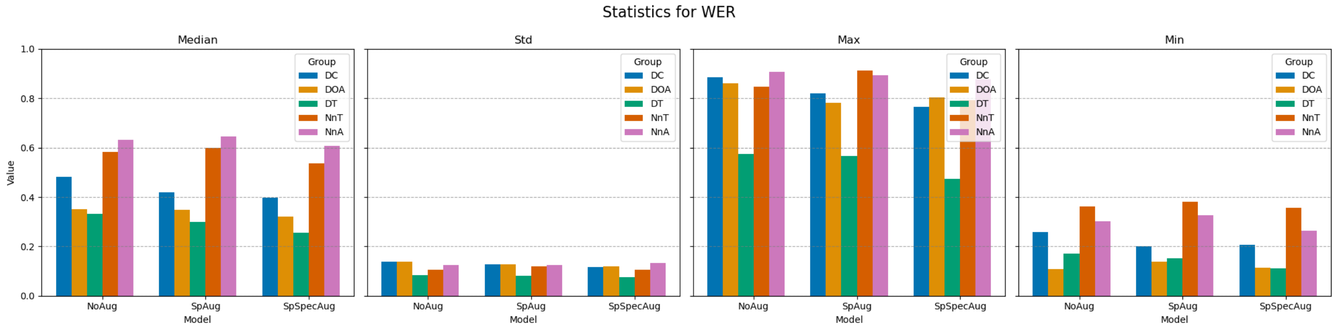

Figure 5: Histogram plot with statistics for WER.

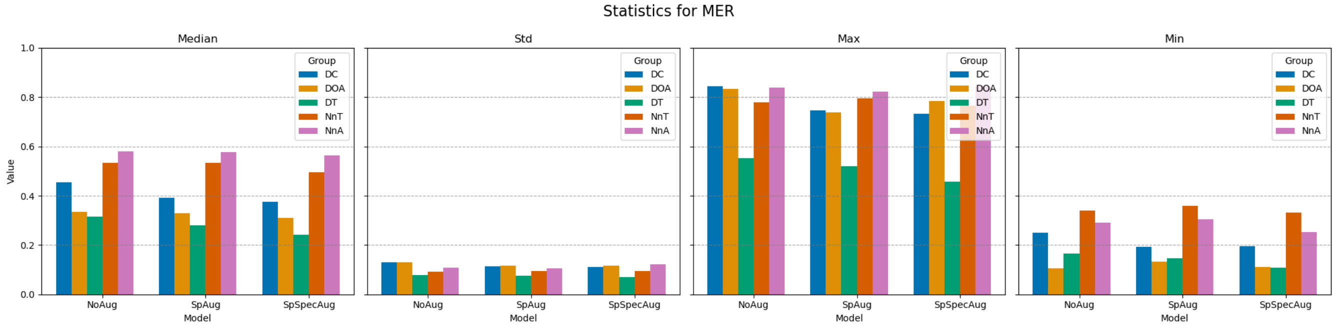

Figure 6: Histogram plot with statistics for MER.

References

[1] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 429–440, 2021.

[2] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix Smits. JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

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

[4] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*, 2021.

[5] F. Filippidou and L. Moussiades. A benchmarking of ibm, google and wit automatic speech recognition systems. In *Artificial Intelligence Applications and Innovations*, volume 583, pages 73–82. Springer, 2020.

[6] Wiqas Ghai and Navdeep Singh. Literature review on automatic speech recognition. *International Journal of Computer Applications*, 41(8), 2012.

[7] Calbert Graham and Nathan Roll. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2024.

[8] Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. Benchmarking evaluation metrics for code-switching automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 999–1005. IEEE, 2023.

[9] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[10] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[11] Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 230–247, New York, NY, USA, 2022. Association for Computing Machinery.

[12] S Karpagavalli and Edy Chandra. A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404, 2016.

[13] Andrew Morris. An information theoretic measure of sequence recognition performance. 01 2002.

[14] 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.

[15] Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton van der Wouden. CGN, an annotated corpus of spoken Dutch. In *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*, 2003.

[16] Hamid Tabani, Jose-Maria Arnau, Jordi Tubella, and Antonio González. Performance analysis and optimization of automatic speech recognition. *IEEE Transactions on Multi-Scale Computing Systems*, 4(4):847–860, 2018.

[17] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[18] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[19] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007.