



**How to measure bias in automatic speech recognition system?
A bias metric without a reference group**

Tereza Ležovičová
Supervisor(s): Odette Scharenborg, Jorge Martinez Castaneda
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: Tereza Ležovičová
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Jorge Martinez Castaneda, Merve Gürel

Abstract

This paper presents a novel approach to measuring bias in Automatic Speech Recognition (ASR) systems by proposing a metric that does not use the conventional approach of a reference group. Current methods typically measure bias through comparison with a 'norm' or minimum error group, potentially introducing additional biases. To address this issue, this study introduces a new metric: a combination of the Group-to-Average Log Ratio and the Sum of Group Error Differences. This metric aims to provide a fair comparison by measuring performance relative to the average of the groups rather than a single reference group. Results indicate that the new metric reveals different aspects of bias not captured by traditional methods.

This study contributes to the ongoing research on fairness in speech technology by challenging the existing bias metrics and proposing alternatives that might offer more equitable evaluations. Future research should explore further refinements of these metrics, and apply them across more varied datasets and environments. Ultimately, this research moves towards making ASR technologies more inclusive, ensuring that they serve all user groups equitably.

Index Terms: speech recognition, bias metrics, reference group

1. Introduction

As speech technology becomes part of our daily routine, so does the importance of ensuring the safety and fairness of these systems. 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 [1] is challenging. Research shows that ASR models show bias, e.g., between genders and races [2].

Firstly, it is important to define the notions of bias and fairness addressed in this work. This research concerns group fairness, which involves accepting similar outcomes across different speaker (demographic) groups [3]. Furthermore, bias is understood as a performance difference, specifically an error rate difference, between various speaker groups, as defined in prior literature [4]. In ASR the bias is estimated by comparing the automatic ASR output (hypothesis) to the reference transcriptions done by human annotators. However, these human annotations can also contain errors and bias so they are also part of the overall bias in ASR systems [5].

Therefore I am working under the assumption that for the ASR system to be unbiased it is needed to achieve statistical parity - a speaker group is treated the same way as the population as a whole [6]. This is achieved by having a similar error rate across all speaker groups.

Effective and fair bias metrics are needed to discover these biases. In [7], these are calculated from statistical base metrics, like error rate or accuracy, therefore our choice of a base metric also influences the end value. The base metric (b) used throughout this paper will be the word error rate (WER). Currently, there is a lack of consensus on which is the most favorable bias metric in ASR. As research in the fairness of ASR systems is still in its early stages [5] also finding an effective bias metric is still an area of ongoing research.

1.1. Problem Description

Existing bias metrics typically use a "reference group". An example of a bias metric with a reference group is in Eq 1, where

bias represents the difference between the WER of the chosen speaker group and the WER of the reference group.

$$G2_{\text{reference group diff}} = b_{\text{spk}_g} - b_{\text{reference group}} \quad (1)$$

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. Choosing a reference group is not trivial. For instance, two approaches are mentioned in [8]: (1) taking the group with minimum WER, and (2) using a norm group as a reference group. Since the former approach might not always be available, the authors chose the second option. This method, however, has several limitations. This approach assumes that there are two types of groups: the privileged one and the non-privileged one [9]. This results in binary outcomes that fail to capture the nuances of multi-group biases. For example, the norm group chosen by [4] consists of only adult native speakers, whereas within the non-native group, there could be significant variations between speakers from different linguistic backgrounds. The second problem is ethical concern regarding the standardization of what constitutes a "normal" speech characteristic, e.g. commonly accepted accents. This standardization can lead to the inappropriate classification of natural speech variations (like in regional accents) as errors, thereby reinforcing biases. Therefore, it is important to not just focus on lowering the error rates of the test sets but to comprehensively evaluate the fairness of the system [10].

1.2. Aim

This paper proposes an approach to measuring bias in ASR systems that eliminates the use of a reference group. I analyzed the suitability of various bias metrics without reference groups for calculating bias in ASR systems and used the conclusions to develop my bias metric. The bias metric analysis in [4] was used as a baseline. Here, the authors evaluated four existing bias metrics (also with a reference group) on end-to-end (E2E) ASR models (Conformer and Whisper). The developed bias metric was applied on the same models and the results were compared to those reported in [4]. I conducted both a qualitative analysis, in the form of assumptions from existing literature, and a quantitative analysis, in the form of results of my experiment, of the bias metric. This research not only challenges existing metrics but also contributes to the broader discussion on the fairness of ASR technologies.

Firstly, in Section 2 Methodology I will explain my approach to implementing a bias metric and evaluating the ASR model on it. I will explain how I set up my experiment and what models and datasets I am working with to ensure my implementation can be fairly compared to results of the existing bias metrics from [4]. Next, in Section 3 The Bias Metric, I will provide a qualitative analysis of the chosen existing bias metrics, discussing the assumptions made about them in the literature. In this section, I will also show the final implementation of the bias metric. In Section 4 Results, I will give an initial explanation of my results of the bias metric, meaning how it performs on the ASR models, and how it compares to bias metrics that use a reference group, and in Section 6 Discussion, I will analyze how these results reflect my assumptions made in the Section 3. In the Responsible Research Section 5, I will explain why my research conforms to responsible academic norms, and that it is reproducible and inclusive. I will close with a Conclusion and Future Recommendation Section 7 where I will state the future recommendations for my research that can enhance its quality.

2. Methodology

In this section, I explain my research by introducing all parts in more detail and all terms. I also describe the experimental setup I conducted.

2.1. Experimental Setup

For the experimental setup I am using the same approach as in [4] to ensure the same baseline. I evaluate my bias metric on two models: the E2E Conformer model [11] and Open AI’s Whisper model [12]. For both models, I evaluate the bias before and after potential augmentation or fine-tuning to determine the impact of these techniques on the fairness of the ASR systems.

The Conformer model I am evaluating with my metric was trained on Corpus Gesproken Nederlands (CGN) [13]. This dataset was trained on adult native Dutch speech, with both Read speech (Rd) and Conversational Telephone Speech (CTS). In my experiment, I will use the Jasmin corpus [14]. Jasmin consists of Read (Rd) Speech and Human Machine Interaction (HMI). Speakers in Jasmin are categorized into multiple groups by age: children (C), teenagers (T), older adults (OA), and by nationality: dutch (D), non-native (Nn). Given that CGN serves as the training dataset for the Conformer model and Jasmin contains data from different demographic groups, distinct from those in CGN, makes Jasmin suitable for our testing purposes.

2.2. Bias metrics

There are 4 bias metrics evaluated in [4].

Group-to-min calculates the difference (Eq 2) and the relative difference (Eq 3) between the WER of the specific speaker group under evaluation spk_g and the group with minimum WER.

$$G2_{\min, \text{diff}} = b_{spk_g} - b_{\min} \quad (2)$$

$$G2_{\min, \text{reldiff}} = (b_{spk_g} - b_{\min})/b_{\min} \quad (3)$$

Next, Group-to-norm calculates the difference (Eq 4) and the relative difference (Eq 5) between the WER of the specific speaker group under evaluation spk_g and the reference (norm) group, which in this case is the CGN test set [4]. So these two methods (minimum WER group and norm group) correspond to the two approaches of choosing a reference group mentioned in [1]. In Eq 2 and 3 we subtract from the WER of the chosen speaker group the minimal WER from the set and in Eq 4 and 5 the WER of the norm group.

$$G2_{\text{norm}, \text{diff}} = b_{spk_g} - b_{\text{norm}} \quad (4)$$

$$G2_{\text{norm}, \text{reldiff}} = (b_{spk_g} - b_{\text{norm}})/b_v \quad (5)$$

3. The Bias Metric

In this section, I explain how I arrived at my implementation of a bias metric and what assumptions I considered in the process.

3.1. Evaluation of Multiple Bias Metrics

As both ASR systems and bias detection are relatively new fields, there is no state-of-the-art bias metric for ASR systems. Some groups estimate bias by taking a difference/ratio between the speaker groups’ WERs [1, 15, 16], or taking the relative

difference between the best-performing group and the worst-performing group [5] or others use the bias metrics described above ($G2_{\min}$ and $G2_{\text{norm}}$) [4].

Therefore, as there are not many bias metrics in the ASR field, I researched other close areas. Firstly, in the field of speaker verification [7] proposes the use of a Group-to-average log ratio (Eq 6)

$$G2_{\text{avg log ratio}}(b)_g = -\ln \frac{b_g}{b_{\text{avg}}} \quad (6)$$

Conceptually, this metric measures how much better or worse a particular group is performing relative to the average performance of all groups. In the context of ASR, this could mean comparing the WERs of different demographic groups to the average WER.

The paper [7] recommends ratio-based bias metrics as difference-based ones like $G2_{\min}$ “cannot be compared across base metrics with different orders of magnitude, and lack sensitivity when base metrics are small” [7][p. 4]. They also recommend the use of a log ratio as by being centered around 0 it is easier to use as a comparison of values across multiple groups.

The second field in which I researched bias metrics was biometric verification. The paper [17] introduces a new bias metric Sum of Group Error Differences (Eq 9).

$$\delta\text{FMR}_g = \left| 1 - \frac{\text{FMR}_g(T)}{\text{FMR}_{\text{global}}(T)} \right| \quad (7)$$

$$\delta\text{FNMR}_g = \left| 1 - \frac{\text{FNMR}_g(T)}{\text{FNMR}_{\text{global}}(T)} \right| \quad (8)$$

$$\text{SED}_g = \delta\text{FMR}_g + \delta\text{FNMR}_g \quad (9)$$

This metric aggregates the differences in error rates for each group compared to the global error rate, giving a comprehensive measure of bias. For ASR, this could involve summing the differences in WERs, as WER measures the overall transcription errors, for various groups to understand the overall bias.

In Eq 7 and 8, FMR (False Match Rate) and FNMR (False Non-Match Rate) represent the proportions of incorrect acceptances and rejections, respectively, within a biometric system. These rates are computed at a specific threshold T , which defines whether a biometric sample is considered a match. The metric calculates the absolute value of the deviation of a demographic group’s error rates ($\delta\text{FMR}_g, \delta\text{FNMR}_g$) from the global averages, emphasizing that any significant deviation—whether positive or negative—is indicative of bias. The final bias value, SED_g , is the sum of these deviations, providing a measure of overall bias for each group. This concept of threshold-based error rate comparison, while standard in biometric verification, requires adaptation for use in ASR systems, where such thresholds do not typically apply.

3.2. Assumptions

In this subsection, I introduce multiple assumptions from prior literature under which I designed my bias metric.

Assumption 1. *For a model to be considered fair between two demographic groups, measuring fairness based on the minimum ratio is better than measuring based on their absolute difference.*

According to [3] there are two common ways to measure the extent of the bias between two groups. Either by taking the absolute difference between the two demographic groups

Table 1: Results in %WER for Conformer and Whisper models when tested on Jasmin.

Model	Approach	Jasmin: Read					Jasmin: HMI					Jasmin: Average		
		DC	DT	NnT	NnA	DOA	DC	DT	NnT	NnA	DOA	Read	HMI	All
Conformer	NoAug	44.55	23.91	57.01	62.95	29.65	52.03	41.74	60.79	63.21	43.52	43.61	52.26	47.94
	SpAug	38.57	22.58	58.46	65.14	29.01	45.60	37.24	61.89	64.10	42.80	42.75	50.33	46.54
	SpSpecAug	38.02	20.66	54.28	62.29	27.47	43.23	31.27	54.49	58.89	40.13	40.54	45.6	43.07
Whisper	W_s	40.30	25.51	53.84	58.14	34.13	54.48	40.64	59.32	73.09	50.93	42.38	55.69	49.04
	$W_s FT_{cgn}$	40.90	22.38	57.73	60.79	28.23	43.45	37.62	59.62	58.32	43.89	42.0	48.58	45.29

($|b_{spk_1} - b_{spk_2}|$) or by taking the minimum ratio between the two groups ($\min\{b_{spk_1}/b_{spk_2}, b_{spk_2}/b_{spk_1}\}$). However, according to their research, it is better in most cases to use the ratio approach as when a model is fair under the ratio approach it will also be fair under the difference approach, but not the other way around.

Assumption 2. *Measuring the absolute value of bias of the speaker group ($|b_{spk}|$) provides a clearer comparison of values of bias between speaker groups.*

In certain contexts, it is important not only to identify where performance falls below a global reference but also where it exceeds this benchmark. This approach conveys that any significant deviation from the global reference, whether positive or negative, constitutes a form of bias. Thus, the use of absolute values in measuring discrepancies is emphasized in [17].

Assumption 3. *Using functions such as the logarithm function in Eq 6 and subtracting 1 in Eq 9 to center bias measurements at 0 improves the clarity of comparing bias across different speaker groups. This approach divides the positively biased groups (bias above 0) and negatively biased groups (bias below 0).*

In [7] it is mentioned that applying a logarithm function (Eq 6) makes the bias centered around 0 - an unbiased speaker group has a bias of 0 - which makes it easier to compare across multiple groups as people find it easier to differentiate between negative values (negatively biased group) and positive values (positively biased group). The same effect also makes subtracting 1 from the bias in Eq 9, as with ratio, the neutral bias is at value 1, subtracting makes it centered again at 0.

Assumption 4. *By having both across-group and across-model comparisons we can have a more comprehensive and informed assessment of bias.*

Looking at the comparison of performance against groups is important to find out which groups are being disadvantaged but it is also important for a bias assessment to be comprehensive to also look at summative-aggregation measures [17]. This will come to place when we want to compare models' overall bias.

Assumption 5. *The approach of not utilizing a reference group in a bias metric avoids the ethical and methodological issues associated with selecting an appropriate reference group, which may inadvertently perpetuate biases.*

Lastly, following the problem description, a bias metric that does not utilize a reference group makes the performance assessment between the speaker groups more fair, as no group gets more advantage in the process. Furthermore, a bias metric without a reference group does not have the disadvantage of not being able to calculate the bias of one of the speaker groups [8].

3.3. Implementation of My Bias Metric

In my implementation, I first calculate the average WER of every group (g) per model and speaking style (MS). Then I calculate the Group-to-Average Log Ratio $G2_{\text{avg log ratio}}(g)$ to measure the bias of each demographic group (Eq 12), where $b_{\text{avg}(MS)}$ means the average WER per model and speaking style and $b_{g(MS)}$ is WER of demographic group g per model and speaking style. G represents the set of all (demographic) groups. The code for this metric can be found in Figure 6 in Appendix A.

$$b_{\text{avg}(MS)} = \frac{1}{G} \sum_{g \in G} \text{WER}_{g(MS)} \quad (10)$$

$$b_{g(MS)} = \text{WER}_{g(MS)} \quad (11)$$

$$G2_{\text{avg log ratio}}(g(MS)) = -\ln \frac{b_{g(MS)}}{b_{\text{avg}(MS)}} \quad (12)$$

To implement a bias metric inspired by the Sum of Groups Error from [17], I used the calculation of the ratio of $b_{g(MS)}$ (Eq 10) and $b_{\text{avg}(MS)}$ (Eq 11) and at first subtracted it from 1 and then took absolute value from the result (Eq 13). This code implementation is shown in Figure 7 in Appendix A.

$$SED_{g(MS)} = \left| 1 - \frac{b_{g(MS)}}{b_{\text{avg}(MS)}} \right| \quad (13)$$

Then, I adapt the concept of the Sum of Group Error Differences to quantify the total bias by summing the biases across all groups per model. This combined metric provides a robust understanding of bias in ASR systems, allowing for a more comprehensive evaluation and subsequent mitigation of biases. After calculating a bias of all groups per model and speaking style with a bias metric B , I summed all biases per model (Eq 14).

$$\text{Total Sum}_M = \sum_{g \in G} B(g_{MRd}) + B(g_{MHMI}) \quad (14)$$

4. Results

In this section, I present the results of the bias metrics mentioned in Section 3.1 and compare them to bias metrics calculations from [4].

4.1. Word Error Rates of the ASR Systems

Firstly, I calculated the WERs as in Eq 15, which compares the reference transcription and the automatic output, where S stands for the number of substituted words, D for the number

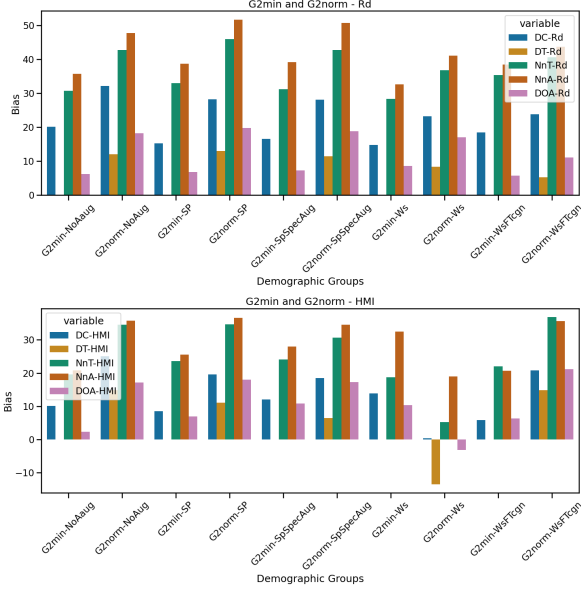


Figure 1: Bias calculated by $G2_{min_diff}$ and $G2_{norm_diff}$ from [4] on three Conformer and two Whisper models.

of deleted words, I for the number of inserted words and N is the number of total words in the reference transcription [18].

$$WER = \frac{S + D + I}{N} \quad (15)$$

In Table 1 the WERs are calculated per model (Conformer - *NoAug*, *SpAug*, *SpSpecAug* or Whisper - W_s , W_sFT_{cgn}), per speaking style (Read or HMI) and per group (DC, DT, NnT, NnA, DOA). In the last column, I calculated the average WERs of Read and HMI and both together. The most important part of this code can be found in Figure 8 in Appendix B.

In [1] the bias is measured by the difference in WER, so this is also analyzed here. The average WERs for HMI (52.26, 50.33, 45.6) are larger than in Read (43.61, 42.75, 40.54). After the augmentations *SpAug*, *SpSpecAug*, W_sFT_{cgn} the WERs did get lower, especially in the HMI speaking style.

4.2. Bias Metrics Results

In this subsection, I explain my results from applying all bias metrics mentioned in Section 2.2 and Section 3.1.

4.2.1. Group-to-min and Group-to-norm

Firstly, in Figure 1 and Figure 2 are the results after evaluating the models on bias metrics $G2_{min}$ and $G2_{norm}$ from [4]. These values are taken directly from that paper to compare their results with this experiment. The values can be found in table form in Tables 2, 3 in Appendix C.

Figure 1 and Figure 2 show the magnitude of bias on the y-axis per model on the x-axis. The higher the magnitude of bias, the more biased the model is on that speaker group. Negative values can only occur with $G2_{norm}$ and they represent speaker groups with lower bias than the reference group, which in these cases are DT-HMI and DOA-HMI.

We can also observe that even though the WER for HMI was higher on average than in Read the bias within the model-speaking style is lower than in Read. This means that if we

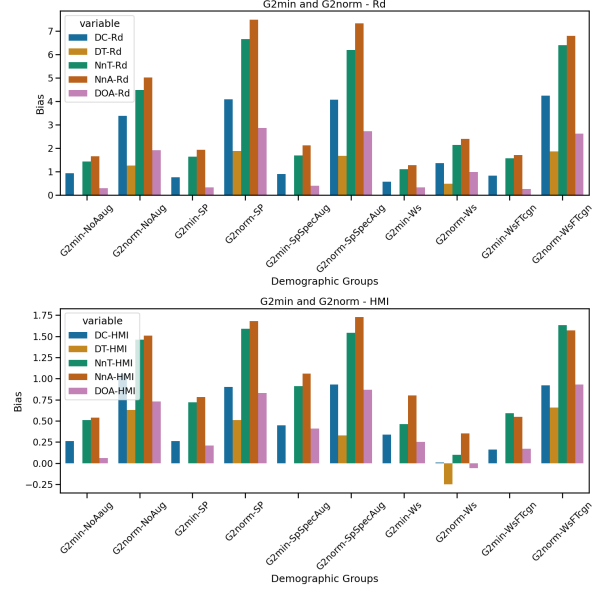


Figure 2: Bias calculated by $G2_{min_reldiff}$ and $G2_{norm_reldiff}$ from [4] on three Conformer and two Whisper models.

compare the speaker groups within HMI their performance is more similar than within Read. This shows that looking at only lower average WERs can obscure some bias characteristics.

4.2.2. Group-to-average and Sum of Group Error Differences

Secondly, Figures 3 and 4 are the results of applying the new metrics within the ASR system, the Group-to-Average Log Ratio $G2_{avg \log \text{ratio}}(g)$ from Eq 12 and the Sum of Group Error Differences SED_{gMS} from Eq 13. In both figures the magnitude of the bias measured by the two bias metrics is on the y-axis, the models are on the x-axis and the speaking style Rd is on the top graph whereas the speaking style HMI is on the bottom one.

In Figure 3 we can observe that the bias can be both positive and negative, as the average bias (unbiased group) has a value of 0. Therefore the speaker groups with negative bias (NnT and NnA) have worse-than-average performance, but we can also observe that the speaker group DT has the best performance everywhere.

For the Sum of Group Error Differences metric in Figure 4 all of the biases are positive because of the absolute value function in Eq 9. Here again, the value of the unbiased group is 0, and the bigger the value, the bigger the bias. This makes the most biased speaker groups by this metric mostly the DT and the NnA.

4.2.3. Total Sum of Biases

In Figure 5 are the results of applying Eq 14 on the values from Figure 3. It shows a comparison of all models, on the x-axis, on how high the sum of all their group biases is. From Figure 5 we can see that the models with the highest summative bias are *SpAug*, *SpSpecAug*.

5. Responsible Research

In this section, I explain how I conducted my research responsibly. In the first part, I talk about how the content of my research

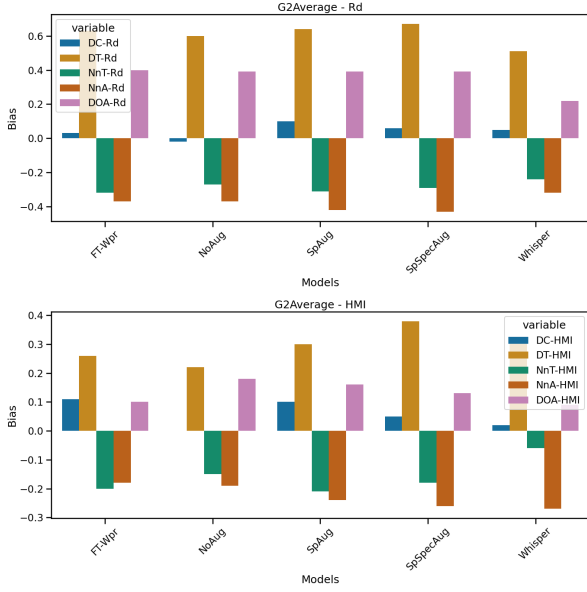


Figure 3: Bias calculated by $G2_{avg} \log ratio$ from [7] on three Conformer and two Whisper models.

is responsible, and in the second part how I conducted my research in a responsible way.

5.1. Ethical Considerations

Throughout conducting my research I actively worked on ensuring the ethical integrity of my experiment. As my research concerned quantifying the fairness of computer systems, I had to make sure I was not reinforcing the bias even more, by creating unfair bias metric. This was done by making sure my new bias metric is designed with fairness principles in mind. The goal was not to just blindly lower the error rates in the model but to evaluate the system thoroughly, looking at various relationships between the groups and also looking at the distribution of the error rates.

Furthermore, the broader goal of my research is to contribute to mitigating the bias of the ASR system by finding the demographic groups that suffer the most. This should help make these systems more fair, i.e. not favoring any of the demographic groups.

5.2. Reproducible Research

Conducting reproducible research is an important part of the responsibility of a researcher. I made sure I wrote my paper in such a way that anyone with enough programming knowledge could recreate my experiment and review it or extend it.

Firstly, I wrote a detailed methodology where I described all the datasets and models I was working with, so the next researchers could compare their data with mine at the same baseline. I also added references to where to find them. Next, I also wrote the steps of how I calculated my bias metrics, and in appendices, I referenced the most important part of my code. My repository is also available publicly with my results in CSV files¹. I made sure the program was also thoroughly documented, the methods explained and the variables names properly. These

¹ Available [at this link](#)

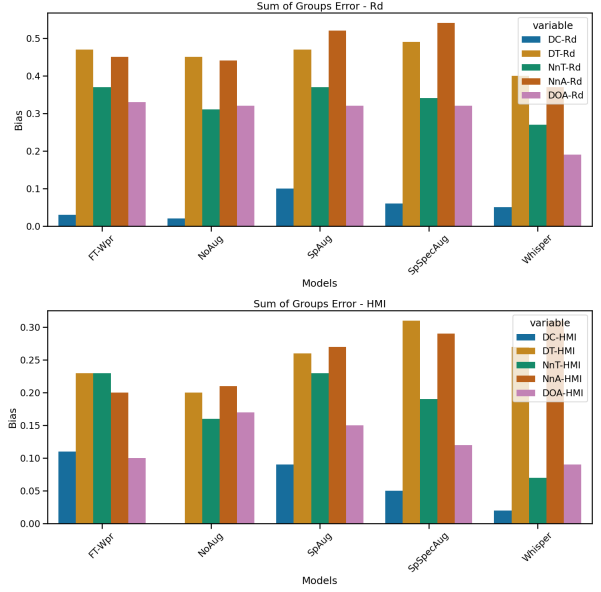


Figure 4: Bias calculated by SED_g from [17] on three Conformer and two Whisper models.

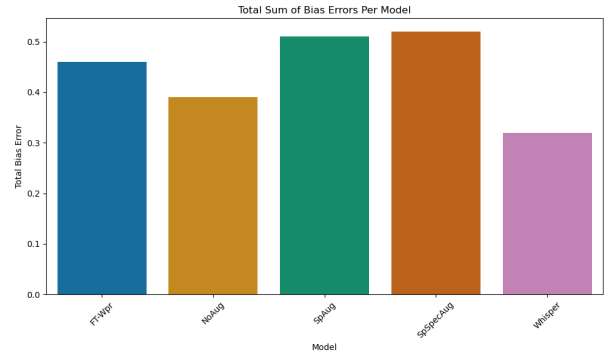


Figure 5: Total bias calculated by $\sum G2_{avg} \log ratio$ from [7, 17] on three Conformer and two Whisper models.

features of my report ensure that the Transparency principle of the Netherlands Code of Conduct for Research Integrity [19] is upheld.

However, what needs to be explained is my choice to use the Jasmin dataset as my test dataset. Although Jasmin is not an open-access dataset, it is free to use for academic purposes². The use of this dataset was important to properly evaluate the models for the Dutch language, to have data from diverse demographic groups, and to have different test set from the training set.

Furthermore, to have the research responsible it was important to uphold the principle of Honesty from the Code of Conduct [19]. This means objectively writing down the limitations of my research, citing my sources properly, being honest about any pre-processing of my data, and also being honest about my usage of any Large Language Models (LLMs) while doing my research. The context of usage of LLMs and prompts I wrote are in Appendix D.

² Available [at this link](#) for research purposes upon request.

Lastly, to make my research more objective I took part in two peer review sessions, one (mandatory for the course) with my classmates and one with my project group. I conducted these peer review sessions with “Standards for good research practices - Assessment and peer review” [19][p. 18] in mind.

6. Discussion

In this section, I discuss my results from Section 4 and how they can relate to my assumptions from Section 3.2. This way, I compare the new approach of calculating bias to one from [4].

6.1. Assumptions

In Section 3.2 I made various assumptions about what properties of a bias metric in the ASR should be prosperous and applied these assumptions to my implementation of a bias metric. Here I explain how these assumptions relate to my bias metric and how the results reflect these assumptions.

The Assumption 1 prefers the ratio approach to measure bias over the difference approach. In Figure 2 where we normalize the bias by dividing by reference group, the biases show a different trend where $G2_{norm} - SP$ show proportionally higher bias than e.g. $G2_{norm} - W_s$ than in Figure 1. This represents the importance of using ratio-based metrics.

In Assumption 2 the absolute value function was meant to give a clearer comparison of the magnitudes of the biases. By plotting the results in Figure 4, the comparison of the magnitude of the bias is clearer than in Figure 3 but it loses the information about whether the bias is positive or negative. This is the reason I chose not to include the absolute value function in my implementation, but there is a way to include it and show the direction of the bias in a different place.

The application from Assumption 3 can be seen in both Figures 3 and 4 as both of these have a neutral bias at value 0. This is more clear than having it centered at 1.

The Assumption 4 was applied in Figure 5, where I sum the biases ($G2_{avg \log \text{ratio}}(g_{MS})$) of all groups for both speaking styles per model. The importance of this assumption is shown here as even though the augmentations in *SpAug*, *SpSpecAug* lower the average WER the summative bias in the models is the highest.

The most important assumption was the Assumption 5 which talked about the importance of not utilizing a reference group when calculating the bias of a model. To evaluate this assumption Figures 1 (using a reference group) and 3 (not using a reference group) need to be compared. Firstly, the speaker group DT always has the smallest bias in all models with bias metric $G2_{min_{diff}}$ and $G2_{norm_{diff}}$ in Figure 1. However, with my new bias metric DT is one of the groups (with NnT and NnA) with the largest bias. This depends on how we calculate our bias metric but because the DT is the best-performing group by far among the speaker groups in our definition this is reflected in the bias. This is one of the examples of how bias metrics with reference groups could hide certain characteristics of a bias.

7. Conclusions and Future Recommendations

In this section, I discuss how my research could be limited, what next steps I propose to make this research better, and lastly, my research is concluded.

7.1. Limitations and Future recommendations

Firstly, as there is no ground truth to bias - no clear definition of what groups are biased - it is challenging to establish if a bias metric is better than another. I only analyzed in my research the limitations of a bias metric with a reference group and showed differences in the results between bias metrics. Future research could take a look at different ways of evaluating bias metrics in ASR. For example in [17] they simulated an environment where either one group is disadvantaged or multiple groups are disadvantaged and evaluated the results of their bias metric on models with these datasets. A similar test simulation could be used also on the bias metrics in ASR.

Secondly, the datasets used in ASR are not completely representative of the population. The CGN dataset [13] only consists of adult native speakers and the Jasmin dataset [14] is missing this speaker group. For this reason, the average WER calculated per model and speaking style (Eq 10) could skew in a certain direction because of the missing speaker group. Furthermore, the term non-native speaker group is too broad to generalize. It is important to take into account also the language background of the non-native speakers as the ones with similar language backgrounds (e.g. Germany) could have better results for WER than people from different language groups. This is especially important to consider for a country like the Netherlands where a large part of the population has also, European (non-Dutch) origin but also non-European origin [20].

7.2. Conclusion

In conclusion, this research introduces a new approach to quantifying bias in automatic speech recognition (ASR) systems without using a reference group in calculation. It combines ideas from the Group-to-Average Log Ratio and the Sum of Group Error Differences approaches to make the new bias metric suitable for our research purposes and make it uncover more bias characteristics than in bias metrics used by [4].

The results of this research were compared to results by Patel et al. [4]. While no bias metric can be universally better, as each problem is different and there is no ground truth to bias, the new metric reduced the bias caused by utilizing the reference group. The results from the new bias metric also show that while the native teenagers (DT) speaker group has the best performance, this could be also seen as a positive bias in the system.

Furthermore, it is important to not just measure the across-group disparity, as we need to look at the bias from different angles, the new bias metric also uses summative bias calculation. With this, it is possible to also compare bias across different models, which shows bias in the models after augmentation (*SpAug*, *SpSpecAug*).

The ultimate goal is to enhance the fairness of ASR systems, making them more accessible and beneficial for all users, irrespective of their linguistic background or demographics.

8. References

- [1] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," April 2021. [Online]. Available: <https://arxiv.org/abs/2103.15122>
- [2] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr 2020.
- [3] M.-H. Yeh, B. Metevier, A. Hoag, and P. Thomas, "Analyzing the relationship between difference and ratio-based fairness metrics," pp. 518–528, Jun. 2024.
- [4] T. Patel, W. Hutiri, A. Ding, and O. Scharenborg, "How to evaluate automatic speech recognition: Comparing different performance and bias measures," 2024, Work in progress.
- [5] P. Dheram *et al.*, "Toward fairness in speech recognition: Discovery and mitigation of performance disparities," in *Interspeech 2022*, Sep 2022, pp. 1268–1272.
- [6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *arXiv*, Nov. 2011. [Online]. Available: <https://doi.org/10.48550/arXiv.1104.3913>
- [7] W. Hutiri, T. Patel, A. Ding, and O. Scharenborg, "As biased as you measure: Methodological pitfalls of bias evaluations in speaker verification research," 2024, submitted to Interspeech 2024.
- [8] T. Patel and O. Scharenborg, "Using data augmentations and vtn to reduce bias in dutch end-to-end speech recognition systems," July 4 2023, available at [arXiv:2307.02009](https://arxiv.org/abs/2307.02009).
- [9] K. Lum, Y. Zhang, and A. Bower, "De-biasing "bias" measurement," June 2022, pp. 379–389.
- [10] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech Language*, vol. 84, 2023.
- [11] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv*, May 16 2020, doi: 10.48550/arXiv.2005.08100.
- [12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv*, Dec 06 2022, doi: 10.48550/arXiv.2212.04356.
- [13] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, Eds. Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: <http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf>
- [14] C. Cucchiari, H. Van hamme, O. van Herwijnen, and F. 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 *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds. Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: <http://www.lrec-conf.org/proceedings/lrec2006/pdf/254.pdf.pdf>
- [15] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers," pp. 146–151, 2023. [Online]. Available: <https://doi.org/10.1109/SpeD59241.2023.10314895>
- [16] C. Liu *et al.*, "Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6162–6166.
- [17] A. Elobaid, N. Ramoly, L. Younes, S. Papadopoulos, E. Ntoutsis, and I. Kompatsiaris, "Sum of group error differences: A critical examination of bias evaluation in biometric verification and a dual-metric measure," *18th International Conference on Automatic Face and Gesture Recognition (FG)*, April 2024. [Online]. Available: <https://arxiv.org/abs/2404.15385>
- [18] I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," IDAP, Martigny, Switzerland, Idiap-RR Idiap-RR-73-2004, 2004.
- [19] KNAW, NFO, NWO, TO2-Federatie, Vereniging Hogescholen, and VSNU, "Nederlandse gedragscode wetenschappelijke integriteit," Data Archiving and Networked Services (DANS), Tech. Rep., 2018.
- [20] "Statline - bevolking; kerncijfers," <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85496NED/table?ts=1719042061753>, accessed: Jun. 22, 2024.

A. Appendix - Bias metric

```
def calculate_bias_metric(df):  
  
    # Calculate the average WER for each Model and Style combination  
    average_wer = df.groupby(['Model', 'Style'])['WER'].mean().reset_index()  
    average_wer.rename(columns={'WER': 'Average_WER'}, inplace=True)  
  
    # Merge this average WER back into the original DataFrame  
    df = pd.merge(df, average_wer, on=['Model', 'Style'])  
  
    # Calculate the adjusted WER  
    df['Bias_Average'] = -np.log(df['WER'] / df['Average_WER'])  
  
    return df
```

Figure 6: Method to calculate $G2_{avg \log ratio}(g_{(MS)})$

```
def calculate_bias_metric(df):  
    # Calculate the average WER for each Model and Style combination  
    average_wer = df.groupby(['Model', 'Style'])['WER'].mean().reset_index()  
    average_wer.rename(columns={'WER': 'Average_WER'}, inplace=True)  
  
    # Merge this average WER back into the original DataFrame  
    df = pd.merge(df, average_wer, on=['Model', 'Style'])  
  
    # Calculate the adjusted WER  
    df['Bias_SumOfErrors'] = abs(1 - (df['WER'] / df['Average_WER']))  
  
    return df
```

Figure 7: Method to calculate $SED_{g_{(MS)}}$

B. Appendix - WER Calculation

```
def calculate_wer(corr, sub, del_, ins):
    return (sub + del_ + ins) / (corr + sub + del_)*100

# Microsoft's Copilot helped me with this function https://copilot.microsoft.com/
def process_error_rate_files(filepath_manager):
    results = []

    for style_idx, (style_folder, style_infix) in enumerate(zip(filepath_manager.
        get_speaking_style_folders(), filepath_manager.get_speaking_style_infixes())):
        for group in filepath_manager.get_speaker_groups():
            for model in filepath_manager.get_asr_models():
                error_rate_path = filepath_manager.get_error_rate_path(style_folder,
                    style_infix, group, model)

                # Read the file and get the summary row (third from the bottom)
                try:
                    df = pd.read_csv(error_rate_path)
                    summary_row = df.iloc[-3]
                    corr, sub, del_, ins = summary_row[['Corr', 'Sub', 'Del', 'Ins']]

                    wer = calculate_wer(corr, sub, del_, ins)
                    results.append({
                        'Style': style_folder,
                        'Group': group,
                        'Model': model,
                        'WER': wer
                    })
                except FileNotFoundError:
                    print(f"File not found: {error_rate_path}")
                except Exception as e:
                    print(f"Error processing file {error_rate_path}: {e}")

    return pd.DataFrame(results)
```

Figure 8: Methods to process the Jasmin files and calculate the WER per group and model

C. Appendix - Bias metrics results

Diff - Rd	DC-Rd	DT-Rd	NnT-Rd	NnA-Rd	DOA-Rd
G2min-NoAug	20.1	0.0	30.7	35.7	6.2
G2norm-NoAug	32.1	12.0	42.7	47.7	18.2
G2min-SP	15.2	0.0	32.9	38.6	6.8
G2norm-SP	28.2	13.0	45.9	51.6	19.8
G2min-SpSpecAug	16.6	0.0	31.2	39.1	7.3
G2norm-SpSpecAug	28.1	11.5	42.7	50.6	18.8
G2min-Ws	14.8	0.0	28.3	32.6	8.6
G2norm-Ws	23.2	8.4	36.7	41.0	17.0
G2min-WsFTcgn	18.5	0.0	35.3	38.4	5.8
G2norm-WsFTcgn	23.8	5.3	40.6	43.7	11.1

Table 2: Table of results Rd after applying bias metric from Eq 2 and 4 from [4].

Diff - HMI	DC-HMI	DT-HMI	NnT-HMI	NnA-HMI	DOA-HMI
G2min-NoAug	10.1	0.0	19.6	20.8	2.3
G2norm-NoAug	25.0	14.9	34.5	35.7	17.2
G2min-SP	8.5	0.0	23.6	25.5	6.9
G2norm-SP	19.6	11.1	34.7	36.6	18.0
G2min-SpSpecAug	12.0	0.0	24.1	28.0	10.8
G2norm-SpSpecAug	18.5	6.5	30.6	34.5	17.3
G2min-Ws	13.9	0.0	18.7	32.5	10.3
G2norm-Ws	0.4	-13.5	5.2	19.0	-3.2
G2min-WsFTcgn	5.9	0.0	22.0	20.7	6.3
G2norm-WsFTcgn	20.8	14.9	36.9	35.6	21.2

Table 3: Table of results HMI after applying bias metric from Eq 2 and 4 from [4].

D. Use of Large Language Models (LLMs)

I utilized Large Language Models (LLMs) for this research for the purposes of:

- brainstorming ideas
- formatting my ideas in latex form

These are the prompts I used for these purposes:

Q: create bibtex: [1] C. Liu et al., 'Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions', in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 6162–6166. doi: 10.1109/ICASSP43922.2022.9747501.

A: Here's a BibTeX entry for the conference paper you've described:

“bibtex @inproceedingsliu2022fairness, title = Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions, author = Liu, C. and others, booktitle = ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), year = 2022, pages = 6162–6166, month = may, doi = 10.1109/ICASSP43922.2022.9747501 ““

This BibTeX entry includes all the necessary details for citing a conference paper, such as author, title, conference name, year, pages, and DOI.

Q: can you generate a structure for my abstract?

A: Certainly! Here's a structured outline to help you craft your abstract for the research report on bias measurement in Automatic Speech Recognition (ASR) systems:

1. **Purpose** - Briefly introduce the main objective of the research. - State the problem with current bias measurement methods in ASR systems.
2. **Methodology** - Outline the methods used to develop and evaluate the new bias metrics. - Mention the datasets (CGN and Jasmin corpus) and ASR models (E2E Conformer and OpenAI's Whisper model) used for testing.
3. **Key Findings** - Highlight significant findings from applying the new metrics compared to traditional methods. - Discuss how these findings reveal different aspects of bias in ASR systems.
4. **Implications** - Explain the implications of these findings for the design and evaluation of ASR systems. - Suggest how these new metrics could lead to more equitable ASR technologies.
5. **Future Research** - Propose directions for future research to refine these metrics and explore their applicability in diverse contexts.

This structure ensures that your abstract is clear, concise, and effectively communicates the purpose, method, findings, and significance of your research to your readers.

Q: i want to have a numbered assumption in latex

A: To create a numbered assumption in LaTeX, you can use the `amsthm` package, which provides a straightforward way to define new theorem-like environments. Here's a simple example of how you could define and use a numbered assumption:

Load the `amsthm` package: Include this package in the preamble of your LaTeX document. Define the Assumption Environment: You can define a new environment for assumptions just after loading the `amsthm` package. Use the Assumption in Your Document: Once defined, you can use this environment in your document just like any standard LaTeX environment.

Q: how do i reference this equation?

A: To reference the given equation properly, you should label it within your LaTeX document. This way, you can refer to it using the label elsewhere in your text. Here's how you can do it:

Add a label to your equation. Reference the equation using the label.

Q: can you add a horizontal divider between table columns in latex?

A: Certainly! To add a horizontal divider between the rows in your LaTeX table, you can insert the `midrule` command between these rows.