# As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research

*Wiebke Hutiri*[1,2], *Tanvina Patel*[1], *Aaron Yi Ding*[1], *Odette Scharenborg*[1]

[1]Delft University of Technology, The Netherlands
[2]Sony AI, Switzerland

wiebke.hutiri@sony.com, t.b.patel@tudelft.nl, aaron.ding@tudelft.nl, o.e.scharenborg@tudelft.nl

## Abstract

Detecting and mitigating bias in speaker verification systems is important, as datasets, processing choices and algorithms can lead to performance differences that systematically favour some groups of people while disadvantaging others. Prior studies have thus measured performance differences across groups to evaluate bias. However, when comparing results across studies, it becomes apparent that they draw contradictory conclusions, hindering progress in this area. In this paper we investigate how measurement impacts the outcomes of bias evaluations. We show empirically that bias evaluations are strongly influenced by base metrics that measure performance, by the choice of ratio or difference-based bias measure, and by the aggregation of bias measures into meta-measures. Based on our findings, we recommend the use of ratio-based bias measures, in particular when the values of base metrics are small, or when base metrics with different orders of magnitude need to be compared.

**Index Terms**: speaker verification, speaker recognition, bias, fairness, evaluation, metrics

## 1. Introduction

Speech technologies are increasingly integrated into services where reliable performance is key for human well-being and safety. One such example is speaker verification, which is used for proof-of-life verification of pensioners [1] and authentication of financial transactions [2]. In these social-security and safety-critical applications, prediction errors can lead to technology failures that cause harms to individuals [3]. In many domains that use machine learning, prediction errors have been found to be systematic, correlating with personal and demographic attributes (e.g. age, gender, accent) [4]. The algorithmic fairness and legal communities refer to this phenomenon as *bias*. Biased speech technologies can carry significant social consequences if they assign undesirable outcomes or deny opportunities to people without reason [5]. New regulations, like the EU AI Act, thus place increasing pressure on technology developers and providers to detect and mitigate bias [6].

Several recent studies have found evidence of bias in speaker recognition systems, for example models that are biased by speakers' gender [7, 8, 9, 10, 11], nationality [7, 8, 12], race [11], accent [13] and age [10, 9]. However, despite having similar experimental setups, the studies draw varying conclusions on which groups are favoured and which are prejudiced against. For example, while one study found systems to perform worse for female speakers and non-US nationals [7], another study that trained and evaluated on the same dataset found the opposite; that models perform better for females, and better for UK nationals than for US nationals [8]. One reason for these divergent claims is that studies use different metrics and measures to compare performance errors across groups of people.

In this paper we study how the metrics and measures used to quantify bias impact the validity of bias evaluations of speaker verification systems. First, we introduce terminology to distinguish base metrics from bias measures and meta-measures. We then compare three bias measures and two meta-measures from the literature, showing empirically how they lead to different bias evaluation outcomes. Finally, we demonstrate with a thought experiment how biased speaker verification systems can impact people in a real-world application, and why reliable bias evaluations are important to prevent this. Based on these insights we make recommendations for more reliable bias evaluations that can lead to fairer and more inclusive speaker verification systems.

## 2. Background and Related Work

Base metrics, bias measures and meta-measures are essential components of bias evaluations [14]. *Bias measures* quantify and thus measure bias for the purpose of bias detection (or diagnosis) and mitigation (or intervention) [15]. This paper focuses on the former. When used for detection purposes, bias measures can be applied during model development or post-hoc to test models and applications in order to gain insights into the limits of their performance. Most bias measures are calculated from statistical *base metrics* that quantify model performance or prediction error rates. Common base metrics used in speaker recognition are the false positive (FPR) and false negative (FNR) error rates, equal error rate (EER) and the minimum detection cost (minCDet) [16]. Base metrics can be disaggregated across groups of people to evaluate model performance across demographic or other protected attributes [17]. To compare model performance across groups, bias measures calculate ratios or differences between the base metrics of a group and a reference group, or overall performance. *Meta-measures* aggregate bias measures across groups into a single score for a model [18] to support the comparison of bias across different models. Like bias measures, meta-measures can be computed for different base metrics.

The most common base metric used to measure performance in studies that investigate bias in speaker recognition is the EER [19, 8, 9, 21, 22]. Other base metrics that have been considered are the minCDet [7], the log likelihood-ratio cost function (Cllr) [13], the FNR at a FPR of 1% [21] the false accept rate (same as FPR) and the false reject rate (same as FNR) [22]. Bias measures can be classified broadly as difference-based and ratio-based measures. Most studies use difference-based measures calculated either from the EER [19, 8, 9] or from statistical fairness measures in ML [21, 12]. However, with the exception of the equalized odds ratio, most statis-

Table 1: *Bias measures evaluated in this study*

| Name | Description | Equation | Reference | In meta-measure |
|---|---|---|---|---|
| Group-to-min Difference | Distance between the base metric (b) of a group (g) and the base metric of the best performing group (m) | $G2min\ diff(b)_g = b_g - b_m$ | [19, 8, 9] | FDR |
| Group-to-average Ratio | Ratio between a group's base metric and the average base metric value across all groups | $G2avg\ ratio(b)_g = \frac{b_g}{b_{average}}$ | [7] | - |
| Group-to-average log Ratio | Negative log of the Group-to-average Ratio | $G2avg\ log\ ratio(b)_g = -ln\,(G2avg\ ratio(b)_g)$ | [20] | NRB |

tical fairness measures consider performance disparities either due to false positive or due to false negative errors. As speaker verification systems trade off the FPR and FNR, this limits the utility of statistical fairness measures for bias evaluations in the speaker verification domain. Only one study used a ratio-based bias measure with the EER and minCDet base metrics [7]. Studies that use a meta-measure have adopted the Fairness Discrepancy Rate (FDR) [13, 22], which was first proposed to assess fairness in biometric verification systems [23].

# 3. Method

This section defines the bias and meta-measures that we compare, and describes the experimental setup. The software used for the analysis has been released as a package on PyPI[1].

## 3.1. Bias and Meta-measures

Table 1 defines three bias measures from the literature that we compare in this study: the *Group-to-min (G2min) Difference*, the *Group-to-average (G2avg) Ratio* and the *G2avg log Ratio*. They can be used with any base metric. In addition we compare two meta-measures, the Fairness Discrepancy Rate (FDR) and the Normalised Reliability Bias (NRB), which we define below.

The FDR [23] performs a pairwise comparison of the FPR and FNR differences across groups at a threshold $\tau$. For each error rate it selects the pair with the maximum difference (i.e. the maximum value of the *G2min Difference* bias measure). The maximum differences are then weighted by $\alpha$, and combined into a joint measure, the FDR.

$$\begin{aligned} max_{\triangle FPR}(\tau) &= max\,(G2min\ diff(FPR(\tau))_G) \\ max_{\triangle FNR}(\tau) &= max\,(G2min\ diff(FNR(\tau))_G) \\ FDR(\tau) &= 1 - (\alpha \times max_{\triangle FPR}(\tau) \\ &\quad + (1-\alpha) \times max_{\triangle FNR}(\tau)); 0 <= \alpha <= 1 \end{aligned} \tag{1}$$

The FDR ranges from 0 (most biased) to 1 (least biased). It can be evaluated at different thresholds $\tau$, which produce different design error rates $FPR_{avg}$. Choosing $\tau$ is a form of choosing a base metric, as each threshold produces a unique $(FPR_{avg},\ FNR_{avg})$ pair. The FPRs and FNRs of groups will deviate from those of the system average, unless the system is unbiased. The system can further be evaluated for different weights $\alpha$. When $\alpha = 0$ the FDR only accounts for FN errors. When $\alpha = 1$, only FP errors are evaluated.

As a second meta-measure we consider Reliability Bias, which was proposed to measure quality-of-service harms in on-device keyword spotting systems [20]. The measure calculates the sum of the absolute values of the *G2avg log Ratio*. To make the Reliability Bias meta-measure comparable across variable numbers of groups, we normalise it by dividing by the number of groups ($G$). The Normalised Reliability Bias (NRB) in Equation 2 has a lower bound of 0 when the performance across all groups is equal and the model is unbiased. The upper limit is infinite. The higher the score, the greater the difference between group and average performance, and the more biased the model. Note that this interpretation is opposite to that of the FDR.

$$NRB(b) = 1/G \sum_{g=1}^{G} |G2avg\ log\ ratio(b)_g| \tag{2}$$

## 3.2. Experiment Setup

To investigate the bias and meta-measures, we use a pre-trained end-to-end ResNet-34 speaker verification model from the Clova baseline[2] [24], trained on the VoxCeleb2 dataset [25] as a black-box predictor. The model is evaluated on two evaluation sets constructed from trial pairs in the VoxCeleb1 dataset: VoxCeleb1-H and VoxCeleb1-I. VoxCeleb1-H consists exclusively of trial pairs where speakers have the same nationality and gender. However, prior research showed that across nationalities 8% - 17% of same speaker pairs in VoxCeleb1-H use trial pairs that come from the same voice recording, making these comparisons trivial [26]. Moreover, the proportion of trivial same speaker pairs is not the same across nationalities, which results in a skewed evaluation setup. We thus also evaluate on the VoxCeleb1-I trial pairs proposed in [26].

We limit our bias evaluation to groups that can be constructed from demographic metadata released with VoxCeleb1, namely binary *gender* (male, female) and the intersection of *gender and nationality*, for the following nationalities: Ireland, India (IN), USA (US), Australia (AUS), Canada, UK, Norway (NO) and Germany (DE).

# 4. Results

We now present our results, starting with an overview of model performance and disaggregated base metrics across groups. Next, we anaylse how the base metrics and bias measures, and then the meta-measures impact the outcomes of the bias evaluation. Our analysis is available as a jupyter notebook[3].

The average EER and minCDet values of the speaker verification model are (2.402, 0.008) and (3.657, 0.012) for VoxCeleb1-H and -I respectively. While the performance measures are 50% greater (i.e. the model performs worse) when evaluating on the more challenging conditions of the VoxCeleb1-I set, the overarching trends are similar. We present our analysis on VoxCeleb1-I going forward. Table 2 shows disaggregated base metrics for *gender* groups in the left column labelled 'All', and for intersectional *gender* + *nationality* groups in the remaining columns. Due to space constraints we only show results for the best and worst performing nationalities (IN, US, AUS, NO, DE).

[1] https://pypi.org/project/bt4vt/

[2] We use the "performance-optimized" model.

[3] https://github.com/wiebket/measuring_bias_speech/

Table 2: *Disaggregated EER and minCDet base metrics on VoxCeleb1-I for* ***gender*** *(col. 'All') and* ***gender + nationality*** *groups (**bold** is best performing base metric in group).*

| Base metric | All | IN | US | AUS | NO | DE |
|---|---|---|---|---|---|---|
| **Male** | | | | | | |
| **EER** | **3.581** | 3.218 | 2.999 | 4.362 | 8.210 | 3.013 |
| **minCDet** | **0.011** | 0.018 | 0.010 | 0.012 | 0.025 | **0.009** |
| **Female** | | | | | | |
| **EER** | 3.757 | 7.028 | 3.250 | **2.788** | 4.588 | 10.641 |
| **minCDet** | 0.012 | 0.023 | 0.011 | 0.011 | 0.014 | 0.019 |

For *gender* groups, the model performs better for males than females for both base metrics. For *gender + nationality* groups, the EER is lowest for Australian females, and the minCDet lowest for German males. For male and female *gender + nationality* groups there are groups with substantially worse than average performance. For example, Norwegian males have an EER that is 2.7 times that of US males. Similarly, German females have 3.8 times the EER of Australian females, but only 1.7 times the minCDet. These results show that the performance of the model varies significantly across genders and nationalities, implying that it is biased. However, the results also suggest that **the extent of bias depends on the base metric**.

### 4.1. Impact of Base Metrics and Bias Measures

Table 3 shows the bias measures evaluated for the EER and minCDet base metrics across *gender* and best and worst performing *gender + nationality* groups. For *gender* groups (i.e. column 'All') the model shows preference for the male group across all base metrics and bias measures. The ratio-based bias measures for males have a *G2avg Ratio* less than 1 and a positive *G2avg log Ratio*, indicating that performance for this group is always better than average. For the female group the inverse is true: the *G2avg Ratio* is greater than 1 and the *G2avg log Ratio* is negative, indicating that performance is always worse than average. The difference-based *G2min Difference* evaluates to 0 for the male group, which has the smaller error rates and is thus used as reference. For the female group, the EER and minCDet base metrics are two orders of magnitude apart. This makes it difficult to compare the *G2min Difference* across the base metrics to establish their impact on the measure. The two ratio-based bias measures, by contrast, are invariant to the order of magnitude of the base metric. We can thus compare the bias measures for the EER and minCDet base metrics to confirm what we observed in Table 2, namely that the extent of bias depends on the base metric used to measure performance.

The impact of the bias measures on bias evaluations becomes more evident for the multicategory *gender + nationality* groups (cols IN, US, AUS, DE). Firstly, we observe that **bias measures preserve the ranking of groups by performance for a particular base metric, but can change it across base metrics**. For example, Australian females have the lowest EER and are used as reference for the *G2min Difference*. This group also has the lowest *G2avg Ratio* of 0.762 and the highest *G2avg log Ratio* of 0.271. All bias measures thus show that this group is strongly favoured when using the EER base metric. However, when computing bias measures with the minCDet, the reference for the *G2min Difference* changes to German males, who have the lowest minCDet value. When analysing the *G2avg Ratio* and *G2avg log Ratio*, German males are now the most favoured group, followed by US males, US females and only then Australian females. In addition to changing the order of preference, **a change in base metric can also lead to a different conclusion about bias**, as is the case with Indian males who change from being favoured to being prejudiced against when the base metric changes from the EER to the minCDet.

Table 3: *Bias measures for* ***gender*** *('All'), best and worst performing* ***gender + nationality*** *groups (**bold** is most favoured).*

| Bias meas. | Base | All | IN | US | AUS | DE |
|---|---|---|---|---|---|---|
| **Male** | | | | | | |
| G2min Diff. | EER | 0.000 | 0.429 | 0.211 | 1.573 | 0.224 |
| | minCDet | 0.000 | 0.010 | 0.001 | 0.003 | **0.000** |
| G2avg Ratio | EER | 0.979 | 0.880 | 0.820 | 1.193 | 0.824 |
| | minCDet | 0.954 | 1.571 | 0.863 | 1.046 | **0.749** |
| G2avg log Ratio | EER | 0.021 | 0.128 | 0.198 | -0.176 | 0.194 |
| | minCDet | 0.047 | -0.452 | 0.148 | -0.045 | **0.289** |
| **Female** | | | | | | |
| G2min Diff. | EER | 0.176 | 4.240 | 0.462 | **0.000** | 7.853 |
| | minCDet | 0.001 | 0.015 | 0.002 | 0.002 | 0.011 |
| G2avg Ratio | EER | 1.027 | 1.922 | 0.889 | **0.762** | 2.909 |
| | minCDet | 1.059 | 1.986 | 0.937 | 0.945 | 1.662 |
| G2avg log Ratio | EER | -0.027 | -0.653 | 0.118 | **0.271** | -1.068 |
| | minCDet | -0.057 | -0.686 | 0.065 | 0.056 | -0.508 |

### 4.2. Impact of Meta-measures

We now compare the FDR and NRB meta-measures to consider how aggregating bias measures into a single meta-measure further impacts bias evaluations. We show results for *gender + nationality* groups. *Gender* groups, which are not shown, follow a similar but weaker trend. Figure 1 visualises the results of a bias evaluation using the FDR from Equation 1 with different $\alpha$ and $\tau$. We selected $\tau$ that calibrate the system to $FPR_{avg} = \{0.001, 0.01, 0.025, 0.05, 0.1\}$, and evaluated the FDR at $\alpha = \{0, 0.25, 0.5, 0.75, 1\}$. The figure shows that at a small $FPR_{avg}$ (e.g. 0.001) the FDR approaches 1 (i.e. least bias) as $\alpha$, which increases the weight of the FPR, increases. This implies that the FNR determines the FDR bias value at small $FPR_{avg}$. This trend is reversed for systems calibrated to larger $FPR_{avg}$ (e.g. 0.1), where larger $\alpha$ reduce the FDR, implying that the FPR determines bias.

Next, we conduct a similar evaluation for the NRB from Equation 2, testing it with a range of base metrics; the EER, minCDet, and FPRs and FNRs at systems calibrated to $FPR_{avg} = \{0.001, 0.01, 0.025, 0.05, 0.1\}$. These $FPR_{avg}$ values have been chosen to correspond with those evaluated for the FDR. Figure 2 shows the NRB values for each base metric. Moving from left to right, $FPR_{avg}$ decreases and the NRB increases (i.e. shows greater bias) for the FPRs (green). For the FNRs (blue-grey) the opposite is true: the NRB decreases as $FPR_{avg}$ decreases. For example, at $FPR_{avg} = 0.1$ (i.e. left

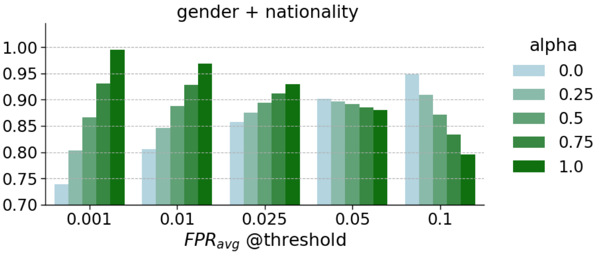

Figure 1: ***FDR meta-measure*** *for* ***gender + nationality*** *groups. The FDR is calculated for different $\alpha$ and for systems calibrated to thresholds that produce pre-determined $FPR_{avg}$. $\alpha = 0$ only considers the FNR, while $\alpha = 1$ only considers the FPR.*

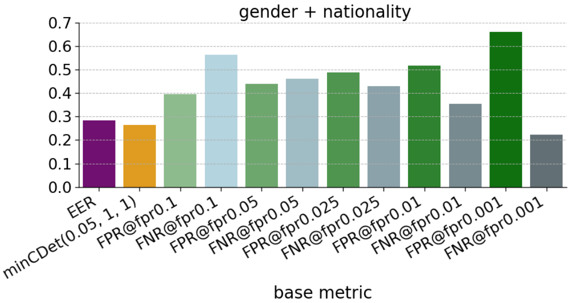

Figure 2: ***NRB meta-measure** for **gender + nationality** groups. The meta-measure is calculated for different base metrics.*

side of the chart), the NRB is lower when calculated with the FPR than with the FNR. However, at $FPR_{avg} = 0.001$ (i.e. right side of the chart), the NRB is substantially greater when calculated with the FPR than the FNR.

**The bias evaluations with the FDR and NRB thus lead to contradictory conclusions.** To illustrate this, consider a hypothetical system that will be used in an application that requires high security. This necessitates a low FPR, and the system is thus calibrated to $FPR_{avg} = 0.001$. Bias is then evaluated specifically for the FPR base metric, given its importance to the use case. We obtain the FDR at $\alpha = 1$, which weights the meta-measure to only consider bias due to the FPR. From Figure 1 we estimate a bias value of $\sim$0.99, which suggests that the model contains minimal bias and is safe to use. Next, from Figure 2 we estimate the NRB of this system as $\sim$0.65 (green bar on the right). This value indicates that substantial performance discrepancies exist across speakers with different genders and nationalities. The system should not be used, as the security of some groups will be severely jeopardised.

### 4.3. Which meta-measure is correct?

How can these two meta-measures lead to opposite conclusions about bias? To investigate this, we decompose the meta-measures into their constituent bias measures and base metric in Table 4. Given the small FPR values, the *G2min Difference* values, which contribute to the calculation of the FDR, are equally small and therefore insensitive to small performance differences across groups. This observation is affirmed by Equation 1, which shows that the FDR is primarily influenced by the order of magnitude of the base metric, and secondly by the value of $alpha$. When the base metric is small, the FDR is thus prone to underestimate performance differences across groups. By contrast, the *G2avg log Ratio*, which is used to calculate the NRB, captures a relative relationship and is unaffected by the order of magnitude of the base metric. The high bias value of the NRB thus reflects the disparity in FPRs across groups shown in Table 4 and correctly identifies the system as biased.

While we have demonstrated that the NRB correctly identifies bias, it remains important to assess if the seemingly small differences in FPR that we observe in Table 4 matter. To explore this, we consider a scenario where an attacker gains access to a device with the previously described speaker verification system. They attempt to access sensitive information on the device by invoking the speaker verification system once a minute (i.e. 60 times / hour). At the design FPR of 0.001 they have a 1 in a 1000 chance of gaining access to the system. After 17 hours (i.e. 1020 attempts), they are likely to succeed. If the device belonged to an Indian male, the FPR of 0.005 would now grant the attackers a 1 in 200 chance of success. This means that they only need to attack the system for 3.5 hours to gain access. This increased exposure to successful attacks presents greater risk of harm to groups that have worse than average performance. The NRB thus correctly identifies this system as biased, while the FDR misrepresents the potential risk.

Table 4: *Disaggregated FPR, G2min Difference and G2avg log Ratio at design $FPR_{avg} = 0.001$ (**bold** > 0.001).*

| | | | | |
|---|---|---|---|---|
| **Male** | | | | |
| **Base metric / Bias measure** | **IN** | **US** | **AUS** | **DE** |
| **FPR@fpr0.001** | **0.005** | 0.000 | 0.001 | **0.002** |
| **G2min Difference** | 0.005 | 0.000 | 0.001 | 0.002 |
| **G2avg log Ratio** | 1.659 | -0.912 | 0.000 | 0.654 |
| **Female** | | | | |
| **FPR@fpr0.001** | **0.003** | 0.001 | **0.002** | **0.001** |
| **G2min Diff** | 0.003 | 0.000 | 0.001 | 0.001 |
| **G2avg log Ratio** | 1.201 | -0.475 | 0.472 | 0.249 |

## 5. Discussion and Limitations

The results in this paper demonstrate that bias evaluations are important to prevent unfair speaker verification systems. However, they also show that evaluations are highly influenced by the choice of base metrics, bias measures and meta-measures. We highlight that the performance ranking of groups depends on the base metric used to measure performance. We further show that bias measures, which are calculated from base metrics, are affected by the order of magnitude of base metrics. Importantly, difference-based measures such as the *G2min Difference* cannot be compared across base metrics with different orders of magnitude, and lack sensitivity when base metrics are small. These shortcomings affect meta-measures based on difference-based bias measures, such as the FDR. We thus recommend the use of ratio-based measures, which are invariant to the magnitude of the base metric, and meta-measures like the NRB that are calculated from ratio-based bias measures. We find the *G2avg log Ratio*, which is centered around 0, easier to compare across multiple groups than the *G2avg Ratio*.

The base metrics and bias measures that we investigate in our empirical study are those that are most frequently used in speaker verification bias evaluations. However, they are not the only metrics and measures that can be used. As our insights pertain to the impact that the magnitude of numbers has on basic arithmetic operations (subtraction and division), our results are broadly applicable to any difference- and ratio-based bias measures, and also do not depend on the model or dataset used for evaluation. Our experimental setup resembles the evaluation scenarios of many prior studies. This made it suitable for our analysis of bias measures used in prior research. However, the VoxCeleb datasets, which have now been retracted by the authors, present numerous ethical and privacy concerns [27]. We thus do not recommend them for evaluation.

## 6. Conclusion

This paper studied the impact of base metrics, bias measures and meta-measures on the outcomes of speaker verification bias evaluations. Our empirical analysis demonstrates that metrics and measures significantly impact evaluation outcomes. We recommend the use of ratio-based bias measures, in particular when the values of base metrics are small, or when base metrics with different orders of magnitude need to be compared.

## 7. References

[1] M. Morrás, "BBVA Mexico allows its pensioner customers to provide proof of life from home thanks to Veridas voice biometrics," 2021. [Online]. Available: https://veridas.com/en/bbva-mexico-allows-pensioner-customers-provide-proof-of-life-from-home/

[2] J. Cox, "How I Broke Into a Bank Account With an AI-Generated Voice," 2023. [Online]. Available: https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice

[3] R. Shelby, S. Rismani, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk, "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," in *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 10 2022.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, jul 2021. [Online]. Available: https://doi.org/10.1145/3457607

[5] B. Friedman and H. Nissenbaum, "Bias in computer systems," *Computer Ethics*, vol. 14, no. 3, pp. 215–232, 1996.

[6] European Parliament, "EU AI Act: first regulation on artificial intelligence," 2023. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[7] W. Hutiri and A. Y. Ding, "Bias in Automated Speaker Recognition," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 230–247. [Online]. Available: https://doi.org/10.1145/3531146.3533089

[8] M. Jin, C. J. T. Ju, Z. Chen, Y.-C. Liu, J. Droppo, and A. Stolcke, "Adversarial Reweighting for Speaker Verification Fairness," in *Proc. Interspeech*, 2022, pp. 4800–4804.

[9] G. Fenu, H. Lafhouli, and M. Marras, "Exploring Algorithmic Fairness in Deep Speaker Verification," *Computational Science and Its Applications – ICCSA 2020*, 2020. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-58811-3_6

[10] G. Fenu, G. Medda, M. Marras, and G. Meloni, "Improving Fairness in Speaker Recognition," *Proceedings of the 2020 European Symposium on Software Engineering*, pp. 129–136, 2020.

[11] X. Chen, Z. Li, S. Setlur, and W. Xu, "Exploring racial and gender disparities in voice biometrics," *Nature Scientific Reports*, vol. 12, no. 1, pp. 1–12, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-06673-y

[12] A. Hajavi and A. Etemad, "A study on bias and fairness in deep speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[13] M. Estevez and L. Ferrer, "Study on the fairness of speaker verification systems across accent and gender groups," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[14] W. Hutiri, "Design Patterns for Detecting and Mitigating Bias in Edge AI," Ph.D. dissertation, Delft University of Technology, 2023.

[15] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[16] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Computer Speech and Language*, vol. 60, 2020.

[17] S. Verma and J. Rubin, "Fairness definitions explained," *FairWare '18*, p. 1–7, 2018. [Online]. Available: https://doi.org/10.1145/3194770.3194776

[18] K. Lum, Y. Zhang, and A. Bower, "De-biasing "bias" measurement," *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 379–389, 2022. [Online]. Available: https://doi.org/10.1145/3531146.3533105

[19] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, and A. Stolcke, "Improving Fairness in Speaker Verification via Group-Adapted Fusion Network," *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 7077–7081, 2022.

[20] W. Hutiri, A. Y. Ding, F. Kawsar, and A. Mathur, "Tiny, Always-on and Fragile: Bias Propagation through Design Choices in On-device Machine Learning Workflows," *ACM Transactions on Software Engineering and Methodology*, 4 2023.

[21] G. Fenu, M. Marras, G. Medda, and G. Meloni, "Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition," *Proc. Interspeech*, pp. 1892–1896, 2021.

[22] R. Peri, K. Somandepalli, and S. Narayanan, "A study of bias mitigation strategies for speaker recognition," *Computer Speech and Language*, vol. 79, 2023. [Online]. Available: https://doi.org/10.1016/j.csl.2022.101481

[23] T. De Freitas Pereira and S. Marcel, "Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2022.

[24] H. S. Heo, B. J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the VoxCeleb speaker recognition challenge 2020," 2020. [Online]. Available: https://arxiv.org/abs/2009.14153

[25] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech and Language*, vol. 60, p. 101027, 2020. [Online]. Available: https://doi.org/10.1016/j.csl.2019.101027

[26] W. Hutiri, L. Gorce, and A. Y. Ding, "Design Guidelines for Inclusive Speaker Verification Evaluation Datasets," in *Proc. Interspeech 2022*. Incheon, Repbulic of Korea: International Speech Communication Association, 2022, pp. 1293–1297.

[27] C. Rusti, A. Leschanowsky, C. Quinlan, M. Pnacek, L. Gorce, and W. Hutiri, "Benchmark Dataset Dynamics, Bias and Privacy Challenges in Voice Biometrics Research," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2023.