Fairness and Privacy in Voice Biometrics: A Study of Gender Influences Using wav2vec 2.0

Oubaïda Chouchane, Michele Panariello, Chiara Galdi, Massimiliano Todisco, Nicholas Evans

Fairness and Privacy in Voice Biometrics: A Study of Gender Influences Using wav2vec 2.0

Oubaïda Chouchane, Michele Panariello, Chiara Galdi, Massimiliano Todisco, Nicholas Evans

EURECOM
Sophia Antipolis, France

firstname [dot] *lastname* [at] *eurecom* [dot] *fr*

***Abstract*—This study investigates the impact of gender information on utility, privacy, and fairness in voice biometric systems, guided by the General Data Protection Regulation (GDPR) mandates, which underscore the need for minimizing the processing and storage of private and sensitive data, and ensuring fairness in automated decision-making systems. We adopt an approach that involves the fine-tuning of the wav2vec 2.0 model for speaker verification tasks, evaluating potential gender-related privacy vulnerabilities in the process. Gender influences during the fine-tuning process were employed to enhance fairness and privacy in order to emphasise or obscure gender information within the speakers' embeddings. Results from VoxCeleb datasets indicate our adversarial model increases privacy against uninformed attacks, yet slightly diminishes speaker verification performance compared to the non-adversarial model. However, the model's efficacy reduces against informed attacks. Analysis of system performance was conducted to identify potential gender biases, thus highlighting the need for further research to understand and improve the delicate interplay between utility, privacy, and equity in voice biometric systems.**

***Index Terms*—Speaker verification, privacy preservation, fairness, gender concealment, wav2vec 2.0**

I. Introduction

The voice is an appealing approach to biometric authentication. Its merits include ease of use, contactless and natural interaction, efficiency, and application to authentication at a distance, e.g. over the telephone. However, the voice is a rich source of personal information and recordings of speech can be used to infer far more than just the speaker's identity, e.g. the speaker's gender [27], ethnicity [10], and health status [22]. The safeguarding of such extraneous personal information is nowadays essential; without it, there is no guarantee that recordings of speech will not be used for purposes beyond person authentication [19].

The General Data Protection Regulation (GDPR)[1] calls for adequate protections for personal data, encompassing both *sensitive* biometric information like voice and *personal* attributes such as gender[2]. In adherence to Art. 4(1) of the GDPR, personal data processing must abide by principles of legality and fairness, managing data in line with reasonable expectations and avoiding unjust harm. Any AI-driven data processing resulting in unfair discrimination violates this principle.

[1]https://gdpr-info.eu/

[2]https://www.gdpreu.org/the-regulation/key-concepts/personal-data/

As mandated by GDPR, this study particularly emphasizes privacy and fairness, focusing on gender due to its demonstrated influence on speaker authentication services [9] and the observed gender bias in voice assistant responses [13]. GDPR aims to protect the rights and freedoms of individuals, including privacy and non-discrimination, with regard to personal data processing. Concealing gender adheres to the principles of data minimization and privacy by design, limiting the risk of misuse or unauthorized data access.

In this research, we grapple with the triple challenge of utility, privacy, and fairness in speaker verification systems. Starting with fine-tuning a pre-trained wav2vec 2.0 for speaker verification tasks, we then evaluate potential vulnerabilities tied to gender privacy and the fairness of Automatic Speaker Verification (ASV) performance across genders. Subsequently, we implement an adversarial technique during the fine-tuning process to conceal gender information in the speaker embeddings, thereby enhancing user privacy. To conclude, we present a comprehensive analysis of the impact of gender information on the utility, privacy, and fairness of the systems we propose.

II. Related work

Significant strides have been made in speaker verification, with efforts concentrated on enhancing user privacy. These strategies prioritize the protection of gender-specific data without sacrificing system utility. Noé et al. [15] suggested an Adversarial Auto-Encoder (AAE) method to separate gender aspects from speaker embeddings while preserving ASV performance. The approach uses an external gender classifier to analyze encoded data. Later, they leveraged a normalizing flow to control gender information in a flexible manner [16]. In another study, Benaroya et al. [2] developed a novel neural voice conversion framework using multiple AEs to create separate linguistic and extra-linguistic speech representations, allowing adjustments during the voice conversion process. Recently, Chouchane et al. [3] used an adversarial approach to hide gender details in speaker embeddings while ensuring their effectiveness for speaker verification. They incorporated a Laplace mechanism layer, introducing noise to obscure gender information and offering differential privacy during inference.

In terms of fairness, research reveals a distinct disparity in ASV system performance based on gender, exposing gen-

der bias [23]. Two primary strategies to mitigate this bias include pre-processing and in-processing. Pre-processing uses balanced datasets for training, as Fenu et al. [7] demonstrated with gender, language, and age-balanced data. In contrast, in-processing infuses fairness directly during training, as seen in Shen et al.'s Group-Adapted Fusion Network (GFN) [21] and Jin et al.'s adversarial re-weighting (ARW) approach [12]. Peri et al. [18] recently proposed adversarial and multi-task learning techniques for bias mitigation, highlighting a potential trade-off between system utility and fairness.

Finally, shifting focus to system utility, a cornerstone in ASV performance, the wav2vec 2.0 [1], a self-supervised framework for speech representation learning, enters the scene. The wav2vec 2.0 can be effectively adapted for speaker verification tasks [6], [25].

III. Automatic speaker verification, gender recognition and suppression using wav2vec 2.0

In this section, we outline our use of the wav2vec 2.0 model, a versatile speech feature encoder that is pre-trained through self-supervision and can be adapted to specific tasks. We fine-tuned wav2vec 2.0 for three distinct tasks: speaker recognition, and gender recognition and suppression. Section 3.1 elaborates on the pre-training process, while Section 3.2 details our contributions to fine-tuning. Both procedures are graphically depicted in Fig. 1.

A. Pre-training

Given a raw audio input signal x, wav2vec 2.0 produces a set of T feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$. The model is split into a 1D-convolutional encoder and a Transformer module [24] two main parts. First, the encoder maps the raw audio $\mathbf{x}$ to latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$. The latent features are then fed into the Transformer module to produce output feature vectors $\mathbf{c}_1, \ldots, \mathbf{c}_T$, and are also used to compute a set of quantised macro-codewords $\mathbf{q}_1, \ldots, \mathbf{q}_T$. Each macro-codeword $\mathbf{q}_t$ is the concatenation of G codewords $\mathbf{q}_{t,1}, \ldots, \mathbf{q}_{t,G}$ selected from G different codebooks $\mathcal{Q}_1, \ldots, \mathcal{Q}_G$, each of size V, learned at training time. Each codeword $\mathbf{q}_{t,j}$ is sampled from $\mathcal{Q}_j$ according to a V-fold categorical distribution. The distribution is optimized during pre-training and computed as $\mathbf{p}_{t,j} = \mathrm{GS}(\mathbf{z}_t)$, where GS indicates a linear layer projecting $\mathbf{z}_t$ to V dimensions followed by a straight-through Gumbel-softmax estimator [11].

During pre-training, the model attempts to simultaneously minimize a *contrastive* loss $\mathcal{L}_m$ and a *diversity* loss $\mathcal{L}_d$. To compute the former, some of the latent feature vectors $\mathbf{z}_1, \ldots, \mathbf{z}_T$ are randomly masked. Then, for each masked $\mathbf{z}_t$, the Transformer module attempts to compute $\mathbf{c}_t$ so that it is as similar as possible to the corresponding quantised macro-codeword $\mathbf{q}_t$, and as dissimilar as possible from other "distractor" macro-codewords $\tilde{\mathbf{q}}$ randomly sampled from the rest of the batch. The quantised macro-codewords are computed with no masking. The *diversity* loss $\mathcal{L}_d$ encourages the model to make uniform use of all the V codewords in each codebook by maximizing the entropy of the average probability distribution $\bar{\mathbf{p}}_g$ produced by all $\mathbf{z}_t$ in a batch for each codebook g. The overall loss is:

$$\mathcal{L} = \underbrace{-\sum_{\substack{\text{masked}\\ \text{steps } t}} \log \frac{\exp\left(s(\mathbf{c}_t, \mathbf{q}_t)/\kappa\right)}{\sum_{\tilde{\mathbf{q}}} \exp\left(s(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa\right)}}_{\mathcal{L}_m} \underbrace{-\alpha \frac{1}{GV} \sum_{g=1}^{G} H\left(\bar{\mathbf{p}}_g\right)}_{\mathcal{L}_d} \tag{1}$$

Where κ is a temperature coefficient, s is the cosine similarity, α is a weight hyperparameter and H indicates entropy.

B. Fine-tuning for speaker verification and gender recognition

In this paper, we fine-tune wav2vec 2.0 for the downstream tasks of speaker verification and gender recognition. In both cases, for each input utterance $\mathbf{x}$, the output features $\mathbf{c}_1, \ldots, \mathbf{c}_T$ are averaged across time to obtain a 1-dimensional embedding $\mathbf{c}$. In the case of gender recognition, $\mathbf{c}$ is then passed through a linear layer f_g which is trained by optimising the cross-entropy loss $\mathcal{L}_g$ between the predicted logits and the true gender label for each utterance (0 for male, 1 for female). For speaker verification, $\mathbf{c}$ is passed through a different linear layer f_s of N output neurons, where N is the number of speakers in the training dataset. The layer is then optimized to perform speaker identification by minimizing the additive angular margin (AAM) softmax loss $\mathcal{L}_s$ [26]. At test time, the final embedding $\mathbf{c}$ is used as a trial or enrollment vector. Overall, the final loss can be formulated as:

$$\mathcal{L} = \lambda \mathcal{L}_s + (1 - \lambda)\mathcal{L}_g \tag{2}$$

where λ is a hyper-parameter between 0 and 1 that controls the weight of each loss component. We experimented with three different model configurations: Model 1 (M_s) is fine-tuned for speaker verification, i.e. $\lambda = 1$; Model 2 (M_{sg}) is fine-tuned for both tasks, i.e. $\lambda = 0.5$; Model 3 (M_{sga}) is optimised in a similar manner, though with a gradient reversal layer [8] g_r to suppress gender information.

The optimization process becomes an adversarial game between f_g, which attempts to minimize $\mathcal{L}_g$, and the backbone, which attempts to maximize it. Meanwhile, the $\mathcal{L}_s$ component is optimized as usual.

IV. Experimental setup

Described in this section are the databases used for all experimental work, the metrics used for evaluation, and the fine-tuning procedure.

A. Databases

We used the VoxCeleb1 and VoxCeleb2 speaker recognition databases [4], [14]. VoxCeleb1 includes over 100,000 utterances from 1,251 celebrities, while VoxCeleb2 contains over a million utterances from 6,112 speakers. Both datasets, compiled from YouTube videos, are widely used for speaker recognition and voice-related machine-learning tasks. Fine-tuning is performed using the VoxCeleb2 development set which contains data collected from 5994 unique speakers of which 3682 are male and 2312 are female, corresponding to an imbalance in favour of male speakers of 22.9%. To assess

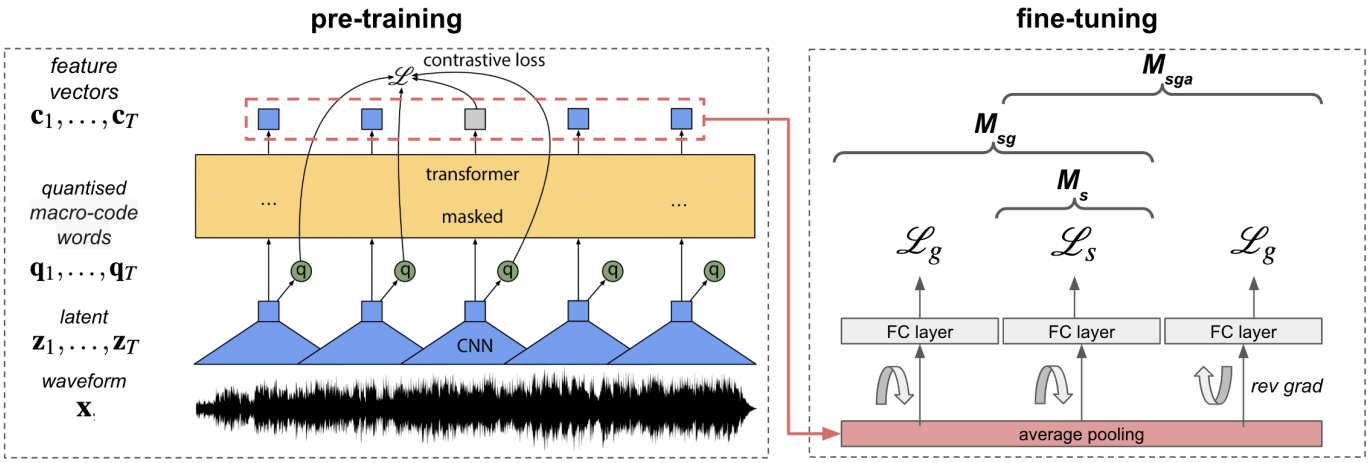

Fig. 1. Graphical depiction of the proposed systems. M_s: fine-tuning the speaker identification task. Msg: fine-tuning gender and speaker identification. M_{sga}: similar to M_{sg}, but the gender identification task is made adversarial.

the performance of our systems, we used the VoxCeleb1 test set, which consists of 40 unique speakers of which 25 are male and 15 are female.

B. Metrics

A range of key metrics was selected, many of which are derived from the evaluation of biometric classification systems, e.g. speaker verification and gender classification. The following describes how they are used to jointly assess the utility, privacy, and fairness of the models under scrutiny.

Utility is measured by assessing the performance for the task of automatic speaker verification (ASV) in terms of equal error rate (EER). EER is the operating point defined by the detection threshold τ at which the false acceptance rate (FAR) and the false rejection rate (FRR) are equal.

Privacy relates to the difficulty of an adversary to infer sensitive attributes. We use AUC (area under the receiver operating characteristic curve) metric to gauge privacy. In contrast to EER, AUC provides a comprehensive view, which is ideal for evaluating system security across diverse threshold selections.

Fairness is aimed at ensuring that a system behaves equally with all subgroups of the target population. Many approaches for measuring fairness have been proposed recently and there is still no agreement on which is the most appropriate. We adopted two different metrics with the aim of giving a more meaningful insight into the fairness of the models.

The first adopted approach aims at ensuring that the error rates for all demographic groups fall within a small margin ϵ. However, for practical purposes, given a pair of demographic groups $D = d_1, d_2$, we calculate $A(\tau)$ and $B(\tau)$, as:

$$A(\tau) = max\left(\left|FAR^{d_1}(\tau) - FAR^{d_2}(\tau)\right|\right) \quad (3)$$

$$B(\tau) = max\left(\left|FRR^{d_1}(\tau) - FRR^{d_2}(\tau)\right|\right). \quad (4)$$

These represent the maximum absolute differences in FAR and FRR across all groups. In a perfect system, both $A(\tau)$ and $B(\tau)$ would equal 0, reflecting identical error rates across all groups.

The Fairness Discrepancy Rate (FDR) [5] is defined as:

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)) \quad (5)$$

where the hyper-parameter $\alpha \in [0, 1]$ determines the relative importance of false alarms. FDR ranges between 0 and 1 and would equal 1 in the case of a perfectly fair system. However, achieving perfect fairness is often unrealistic, leading to the introduction of ϵ which allows for certain discrepancies. Though ϵ isn't included in the FDR calculation, it's vital for defining an acceptable level of fairness and interpreting FDR results.
Given the absence of a universal ϵ and the complexities of biometrics, absolute fairness often isn't achievable. Thus, FDR and Area Under FDR (auFDR) are used to compare the fairness of different biometric systems. The auFDR is calculated by integrating the FDR over a specific threshold range τ, denoted as FAR_x. To fairly compare the auFDR between different systems, the specific range of τ used must be reported, as the value of the auFDR depends on this range. Like the FDR, the auFDR varies from 0 to 1, with higher values denoting better fairness. In our experiments, we set the range to FARs below 0.1; FARs above this value correspond to a system with little practical interest.

The second metric is the fairness activation discrepancy (FAD), which we use to investigate fairness *within* the network. FAD is inspired by *InsideBias* [20], a fairness metric developed originally for the study of face biometrics and which we adapt to our study of voice biometrics. Notably, this adaptation of FAD for voice biometrics is a novel metric in this context.

InsideBias is based upon the examination of neuron activations and the comparison of model responses to demographic groups within distinct layers. In [20], the authors observed that underrepresented groups corresponded to lower average activations. In the case of voice biometrics, the output of each network layer can be viewed as a bi-dimensional tensor of neurons over temporal frames:

$$A_{ij}^{[l]} = \Psi^{[l]}(\cdot) \tag{6}$$

where $i = 1, ..., N$, $j = 1, ..., M$, A_{ij} is the activation of the i^{th} neuron for the j^{th} temporal frame, $\Psi^{[l]}$ is the activation function at layer l, and N and M are the total number of neurons and frames respectively. For each layer l we calculate the root mean square of A_{ij} over the j^{th} frame which serves to account for large positive or negative activations. Then, we take the maximum along the i^{th} feature dimension:

$$\Lambda^{[l]} = \max_i \sqrt{\left(\frac{1}{M}\sum_j A_{ij}^2\right)} \tag{7}$$

The FAD is defined as the absolute difference between Λ for a pair of two distinct groups and is given by $FAD = |\Lambda_{d_1} - \Lambda_{d_2}|$. Near-zero values of FAD indicate better fairness.

C. Fine-tuning procedure

M_s, M_{sg} and M_{sga} models are fine-tuned as described in Section III-B. An initial warm-up is applied to the linear classification heads for the first $10k$ optimization steps, keeping the wav2vec 2.0 backbone frozen. The entire model is then fine-tuned in an end-to-end fashion for the remaining steps. We use the pre-trained model provided by Baevski et al. [17][3]. Performance for the speaker identification task exceeded 95% accuracy for all three models whereas the adversarial system delivered a gender recognition accuracy of only 47%.

D. Gender privacy threat models

The ability of the systems to conceal the gender information contained in its embeddings is measured by simulating the presence of a third party (an *attacker*) training a 2-layer fully-connected neural network $\mathcal{N}$ to infer the speaker gender from utterance embeddings. We consider two threat models. In the first one, the attacker is not aware that gender concealment has taken place (*uninformed attack* (uIA)) and therefore trains $\mathcal{N}$ on embeddings that are not gender-protected (in this case, those produced by M_s and M_{sg}). In the second one, the attacker is aware that model M_{sga} was used to protect the gender identity (*informed attack* (IA)), has access to that model, and trains $\mathcal{N}$ on embeddings produced by that same model. We expect this to result in a more effective attack.

[3]https://github.com/facebookresearch/fairseq/tree/main/examples/

			Models		
			M_s	M_{sg}	M_{sga}
EER(%)	Overall		2.36	3.23	3.89
	Male		3.12	4.22	4.98
	Female		3.05	4.21	5.26
auFDR	α	0	0.98	0.97	0.96
		0.25	0.97	0.97	0.95
		0.5	0.97	0.96	0.94
		0.75	0.96	0.95	0.92
		1	0.95	0.94	0.91

TABLE I
PERFORMANCE ANALYSIS OF THE THREE MODELS FOR UTILITY AND FAIRNESS, INCLUDING EER BREAKDOWN BY GENDER AND AUFDR ACROSS VARIOUS α VALUES (REFER TO EQ.5) FOR τ RANGING FROM 0.1% TO 10%.

	Data		Attack
	Training	Test	AUC (%)
uIA	M_s	M_s	97.09
	M_s	M_{sga}	**46.80**
	M_{sg}	M_{sg}	98.07
	M_{sg}	M_{sga}	**40.76**
IA	M_{sga}	M_{sga}	96.27

TABLE II
ASSESSMENT OF GENDER CONCEALMENT EFFECTIVENESS UNDER DIFFERENT THREAT SCENARIOS IN TERMS OF AUC.

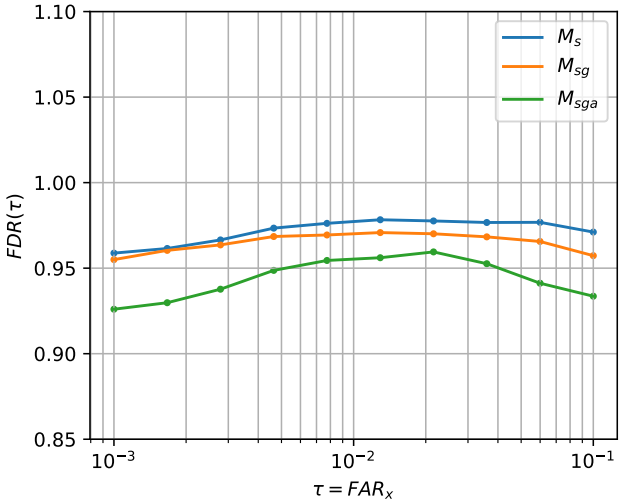

Fig. 2. FDR of different ASV systems for different decision thresholds for τ from 0.1% to 10%

V. EXPERIMENTAL RESULTS

We present results for each of the three models M_s, M_{sg}, and M_{sga}. Performance is assessed in terms of utility, privacy, and fairness.

In terms of utility, the performance of model M_s is in line with state-of-the-art automatic speaker verification systems, achieving an EER of 2.36% as shown in Table I. The performance of model M_{sg} and M_{sga} are slightly worse, 3.23% and 3.89% respectively, showing that gender influence does not improve speaker recognition. Furthermore, an analysis of

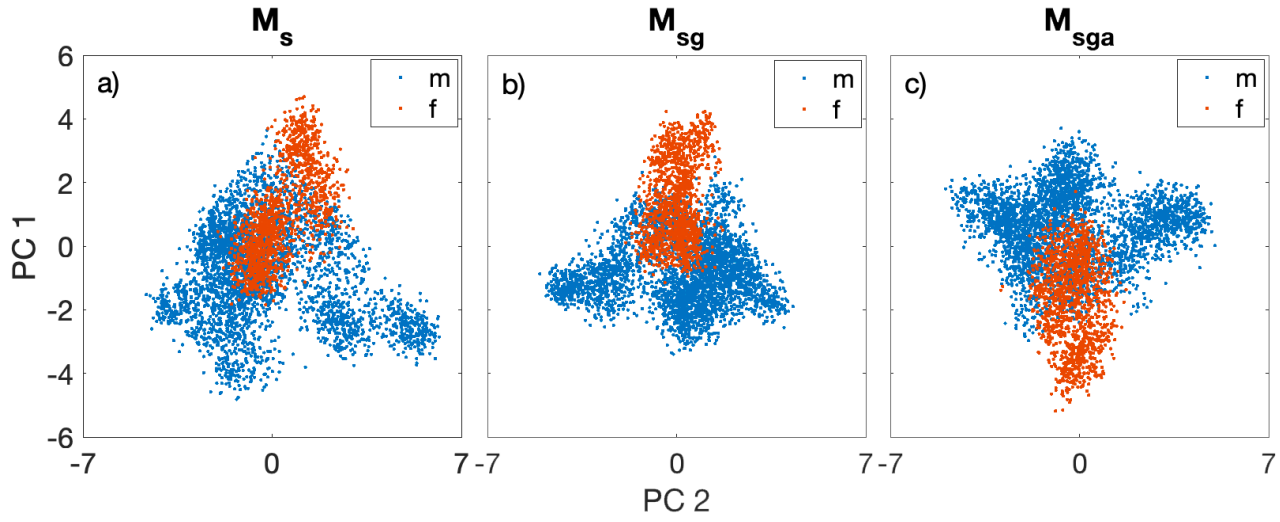

Fig. 3. PCA visualizations of features from three models illustrating gender recognition capabilities. Blue points correspond to males and red to females.

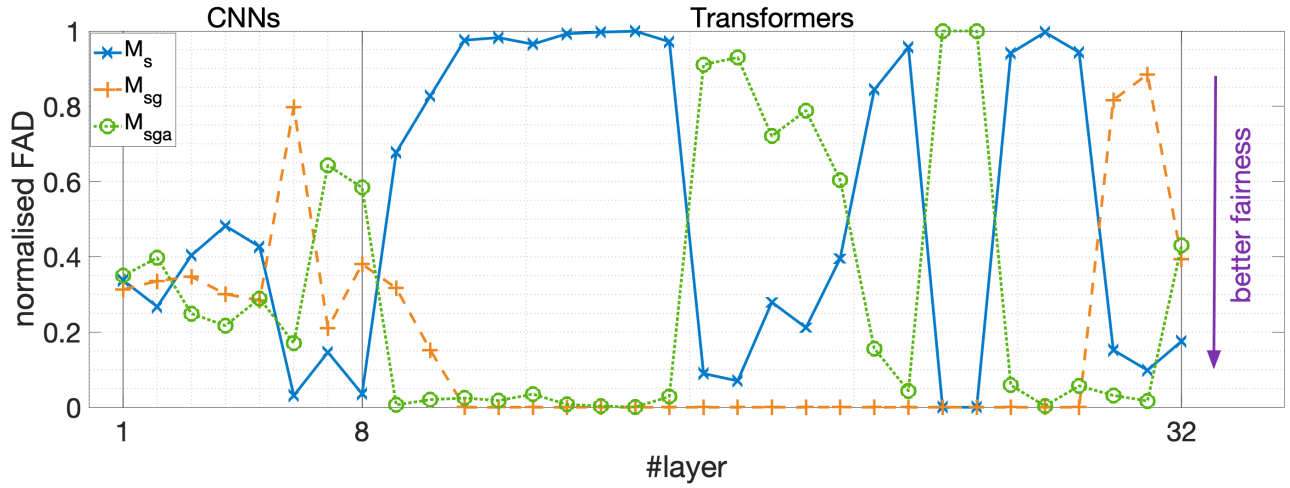

Fig. 4. Normalised Fairness Activation Discrepancy (FAD) of different systems at different wav2vec 2.0 module layers.

the EER broken down by gender shows small differences in speaker recognition for the two genders.

Fairness performances are shown at the bottom of the Table I in terms of the auFDR for different values of α. All auFDR results are close to 1, indicating reasonable fairness for each group. Fig. 2 depicts a plot of the FDR against the threshold for $\alpha = 0.5$. Profiles are shown for all three systems. The FDR is in all cases above 0.9, and the M_s system is always the fairest for each τ. Again, gender influence does not improve fairness.

Privacy performances are presented in Table II. AUC results for uninformed attacks (uIA) are shown at the top. When training and testing are performed using embeddings generated using the same, unprotected models, the AUC is 97.09% and 98.07% for M_s and M_{sg} models, respectively, demonstrating a lack of privacy protection. In contrast, when the same uninformed attack is made on the gender-protected model M_{sga}, the AUC drops to 46.80% and 40.76% respectively. This significant decrease indicates that the gender classifier predictions become nearly random, successfully concealing the gender information, demonstrating effective protection of privacy.

Performances for the informed attack (IA) are shown in the last row of Table II. When embeddings are extracted with the M_{sga} model, the AUC is much higher, at 96.27%. This result underlines the difficulty of obfuscating gender information from embeddings. Fig. 3 reveals an explanation. It illustrates a projection by principal component analysis of the embeddings generated by each of the three models. While the M_{sga} model is adversely trained with respect to gender cues, Fig. 3c shows that they persist. We see that, rather than fully obfuscating gender cues, M_{sga} only rotates the principal components hence why, when trained on similarly-treated training data, gender can still be recognised.

Finally, an analysis of internal bias in terms of FAD has been performed at different network layers considering male and

female groups. This analysis aims to provide insights into the comparative measures of fairness across three distinct models and how they dynamically propagate through the various layers. By examining the internal bias at each layer, we can better understand the impact of model architecture and training data on fairness outcomes. As illustrated in Fig. 4, 32 layers were selected in total from the wav2vec 2.0 model. These include 8 layers from the 1D-convolutional encoder and 24 intermediate activation layers from the Transformer modules.

Fig. 4 shows the FAD values calculated at different layers. The first layers of the CNNs display similar fairness, likely due to their focus on low-level features.

Contrastingly, Transformer layers, which handle high-level features, have wider fairness variations. M_s and M_{sga} show a complementary behavior as when one achieves high FAD, the other has lower FAD, and vice versa. This could be because M_s was fine-tuned for speaker verification, while M_{sga}, with its gradient reversal layer, was trying to suppress gender information. As layers progress, all models converge to FAD values, with M_s being the fairest at the end, confirming what is observed in terms of auFDR.

VI. Conclusions and Future Directions

This research explored the influence of gender information while fine-tuning wav2vec 2.0 for speaker verification. We proposed three models: M_s, M_{sg}, and M_{sga}, each with a different focus: speaker recognition, speaker recognition with gender classification, and speaker recognition with gender obfuscation, respectively. Our experiments revealed that M_s succeeds in speaker verification (EER of 2.36%), while M_{sga}, designed to hide gender information, performed much worse (EER of 3.89%). Interestingly, improving gender recognition in the M_{sg} model did not lead to better speaker verification performance (EER of 3.23%). Privacy evaluations showed effective gender obfuscation against uninformed attacks, but informed attackers could still extract gender information. Fairness evaluations, based on FDR, revealed that highlighting or hiding gender did not significantly impact the fairness of the systems. Furthermore, an analysis of FAD across model layers showed more disparities within Transformer layers, but all systems eventually converged to FAD values that match the auFDR assessment, with system M_s showing superior fairness.

In summary, while we achieved notable results in utility and privacy protection against uninformed attacks, future work includes strengthening gender obfuscation against informed attacks and enhancing fairness across systems.

VII. Acknowledgements

This work is supported by the TReSPAsS-ETN project funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 and partly supported by the VoicePersonae project funded by the French Agence Nationale de la Recherche (ANR) and the Japan Science and Technology Agency (JST).

References

[1] Baevski, Alexei; Zhou, Yuhao; Mohamed, Abdelrahman; Auli, Michael: wav2vec 2.0: A framework for Self-supervised learning of speech representations. In (Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.F.; Lin, H., eds): Advances in Neural Information Processing Systems. volume 33. Curran Associates, Inc., pp. 12449–12460, 2020.

[2] Benaroya, Laurent; Obin, Nicolas; Roebel, Axel, "Beyond Voice Identity conversion: manipulating voice attributes by adversarial learning of structured disentangled representations," arXiv preprint arXiv:2107.12346, 2021.

[3] Chouchane, Oubaïda; Panariello, Michele; Zari, Oualid; Kerenciler, Ismet; Chihaoui, Imen; Todisco, Massimiliano; Önen, Melek, "Differentially private adversarial auto-encoder to protect gender in voice biometrics", In: Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security. pp. 127–132, 2023.

[4] Chung, Joon Son; Nagrani, Arsha; Zisserman, Andrew, "Voxceleb2: deep speaker recognition", arXiv preprint arXiv:1806.05622, 2018.

[5] de Freitas Pereira, Tiago; Marcel, Sébastien, "Fairness in biometrics: a figure of merit to assess biometric verification systems", IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1):19–29, 2021.

[6] Fan, Zhiyun; Li, Meng; Zhou, Shiyu; Xu, Bo, "Exploring wav2vec 2.0 on speaker verification and language identification", arXiv preprint arXiv:2012.06185, 2020.

[7] Fenu, Gianni; Medda, Giacomo; Marras, Mirko; Meloni, Giacomo, "Improving fairness in speaker recognition", In: Proceedings of the 2020 European Symposium on Software Engineering. pp. 129–136, 2020.

[8] Ganin, Yaroslav; Ustinova, Evgeniya; Ajakan, Hana; Germain, Pascal; Larochelle, Hugo; Laviolette, François; March, Mario; Lempitsky, Victor, "Domain-adversarial training of neural networks", Journal of Machine Learning Research, 17(59):1–35, 2016.

[9] Hutiri, Wiebke Toussaint; Ding, Aaron Yi, "Bias in automated speaker recognition", In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 230–247, 2022.

[10] Hanani, Abualsoud; Russell, Martin J; Carey, Michael J, "Human and computer recognition of regional accents and ethnic groups from British English speech", Computer Speech & Language, 27(1):59–74, 2013.

[11] Jang, Eric; Gu, Shixiang; Poole, Ben, "Categorical reparameterization with Gumbel softmax", In: International Conference on Learning Representations. 2017.

[12] Jin, Minho; Ju, Chelsea J-T; Chen, Zeya; Liu, Yi-Chieh; Droppo, Jasha; Stolcke, Andreas, "Adversarial reweighting for speaker verification fairness", arXiv preprint arXiv:2207.07776, 2022.

[13] Lima, Lanna; Furtado, Vasco; Furtado, Elizabeth; Almeida, Virgilio," Empirical analysis of bias in voice-based personal assistants", In: Companion Proceedings of the 2019 World Wide Web Conference. pp. 533–538, 2019.

[14] Nagrani, Arsha; Chung, Joon Son; Zisserman, Andrew, "Voxceleb: a large-scale speaker identification dataset", arXiv preprint arXiv:1706.08612, 2017.

[15] Noé, Paul-Gauthier; Mohammadamini, Mohammad; Matrouf, Driss; Parcollet, Titouan; Nautsch, Andreas; Bonastre, Jean-François", Adversarial disentanglement of speaker representation for attribute-driven privacy preservation", arXiv preprint arXiv:2012.04454, 2020.

[16] Noé, Paul-Gauthier; Nautsch, Andreas; Matrouf, Driss; Bousquet, Pierre-Michel; Bonastre, Jean-François, "A bridge between features and evidence for binary attribute-driven perfect privacy", In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3094–3098, 2022.

[17] Ott, Myle; Edunov, Sergey; Baevski, Alexei; Fan, Angela; Gross, Sam; Ng, Nathan; Grangier, David; Auli, Michael, "Fairseq: a fast, extensible toolkit for sequence modeling", In: Proceedings of NAACL-HLT 2019: Demonstrations. 2019.

[18] Peri, Raghuveer; Somandepalli, Krishna; Narayanan, Shrikanth, "Study of bias mitigation strategies for speaker recognition", Computer Speech & Language, 79:101481, 2023.

[19] Shaqra, Ftoon Abu; Duwairi, Rehab; Al-Ayyoub, Mahmoud, "Recognizing emotion from speech based on age and gender using hierarchical models", Procedia Computer Science, 151:37–44, 2019.

[20] Serna, I.; Pena, A.; Morales, A.; Fierrez, J., "InsideBias: measuring bias in deep networks and application to face gender biometrics", In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE Computer Society, Los Alamitos, CA, USA, pp. 3720–3727, jan 2021.

[21] Shen, Hua; Yang, Yuguang; Sun, Guoli; Langman, Ryan; Han, Eunjung; Droppo, Jasha; Stolcke, Andreas, "Improving fairness in speaker verification via group-adapted fusion network", In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7077–7081, 2022.

[22] Solana-Lavalle, Gabriel; Rosas-Romero, Roberto, "Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation", Biomedical Signal Processing and Control, 66:102415, 2021.

[23] Toussaint, Wiebke; Ding, Aaron Yi, "Sveva fair: A framework for evaluating fairness in speaker verification", arXiv preprint arXiv:2107.12049, 2021.

[24] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Ł ukasz; Polosukhin, Illia, "Attention is All you Need", In (Guyon, I.; Luxburg, U. Von; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R.,eds): Advances in Neural Information Processing Systems. volume 30. Curran Associates, Inc., 2017.

[25] Vaessen, Nik; Van Leeuwen, David A, "Fine-tuning wav2vec2 for speaker recognition", In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7967–7971, 2022.

[26] Xiang, Xu; Wang, Shuai; Huang, Houjun; Qian, Yanmin; Yu, Kai: "Margin matters: towards more discriminative deep neural network embeddings for speaker recognition", In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1652–1656, 2019.

[27] Zaman, Syed Rohit; Sadekeen, Dipan; Alfaz, M Aqib; Shahriyar, Rifat, "One Source to detect them all: gender, age, and emotion detection from voice", In: 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). pp. 338–343, 2021.