INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition

Gianni Fenu[1], *Mirko Marras*[2], *Giacomo Medda*[1], *Giacomo Meloni*[1]

[1]University of Cagliari, Italy
[2]EPFL, Switzerland

{fenu, giacomo.medda}@unica.it, mirko.marras@acm.org, g.meloni31@studenti.unica.it

Abstract

Speaker recognition systems are playing a key role in modern online applications. Though the susceptibility of these systems to discrimination according to group fairness metrics has been recently studied, their assessment has been mainly focused on the difference in equal error rate across groups, not accounting for other fairness criteria important in anti-discrimination policies, defined for demographic groups characterized by sensitive attributes. In this paper, we therefore study how existing group fairness metrics relate with the balancing settings of the training data set in speaker recognition. We conduct this analysis by operationalizing several definitions of fairness and monitoring them under varied data balancing settings. Experiments performed on three deep neural architectures, evaluated on a data set including gender/age-based groups, show that balancing group representation positively impacts on fairness and that the friction across security, usability, and fairness depends on the fairness metric and the recognition threshold.
Index Terms: Speaker Recognition, Speaker Verification, Discrimination, Fairness, Biometrics, Bias, Data Imbalance.

1. Introduction

Increasingly adopted in online and onlife applications, speaker recognition systems aim to confirm or refute the user's identity based on the characteristics of the user's voice [1, 2]. The user is asked to provide samples of his speech, and the resulting utterances are processed to create the enrolled speech model for that user. The vocal sample presented at authentication time is then compared with the enrolled speech model to make the decision. With the wider availability of speech data and increasingly efficient computing resources, these systems have achieved impressively high accuracy by leveraging acoustic representations extracted from deep neural networks, such as X-Vector [3] and ResNets [2]. As in other domains, achieving the highest possible accuracy has been a primary goal along years [4, 5, 6, 7].

However, recent literature in the machine-learning community uncovered algorithmic discrimination, showing that achieving impressive accuracy cannot be the sole goal for machine-learning models shaped for our society [8, 9, 10]. Therefore, fairness-aware models have been proposed, often referring to fairness as a concept of non-discrimination on the basis of the membership to protected groups [11, 12, 13, 14, 15, 16]. Groups are distinguished by a protected feature, e.g., gender and age in anti-discrimination legislation[1]. Hence, group fairness is the absence of group discrimination, and group discrimination is evidenced by disparate outcomes between demographic groups.

Extensive definitions of algorithmic fairness, proposed for the purpose of using them in data-driven algorithms, have led to a long list of criteria and metrics [17]. These criteria are often linked to trade-offs between fairness and other objectives, such as accuracy. Imposing this kind of constraints makes finding an optimum challenging, and fairness criteria are often incompatible under the traditional assumptions [17, 18]. Since different fairness notions lead to different fairness criteria, and not all notions can be fulfilled with one criterion, stakeholders are hence left with the decision among value-concepts when considering a fairness metric. As for now, research in speaker recognition has investigated the susceptibility of these systems to unfairness by mostly focusing on differences in equal error rate across groups [11, 12], not accounting for other important fairness criteria that are still left under-explored in speaker recognition.

In this paper, we investigate the relationship among group fairness metrics in speaker recognition, such as by determining the metric yielding more disparity between groups and explaining the type of discrimination. Specifically, our study aims at answering two key research questions:

- *RQ1: How fair speaker recognition models are under different training data balancing and fairness metrics?*
- *RQ2: What impact does the recognition threshold have on the trade-off among fairness, security, usability?*

To answer these questions, we define a mathematical formalization that serves as a ground for the assessment of disparate outcomes based on a varied set of group fairness metrics. Second, we study the extent to which speaker recognition models emphasize group discrimination according to the considered fairness metrics, under different group balancing settings of the training set. Third, we assess the trade-off between fairness and the traditional security and usability (accuracy) objectives, according to the recognition threshold. Experiments on a large-scale public data set, three deep architectures, and gender/age-based groups[2], show that balancing users' representation across groups positively impacts on all fairness metrics, but it is still not enough to provide fairness guarantees. The trade-off between (security, usability) and fairness depends on the fairness metric and the recognition threshold. Compared to [11, 12], we focus on assessing fairness under additional metrics and extend the considered set of models. With this study, we aim to emphasize the attention required by fair demographic treatments in speaker recognition systems. Code and models are publicly available at https://mirkomarras.github.io/fair-voice/.

[1]Explicit mentions are given in Art. 21 of the EU Charter of Fundamental Rights, Art. 14 of European Convention on Human Rights, and Art. 18-25 of the Treaty on the Functioning of the European Union.

[2]Though gender and age are two important sensitive attributes, there are many other sensitive attributes, such as the geographic provenience, the language, the regional accent, whose analysis is left as future work.

http://dx.doi.org/10.21437/Interspeech.2021-1857

2. Speaker Recognition Formalization

For clarity, we first present and formalize the addressed task and the evaluation protocol. Let $A \subset \mathbb{R}^*$ denote the domain of audio signals with unknown length. We can consider a traditional processing pipeline with an intermediate acoustic representation $S \subset \mathbb{R}^{k \times *}$ (e.g., a spectrogram), where k is the feature vector dimensionality, and/or an explicit feature extraction step which produces fixed-length representations in $D \subset \mathbb{R}^e$, where e is the embedding dimensionality. We denote the respective stages as $\mathcal{F} : A \rightarrow S$ and $\mathcal{D}_\theta : S \rightarrow D$. Given a *decision threshold* τ, a verification trial can be defined as:

$$v_\tau : D_{\theta,p} \times D_{\theta,u} \rightarrow \{0, 1\} \quad (1)$$

which, under the feature extraction hyper-parameters θ, an input feature vector d_p from an unknown user p is compared with a feature vector d_u from user u to confirm or refute the speaker's identity (1 and 0, respectively). In our study, we consider a one-shot verification protocol[3] to align the fairness-aware evaluation proposed in this paper with those of relevant prior works in traditional speaker recognition, that assess the performance of the model by evaluating it along a list of trial verification pairs, such as the trial test pairs in the VoxCeleb-1 set [2]. Our verification protocol relies on a similarity function $\mathcal{S} : D \times D \rightarrow \mathbb{R}$:

$$v_\tau(d_p, d_u) = \mathcal{S}(d_p, d_u) > \tau \quad (2)$$

Therefore, training a speaker recognition model becomes an optimization problem. Given users $\mathbb{U}$, this means finding the model hyper-parameter θ and verification threshold τ that maximize the expectation on the following objective function:

$$(\tilde{\theta}, \tilde{\tau}) = \underset{(\theta,\tau)}{\operatorname{argmax}} \underset{u,p \in U}{\mathbb{E}} \begin{cases} v_\tau\left(D_{\theta,p}, D_{\theta,u}\right) & p = u \\ 1 - v_\tau\left(D_{\theta,p}, D_{\theta,u}\right) & p \neq u \end{cases} \quad (3)$$

In other words, we aim at maximizing the cases where $v_\tau = 1$ when $p = u$ and those where $v_\tau = 0$ when $p \neq u$.

3. Group Fairness Framework

In this section, we describe the data set, the strategies adopted to control the representation of demographic groups in the training set, the set of group fairness metrics part of the framework, the deep architectures considered for the extraction of acoustic feature representations and the underlying security levels.

3.1. Data Set: FairVoice

Despite the existence of several data sets for speaker verification evaluation [19, 20, 2], our fairness study was conducted on FairVoice [11, 12], a data set that offers a large number of utterances and labeled users across several sensitive attributes and languages[4]. This design choice was motivated by the fact that FairVoice already includes baselines and evaluation protocols tailored for fairness analysis in speaker recognition, and our study in this paper aims to extend them with more group fairness metrics and models coming from the fair machine learning community. Each user in FairVoice is identified by the language, gender, and age. Due to the relatively small size of the populations of languages other than English, which may prevent to provide statistically significance findings, we consider only users who provided utterances in English (6,321 speakers). To ensure consistency with the experimental setting provided in [11, 12], our analyses consider four demographic groups based on the users' gender (female, male) and age (users younger than 40 years or not). Further details on the representation of each demographic group are provided later on in the paper.

3.2. Data Splitting and Balancing Strategy

Given that we aim at extending the baselines and benchmarks defined in [11], the strategy adopted to split data in training and test sets follows the same protocol of the original paper, summarized for convenience as follows. We use the same test set proposed in the original paper, considering 100 users (i.e., 25 young females, 25 old females, 25 young males, 25 old males). We also considered the same trial verification pairs. For each test user u_i, 64 trials pairs against utterances of the same user u_i and 64 trial pairs against utterances from a different user u_j with $i \neq j$ are considered, as reported in the original paper. The trial verification pairs are divided into two separate lists, according to an intra-gender and intra-age group scenario. Specifically, the set of intra-age trial pairs have been constructed such that u_i and u_j belong to the same age group, while the intra-gender trial pairs have been constructed to have u_i and u_j with the same gender group. Intra-group trial pairs have been often proved to be the most challenging ones to recognize, so our study in this paper uses the intra-age trial pairs when assessing unfairness on age and intra-gender trial pairs when assessing unfairness on gender. This ensures an adequate representation of each demographic group, making these test sets a suitable tool for fairness evaluation in speaker recognition. Users who are not part of the test set are used to train the speaker verification models. To study the effect of demographically balancing the training set on the fairness achieved by the speaker recognition model, we consider the same training sets defined in [11], as follows:

- **NB** refers to the original unbalanced training set including 718 young females (11.4%), 400 old females (6.3%), 3,935 young males (62.3%), 1,093 old males (17.3%). Each user contributes with at least 20 utterances.
- **UB** refers to a user-balanced training set including 155 young females (25%), 155 old females (25%), 155 young males (25%), and 155 old males (25%). Each user has between 20 and 50 utterances, for consistency[5].

3.3. Group Fairness Metrics

Recent literature in fair machine learning has proposed a varied set of group fairness metrics to assess the discrimination of automated systems against demographic groups characterized by a protected sensitive attribute [17, 18]. In this paper, we aim to analyze fairness in speaker recognition in terms of fairness metrics widely studied in other machine learning domains. extending those covered in [11]. For verification trials based on v_τ (Sect. 2) and two groups a_i, a_j defined based on the sensitive attribute A (gender or age, Sect. 3.2), we consider the following group fairness metrics.

[3] We leave the experiments on other verification protocols, such as the averaging of the enrollment speaker embeddings or the creation of a single embedding by pooling utterances, as future work.

[4] The data set was sampled from *Mozilla Common Voice*, one of the largest corpora including unconstrained speech from diverse acoustic environments. Further details can be found in [12].

[5] The UB training set is very small compared to the NB training set. However, we will show that the difference in overall performance of models trained with these two sets is negligible, while the fairness outcomes of the resulting models is significantly different. The further analysis on the impact of the overall data set size is left as future work.

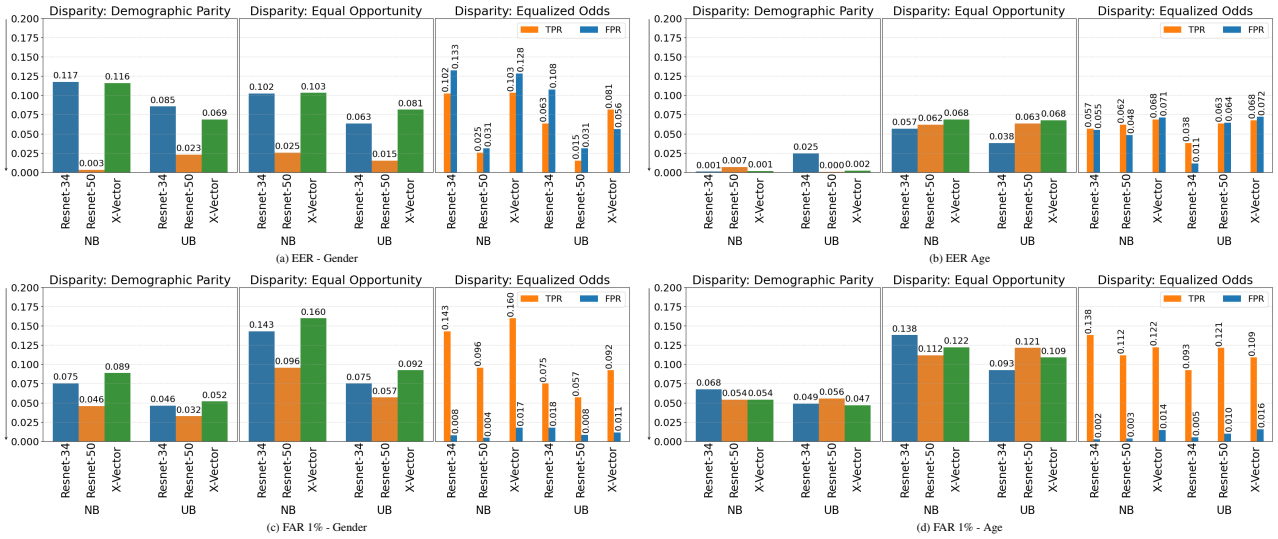

Figure 1: *[RQ1] Fairness estimates of three deep neural architectures (X-Vector, ResNet34, ResNet50), under different training data balancing (NB: unbalanced; UB: user-based balance across demographic groups). The lower the metric is, the fairer the model is.*

- ***Disparity in Demographic Parity [DP]*** implies that the likelihood of a speaker being positively recognized ($PR = (TP + FP)/(TP + FP + TN + FN)$) should be the same regardless of the group. This notion is instantiated as follows:

$$P(v_\tau|A = 0) = P(v_\tau|A = 1). \quad (4)$$

$$DP(\tau) = |PR_{a_i} - PR_{a_j}|, i \neq j. \quad (5)$$

- ***Disparity in Equal Opportunity [EOpp]*** implies that the probability of a speaker being correctly verified should be equal across demographic groups. In other words, the equal opportunity definition states that all the demographic groups should have equal true positive rates (TPR). This notion is defined and operationalized as follows:

$$P(v_\tau{=}1|A{=}0, Y{=}1){=}P(v_\tau{=}1|A{=}1, Y{=}1). \quad (6)$$

$$EOpp(\tau) = |TPR_{a_i} - TPR_{a_j}|, i \neq j. \quad (7)$$

- ***Disparity in Equalized Odds [EOdd]*** implies that the probability of a speaker being correctly verified and of being incorrectly verified should both be the same for the demographic groups. In other words, the equalized odds definition states that the demographic groups should have equal rates for true positives (TPR) and false positives (FPR). This notion is defined and operationalized as follows:

$$P(v_\tau{=}1|A{=}0, Y{=}y){=}P(v_\tau{=}1|A{=}1, Y{=}y), y{\in}0, 1. \quad (8)$$

$$EOdd_{TPR}(\tau) = |TPR_{a_i} - TPR_{a_j}|, i \neq j. \quad (9)$$

$$EOdd_{FPR}(\tau) = |FPR_{a_i} - FPR_{a_j}|, i \neq j. \quad (10)$$

It should be noted that we also considered other fairness metrics, e.g., the fairness discrepancy rate [16]. However, the resulting patterns were similar to those obtained by DP, and so we do not present them in this paper due to space constraints.

3.4. Speaker Recognition Models

For our experiments, we relied on two speaker recognition models, *X-Vector* and *ResNet-34*, whose fairness in terms of equal error rates has been already studied in [11, 12], and an additional model *ResNet-50* [2] whose fairness has not been yet studied, to the best of our knowledge. The *X-Vector* [3] model takes as an input 24 dimensional filterbanks as features with a frame-length of 25ms, mean-normalized over a sliding window of up to 3s. The *ResNet-34* model is similar to a standard multi-layer CNN, but with added skip connections [2]. Composed by 34 residual layers, this model takes as an input spectrograms of size 512 x 300 for 3s of speech using a hamming window of width 25ms and step 10ms. The *ResNet-50* model follows the same specifications of ResNet-34 [2] but is composed by 50 residual layers. The experimental setup and hyper-parameters are provided in the repository shared with this paper.

Table 1: *Performance of the considered speaker recognition models at the EER and FAR1% security levels on FairVoice.*

		ResNet-34		ResNet-50		X-Vector	
	Test set	EER	$FRR_{FAR1\%}$	EER	$FRR_{FAR1\%}$	EER	$FRR_{FAR1\%}$
NB	Intra-age	0.08	0.27	0.09	0.2	0.08	0.2
NB	Intra-gender	0.11	0.43	0.13	0.36	0.11	0.3
UB	Intra-age	0.07	0.18	0.07	0.23	0.08	0.2
UB	Intra-gender	0.11	0.41	0.1	0.33	0.11	0.31

Fairness and accuracy estimates are examined at two well-known security levels: *Equal Error Rate (EER)* and *False Acceptance Rate (FAR) 1%*. Table 1 reports the performance of the speaker recognition models on the data and settings described in Sect. 3.2 and 3.4 (FRR indicates the False Rejection Rate).

4. Experimental Results

In this section, we empirically evaluate the extent to which each speaker recognition model is susceptible to unfairness, in order to answer the two research questions defined in Sect. 1.

4.1. RQ1: Fairness across Data Set Balancing Settings

Our first experiments aim to assess the extent to which speaker recognition models are fair under different training data balancing and group fairness notions. To this end, we consider the three speaker recognition architectures listed in Sect. 3.4, each trained with one of the two training sets *NB* or *UB* defined in

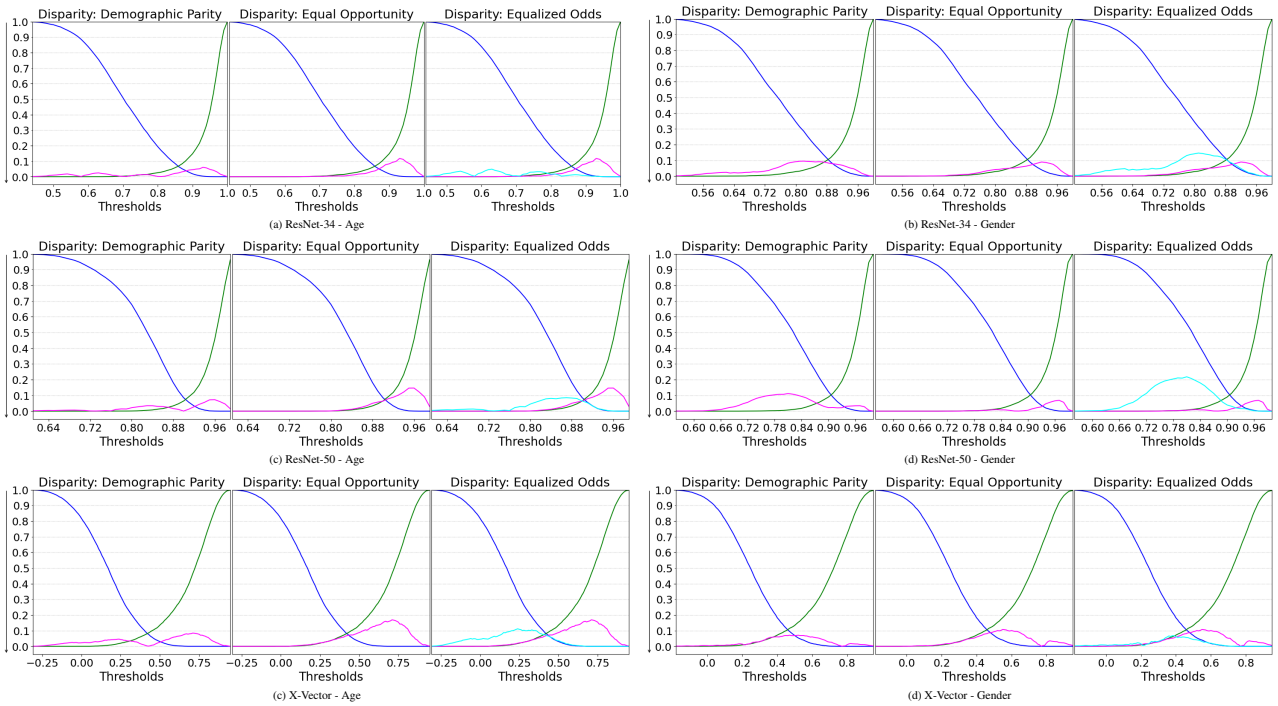

Figure 2: *[RQ2] The impact of the recognition threshold on the trade-off between fairness, security, and usability for three deep architectures, three fairness metrics, for the user-based balanced training set (UB).* ■ *represents the FAR,* ■ *represents the FRR,* ■ *represents the respective fairness metric value. For Equalized Odds* ■ *represents* $EOdd_{TPR}(\tau)$*, while* ■ *represents* $EOdd_{FPR}(\tau)$*.*

Sect. 3.2. For each resulting model, we analyze its fairness under the definitions formalized in Sect. 3.3.

Fig. 1 reports the fairness metric scores on each sensitive attribute (gender: male/female, age: under/over-40), under different training balancing setups and security levels. It can be observed that balancing users across demographic groups often helps mitigating unfairness. Specifically, the disparities between males and females are mitigated for all models under all fairness metrics (except DP for ResNet-50 at EER, and FPR in EOdd for ResNets at FAR 1%). The fairness scores on the age-based groups highlight a good level of mitigation of the disparity between under- and over-40 users as well, but not for all models (e.g., see DP for ResNet-34 at EER). Indeed, ResNet-34 is the one being influenced the most by the balancing of the data set, followed by X-Vector. Surprisingly, regardless of the balancing, the ResNet-50 architecture tends to be fairer on the gender-based groups, while the other two architectures are often fairer than ResNet-50 for age-based groups.

4.2. RQ2: Impact of Recognition Thresholds on Fairness

Our second experiments aim to assess the impact of the recognition threshold on the trade-off between (security, usability) and fairness. Given that models trained on *UB* often led to the fairest results, we focused only on these models. For each model, for each threshold between 0 and 1, we computed the false acceptance rate (FAR), the false rejection rate (FRR), and the fairness estimate, to understand the relation between security (FAR), usability (FRR), and fairness under different fairness notions.

Fig. 2 reports the fairness score, false acceptance rate, and false rejection rate as a function of the recognition threshold, for each speaker recognition model. It can be observed that, for almost all settings, the disparity scores show their peaks nearby the EER and FAR 1% security levels. Our analyses on the age-based groups shed light on unfairness near the FAR 1% threshold, while experiments with the gender-based groups often show unfairness at thresholds close to the EER one. On gender-based groups, the thresholds at which a model achieves lower disparities vary across models. On age-based groups, the thresholds close to EER lead to a degree of unfairness, but not as much as thresholds slightly higher than EER, where the disparities achieve the highest peak. These results highlight the friction between fairness and accuracy (FAR and FRR), confirming the trade-off usually experienced in this task.

5. Conclusions and Future Works

In this paper, we extended an existing fairness framework for speaker recognition with additional group fairness metrics and models, such that the framework can assess disparate outcomes across demographic groups under different fairness notions. We then studied the extent to which speaker recognition models emphasize group discrimination according to the considered fairness notions, under different balancing settings of the training set and recognition thresholds, focusing on two main sensitive attributes for demographic group fairness: gender and age. Our results show that demographically balanced training sets positively influence the fairness of speaker recognition systems. On average, balancing the training set leads to a lower disparity between males and females, compared to the disparity between under- and over-40 users. Models tend to increase the disparity between under/over-40 users at thresholds higher than the EER; the disparity between males and females increases at thresholds below the EER in certain cases. However, balancing the data set does not necessarily guarantee fairness. Future works will focus on approaches able to better mitigate unfairness and on multi-class sensitive attributes beyond gender and age.

6. References

[1] M. Ivanova, S. Bhattacharjee, S. Marcel, A. Rozeva, and M. Durcheva, "Enhancing trust in eassessment - the tesla system solution," in *Technology Enhanced Assessment Conf.*, Dec. 2018.

[2] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[4] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," National Inst of Standards and Technology Gaithersburg Md, Tech. Rep., 1998.

[5] J. Kahn, N. Audibert, S. Rossato, and J.-F. Bonastre, "Intra-speaker variability effects on speaker verification performance." in *Odyssey*, 2010, p. 21.

[6] A. Moez, B. Jean-François, B. K. Waad, R. Solange, and K. Juliette, "Phonetic content impact on forensic voice comparison," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 210–217.

[7] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castan, and A. Lawson, "Analysis of critical metadata factors for the calibration of speaker recognition systems." in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 4325–4329.

[8] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation" ," *AI Magazine*, 2017.

[9] J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Proc. of the European Conference on Speech Communication and Technology (ECSCT)*, 2003.

[10] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[11] G. Fenu, G. Medda, M. Marras, and G. Meloni, "Improving fairness in speaker recognition," in *Prof. of the European Symposium on Software Engineering (ESSE)*. ACM, 2020, p. 129–136.

[12] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *Proc. of the International Conference eon Computational Science and Its Applications (ICCSA)*, 2020, pp. 77–93.

[13] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[14] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning," *arXiv preprint arXiv:2004.11246*, 2020.

[15] I. Serna, A. Peña, A. Morales, and J. Fierrez, "Insidebias: Measuring bias in deep networks and application to face gender biometrics," in *Proc. of IAPR International Conference on Pattern Recognition (ICPR)*, 01 2021.

[16] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," 2020.

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[18] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. of the IEEE/ACM International Workshop on Software Fairness (FAIRWARE)*. IEEE, 2018, pp. 1–7.

[19] M. P. Alvin and A. Martin, "Nist speaker recognition evaluation chronicles," in *Proc. of the IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2004, pp. 12–22.

[20] M. A. Przybocki, A. F. Martin, and A. N. Le, "Nist speaker recognition evaluation chronicles - part 2," in *Proc. of the IEEE Odyssey - Speaker and Language Recognition Workshop*, 2006, pp. 1–6.