Fair-Gate: Fairness-Aware Interpretable Risk Gating for Sex-Fair Voice Biometrics

Yangyang Qu
EURECOM
Sophia Antipolis, France
quy@eurecom.fr

Massimiliano Todisco
EURECOM
Sophia Antipolis, France
todisco@eurecom.fr

Chiara Galdi
EURECOM
Sophia Antipolis, France
galdi@eurecom.fr

Nicholas Evans
EURECOM
Sophia Antipolis, France
evans@eurecom.fr

***Abstract*—Voice biometric systems can exhibit sex-related performance gaps even when overall verification accuracy is strong. We attribute these gaps to two practical mechanisms: (i) demographic shortcut learning, where speaker classification training exploits spurious correlations between sex and speaker identity, and (ii) feature entanglement, where sex-linked acoustic variation overlaps with identity cues and cannot be removed without degrading speaker discrimination. We propose Fair-Gate, a fairness-aware and interpretable risk-gating framework that addresses both mechanisms in a single pipeline. Fair-Gate applies risk extrapolation to reduce variation in speaker-classification risk across proxy sex groups, and introduces a local complementary gate that routes intermediate features into an identity branch and a sex branch. The gate provides interpretability by producing an explicit routing mask that can be inspected to understand which features are allocated to identity versus sex-related pathways. Experiments on VoxCeleb1 show that Fair-Gate improves the utility–fairness trade-off, yielding more sex-fair ASV performance under challenging evaluation conditions.**

***Index Terms*—voice biometrics, speaker verification, fairness, sex bias, risk extrapolation, representation disentanglement**

I. Introduction

Recent progress in deep speaker embedding architectures has led to substantial gains in automatic speaker verification (ASV), a core technology for voice biometrics. Despite these improvements, ASV systems can still exhibit systematic performance differences across demographic groups, especially sex. Although subgroup performance may appear similar when each group is evaluated at its own operating point, such reporting can mask disparities that emerge when deploying a single global decision threshold shared by all users, which is the common practical setting [1], [2]. Recent studies have documented these disparities and examined how dataset composition, demographic imbalance, and evaluation design affect fairness, commonly defined as comparable verification error rates across groups under a shared operating point [1]–[8]. Importantly, these effects persist under standard protocols and operating points, suggesting systematic group-dependent model behavior rather than evaluation artifacts [1], [2].

A common mitigation strategy is to reduce the extent to which ASV embeddings encode sensitive attributes such as sex. This is often implemented through adversarial objectives or auxiliary prediction losses, sometimes combined with reweighting or group-aware fusion [9]–[14]. However, in ASV, sex-linked acoustic cues are not purely nuisance factors: pitch, timbre, and resonance correlate with sex but also carry identity-relevant information. Consequently, enforcing strong sex invariance can suppress useful speaker cues and degrade verification performance [11]. Related trade-offs have also been observed in privacy-oriented speaker anonymization and sex obfuscation, where reducing sex information can harm downstream utility [15]–[17]. This motivates a different goal: rather than enforcing global sex invariance, we seek to control where sex-linked variation is represented and how strongly it perturbs verification behavior under a shared decision threshold.

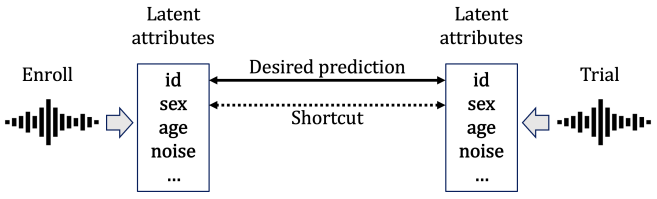

Fig. 1. Desired decision vs. demographic shortcut in speaker verification under a shared threshold. A verifier should base its decision on identity evidence by comparing enrollment and test utterances (solid arrow). However, because sex affects acoustics (e.g., F_0 and formant structure) and can be spuriously correlated with speaker identity in the training data, the model may also exploit sex-linked cues as a shortcut (dashed arrow). Such shortcut reliance can shift score distributions differently for male and female speakers, leading to subgroup error-rate gaps when deploying a single global decision threshold.

Inspired by causal fairness formulations [18], we distinguish (i) the causal effect of sex on speech acoustics (e.g., F_0 and formants) from (ii) dataset-induced correlations between sex and speaker identity in the training data. As illustrated in Fig. 1, standard discriminative training can exploit such correlations as demographic shortcuts, shifting score distributions differently across groups, and producing unequal subgroup error rates under a shared threshold.

This analysis highlights two fairness-relevant failure modes: (i) *shortcut learning*, where speaker classification relies on sex-conditioned correlations that do not generalize equally across groups; (ii) *feature entanglement*, where sex-related variation mixes with identity cues in the embedding used at inference, making it difficult to reduce subgroup error gaps

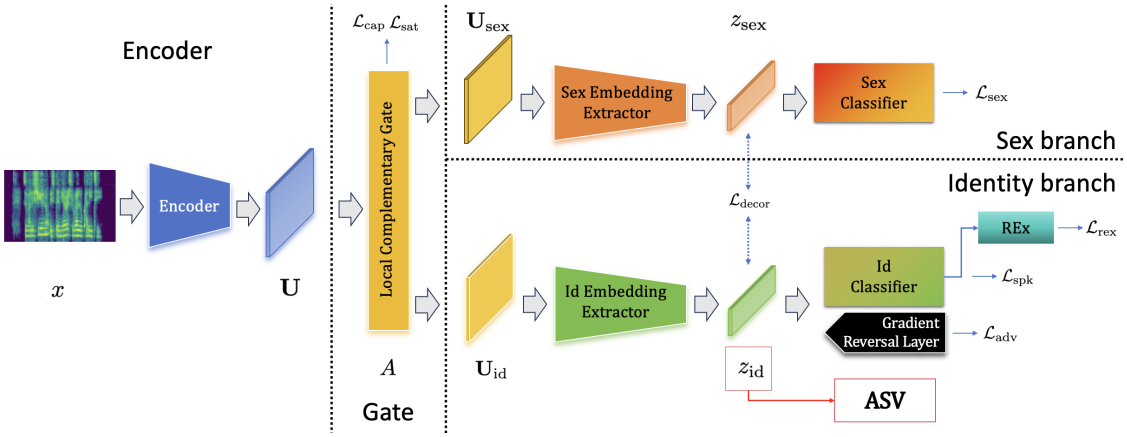

Fig. 2. Overview of Fair-Gate. The encoder produces frame-level features $\mathbf{U}$, which are complementarily soft-routed by a local mask A (gate) into an identity branch and a sex branch. The identity branch produces the embedding z_{id}, which is the only representation used for automatic speaker verification (ASV) at inference. During training, the sex branch learns a sex embedding z_{sex} and predicts proxy sex labels $\hat{s}$ via a sex classifier ($\mathcal{L}_{\text{sex}}$). The identity branch is optimized for speaker classification ($\mathcal{L}_{\text{spk}}$), regularized by Risk Extrapolation (REx) across proxy sex groups ($\mathcal{L}_{\text{rex}}$), and constrained by an adversarial sex classifier implemented through a Gradient Reversal Layer (GRL) ($\mathcal{L}_{\text{adv}}$). A decorrelation loss ($\mathcal{L}_{\text{decor}}$) encourages separation between z_{id} and z_{sex}, while gate regularizers ($\mathcal{L}_{\text{cap}}$, $\mathcal{L}_{\text{sat}}$) prevent degenerate routing.

without sacrificing verification performance.

Motivated by these mechanisms, we propose **Fair-Gate**, a unified training framework that improves the utility–fairness trade-off by addressing shortcut learning and feature entanglement in a single pipeline. In the following, we use the term *utility* to refer specifically to verification performance, in order to clearly distinguish it from fairness performance. First, we apply Risk Extrapolation (REx) [19] across proxy sex groups to penalize differences in speaker-classification risk between groups, discouraging group-specific shortcuts. Second, we introduce a complementary local soft-routing gate that partitions intermediate features into two additive components routed to an identity branch and a sex branch, while preserving the original feature dimensionality. The sex branch provides an explicit pathway for sex-linked variation during training, reducing its leakage into the identity embedding used for verification at inference.

Our contributions are:

- We provide a causal analysis of sex-related bias in ASV, separating inherent sex-linked acoustic variation from dataset-induced correlations.
- We propose Fair-Gate, combining Risk Extrapolation (REx) across proxy sex groups with a complementary local gating mechanism and branch-specific objectives to limit sex leakage into the deployed embedding.
- We demonstrate improved utility–fairness trade-offs on VoxCeleb, and provide ablations that clarify the roles of complementary routing and branch-specific training objectives.

Despite sex and gender terms being commonly used interchangeably, there is a generally, though not universally accepted distinction. We adopt the terminology in [20] and expressly avoid references to gender which refers to socially constructed roles and behaviour. Instead, we refer only to sex, which refers to biological attributes [21].

II. Fair-Gate Framework

Fair-Gate extends a standard ECAPA-style speaker verification pipeline with complementary feature routing and fairness-aware training objectives. The framework illustrated in Fig. 2 consists of three key components: (i) a shared encoder that extracts frame-level representations, (ii) a local complementary gate that routes these representations into an identity branch and a sex branch, and (iii) branch-specific objectives, including risk variance equalization, that promote speaker discrimination while reducing sex-dependent disparities under a shared operating threshold.

At inference, only the identity branch is retained for verification. Given an enrollment–test pair (x_a, x_b), the verification score is computed using cosine similarity between identity embeddings:

$$\text{score}(x_a, x_b) = \cos\big(z_{\text{id}}(x_a), z_{\text{id}}(x_b)\big). \tag{1}$$

A single shared threshold is then applied across all users.

A. Encoder

Given an input log-Mel spectrogram $x \in \mathbb{R}^{F\times T}$, where F is the number of Mel-frequency bins and T the number of frames, the encoder $E(\cdot)$ produces frame-level features in $\mathbf{U} \in \mathbb{R}^{C\times T}$, where C denotes the number of feature channels.

B. Local Complementary Gating

The local complementary gate (center of Fig. 2) softly allocates intermediate features between an identity branch and a sex branch while preserving dimensionality. Unlike global channel attention mechanisms, the gate operates at each time–channel location, enabling fine-grained routing. Preserving dimensionality is important here because the goal

is information reallocation rather than compression: the gate should redistribute partially entangled cues between branches without imposing a fixed feature split or discarding speaker-discriminative content a priori.

1) Mask computation and routing: Given frame-level features $\mathbf{U}$, we compute a soft mask

$$A = \sigma\big(\mathrm{DWConv}_t(\mathbf{U})\big), \tag{2}$$

where DWConv_t is a depthwise temporal convolution applied independently per channel and configured to preserve temporal length, and $\sigma(\cdot)$ is the sigmoid. Features are routed complementarily as

$$\mathbf{U}_{\mathrm{id}} = A \odot \mathbf{U}, \qquad \mathbf{U}_{\mathrm{sex}} = (1 - A) \odot \mathbf{U}. \tag{3}$$

Both branches retain dimensionality and satisfy $\mathbf{U}_{\mathrm{id}} + \mathbf{U}_{\mathrm{sex}} = \mathbf{U}$ element-wise. This additive decomposition ensures that the gate reallocates information between branches, allowing Fair-Gate to learn where information should be represented rather than forcing identity- and sex-related cues into fixed disjoint subspaces.

2) Gate regularization: To avoid degenerate routing (e.g., collapsing all features to one branch or producing ambiguous allocations), we regularize the mask A through two complementary terms defined below: a routing-mass control term $\mathcal{L}_{\mathrm{cap}}$ and a saturation term $\mathcal{L}_{\mathrm{sat}}$.

3) Routing mass control: Let

$$\bar{a} = \frac{1}{BCT} \sum_{i,c,t} A_{i,c,t}.$$

We regulate the average routing mass via

$$\mathcal{L}_{\mathrm{cap}} = (\bar{a} - \rho_{\mathrm{id}})^2,$$

where ρ_{id} controls the desired proportion of features allocated to the identity branch.

4) Saturation constraint: To encourage confident (near-binary) routing,

$$\mathcal{L}_{\mathrm{sat}} = \frac{1}{BCT} \sum_{i,c,t} A_{i,c,t}(1 - A_{i,c,t}).$$

C. Identity and Sex Branches Objectives

Each routed feature sequence is mapped to an utterance-level embedding using a branch-specific embedding extractor (Fig. 2):

$$z_{\mathrm{id}} = f_{\mathrm{spk}}(\mathbf{U}_{\mathrm{id}}), \qquad z_{\mathrm{sex}} = f_{\mathrm{sex}}(\mathbf{U}_{\mathrm{sex}}).$$

Here, $f_{\mathrm{spk}}(\cdot)$ and $f_{\mathrm{sex}}(\cdot)$ denote the identity- and sex-branch utterance-level embedding extractors, respectively. The identity extractor uses attentive statistics pooling [22] to form z_{id}, which is the only embedding used for ASV at inference. The sex branch is used only during training to capture sex-related variation and reduce its leakage into z_{id}.

We do not use human-annotated sex labels. Instead, we obtain binary proxy sex labels $\hat{s} \in \{\mathrm{M}, \mathrm{F}\}$ from a frozen pre-trained classifier and use them only during training. These proxy labels supervise the sex-classification objective, the adversarial sex objective, and the proxy-group partition used by REx. At inference, neither the proxy labels nor the sex branch is required.

Fair-Gate jointly optimizes speaker discrimination and fairness through branch-specific objectives. We denote cross-entropy by $\mathrm{CE}(\cdot, \cdot)$.

1) Speaker classification: The identity embedding is trained with an AAM-Softmax classifier:

$$\mathcal{L}_{\mathrm{spk}} = \mathbb{E}_{(x,y)} \mathrm{CE}\big(h_{\mathrm{spk}}(z_{\mathrm{id}}(x)), y\big), \tag{4}$$

where y denotes the speaker label and $h_{\mathrm{spk}}(\cdot)$ the speaker-classification head.

2) Adversarial constraint: To discourage encoding of sex information in z_{id}, we attach an adversarial sex classifier via a Gradient Reversal Layer (GRL) [9]:

$$\mathcal{L}_{\mathrm{adv}} = \mathbb{E}_{(x,\hat{s})} \mathrm{CE}\big(h^{\mathrm{adv}}_{\mathrm{sex}}(\mathrm{GRL}_\gamma(z_{\mathrm{id}}(x))), \hat{s}\big). \tag{5}$$

Here, $h^{\mathrm{adv}}_{\mathrm{sex}}(\cdot)$ denotes the adversarial sex-classification head and $\mathrm{GRL}_\gamma(\cdot)$ a gradient reversal layer with reversal strength γ. Together, $\mathcal{L}_{\mathrm{spk}}$ and $\mathcal{L}_{\mathrm{adv}}$ encourage identity discrimination while reducing direct sex predictability in the deployed embedding.

3) Sex classification: The sex branch is trained to explicitly capture sex-related variation:

$$\mathcal{L}_{\mathrm{sex}} = \mathbb{E}_{(x,\hat{s})} \mathrm{CE}\big(h_{\mathrm{sex}}(z_{\mathrm{sex}}(x)), \hat{s}\big), \tag{6}$$

where $h_{\mathrm{sex}}(\cdot)$ denotes the sex-classification head attached to z_{sex}.

4) Embedding decorrelation: To further reduce information overlap between branches, we penalize similarity between normalized embeddings:

$$\mathcal{L}_{\mathrm{decor}} = \mathbb{E}_x \big\langle \bar{z}_{\mathrm{id}}(x), \bar{z}_{\mathrm{sex}}(x) \big\rangle^2, \quad \bar{z} = \frac{z}{\|z\|_2}. \tag{7}$$

This term does not enforce full statistical independence; rather, it discourages direct overlap between the two embeddings and complements the routing and adversarial objectives.

5) Risk Variance Equalization: Risk Extrapolation (REx) reduces discrepancies in speaker-classification risk loss across proxy sex groups. Let $\mathcal{E} = \{\mathrm{M}, \mathrm{F}\}$. The per-group risk is

$$\mathcal{R}_e = \mathbb{E}_{(x,y)\sim\mathcal{E}} \mathrm{CE}\big(h_{\mathrm{spk}}(z_{\mathrm{id}}(x)), y\big), \qquad \bar{\mathcal{R}} = \tfrac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} \mathcal{R}_e. \tag{8}$$

The REx penalty is

$$\mathcal{L}_{\mathrm{rex}} = \tfrac{1}{|\mathcal{E}|} \sum_{e\in\mathcal{E}} (\mathcal{R}_e - \bar{\mathcal{R}})^2. \tag{9}$$

Intuitively, if the speaker classifier relies on group-specific shortcuts, speaker-classification risk will differ systematically across proxy sex groups. Penalising the variance of $\{\mathcal{R}_e\}$, therefore, encourages the model to rely on speaker evidence that transfers more uniformly across groups.

In practice, group risks are estimated within mini-batches, and the penalty is applied only when both groups are sufficiently represented in the batch.

TABLE I
VOXCELEB1 VERIFICATION PROTOCOLS USED FOR EVALUATION.

Protocol	Speakers	# Trials	Description
Vox1-O	40	37,720	Original test list
Vox1-E	1,251	581,480	Expanded trial list
Vox1-H	1,251	552,536	Hard (same nationality & sex impostors)

D. Overall Training Objective

The full objective combines utility (i.e. verification performance), fairness, routing, and separation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spk}} + \lambda_s \mathcal{L}_{\text{sex}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{decor}} \mathcal{L}_{\text{decor}} + \lambda_{\text{cap}} \mathcal{L}_{\text{cap}} + \lambda_{\text{sat}} \mathcal{L}_{\text{sat}} + \lambda_{\text{rex}} \mathcal{L}_{\text{rex}}. \quad (10)$$

III. EXPERIMENTAL SETUP AND METRICS

A. Datasets and Protocols

We train all models on the VoxCeleb2 development set [25], [26], which contains over one million utterances from more than 6,000 speakers collected in unconstrained, real-world conditions. Evaluation is performed on VoxCeleb1 [26], [27] using the official verification protocols Vox1-O, Vox1-E, and Vox1-H, ensuring comparability with prior ASV and fairness studies.

Vox1-O is the *O*riginal test list, Vox1-E is an *E*xpanded trial set, and Vox1-H is a *H*ard protocol in which non-mated trials are constructed from speakers matched in nationality and proxy sex. These protocols differ in the number and difficulty of verification trials, as summarized in Table I.

B. Metrics and baselines

The evaluation of the speaker verification system for both utility and fairness is reported in terms of classification error analysis. In particular we observe:

- False match rate (FMR): proportion of the completed biometric non-mated comparison trials that result in a false match;
- False non-match rate (FNMR): proportion of the completed biometric mated comparison trials that result in a false non-match;

For the sake of comparison with the state of the art, system's utility is reported in terms of the equal error rate (EER, where FMR=FNMR) and minimum detection cost function (minDCF) [28] as well:

$$\begin{aligned} \mathrm{DCF}(\tau) &= C_{\mathrm{FNMR}}\, P_{\mathrm{target}}\, \mathrm{FNMR}(\tau) \\ &\quad + C_{\mathrm{FMR}}\, (1 - P_{\mathrm{target}})\, \mathrm{FMR}(\tau), \\ \mathrm{minDCF} &= \min_{\tau}\, \mathrm{DCF}(\tau). \end{aligned} \quad (11)$$

where τ is the decision threshold and $\mathrm{FMR}(\tau)$ / $\mathrm{FNMR}(\tau)$ are computed from pooled trials (male and female speakers); P_{tar} is the prior probability of a genuine attempt (target); and C_{FNMR} and C_{FMR} are weighting parameters that adjust the relative importance of false non-matches and false matches. We use the standard NIST SRE setting with $P_{\mathrm{tar}} = 0.01$ and $C_{\mathrm{FNMR}} = C_{\mathrm{FMR}} = 1$. Unlike the EER, where equal error importance is assumed and no prior information is provided, the minDCF measures the minimum expected application cost by incorporating error costs and class priors, and selecting the optimal threshold for that specific operating scenario.

Following prior ASV fairness evaluations [7], we report fairness using GARBE [29] at a fixed operating point specified by a threshold τ.

GARBE is defined as follows:

$$\mathrm{GARBE}(\tau) = \alpha\, G_{\mathrm{FMR}_\tau} + (1 - \alpha)\, G_{\mathrm{FNMR}_\tau} \quad (12)$$

where G_{FMR_τ} and G_{FNMR_τ} are the Gini coefficients computed over subgroup FMRs and FNMRs at threshold τ, respectively. We set $\alpha = 0.5$ so that FMR and FNMR disparities are equally weighted.

While identical thresholds are used for male and female score distributions for all experiments, different thresholds τ are computed according to the system evaluated (baselines and proposed) and the metric. For EER assessment, τ corresponds to the decision threshold where the system achieves $FMR = FNMR$. For the computation of $minDCF$, τ is selected such that DCF is minimised. Finally, for fairness, the threshold for GARBE is fixed at $\tau = 1\%$, such that $FMR(\tau) = 10^{-2}$, which we indicate in the following with $\mathrm{GARBE}(\tau_{1\%})$.

We compare our approach against three representative baselines covering utility-oriented training, adversarial invariance learning, and disentanglement-based modeling:

1) **ECAPA-TDNN** [23]. We adopt ECAPA-TDNN as a strong utility-oriented backbone. ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network) enhances the conventional TDNN architecture through channel-dependent frame attention, Res2Net multi-scale feature extraction, and attentive statistical pooling. These components improve the modeling of speaker-discriminative information while maintaining robustness to channel and acoustic variability. In our comparison, this model serves as the primary utility-driven baseline, optimized solely for speaker verification without any explicit fairness or invariance constraint.
2) **ECAPA-TDNN + GRL**. To provide an adversarial invariance baseline, we augment ECAPA-TDNN with a gradient reversal layer (GRL) [9] to the speaker embedding. A sex classifier is trained adversarially through the GRL to discourage the encoding of sex-related information in the embedding space. During training, the speaker classification objective and the adversarial sex classification objective are jointly optimized, with the GRL reversing the gradient from the adversary to promote invariance. This setup follows the standard domain-adversarial learning paradigm and represents a commonly used strategy for mitigating demographic leakage while preserving speaker verification utility.

TABLE II
RESULTS ON VOXCELEB1 (O/E/H). WE REPORT EER [%], MINDCF, AND SEX FAIRNESS MEASURED BY GARBE AT THE THRESHOLD $\tau_{1\%}$.

Model	Vox1-O			Vox1-E			Vox1-H		
	EER ↓	minDCF ↓	GARBE ↓	EER ↓	minDCF ↓	GARBE ↓	EER ↓	minDCF ↓	GARBE ↓
ECAPA-TDNN [23]	1.12	0.14	**0.16**	1.34	0.17	0.11	2.62	**0.25**	0.10
ECAPA-TDNN + GRL	0.98	0.13	0.22	1.25	0.14	0.12	2.57	0.28	0.10
VoxDisentangler [24]	**0.82**	0.12	0.17	1.15	0.14	0.11	2.40	0.26	0.10
Fair-Gate (ours)	0.92	**0.11**	0.26	**1.11**	**0.14**	**0.05**	**2.25**	0.26	**0.07**

TABLE III
ABLATION RESULTS ON VOX1-H. WE REPORT EER [%], $\mathrm{GARBE}(\tau_{1\%})$, AND SUBGROUP-SPECIFIC FNMR/FMR [%] AT THE SAME SHARED THRESHOLD $\tau_{1\%}$. BOLD INDICATES THE BEST VALUE IN EACH COLUMN.

Setting	Cap	Sat	Gs	Adv	REx	EER↓	GARBE↓	FNMR_M	FNMR_F	FMR_M	FMR_F
Main w/o Cap	–	✓	✓	✓	✓	2.66	0.09	1.03	**0.95**	4.76	6.30
Main w/o Sat	✓	–	✓	✓	✓	2.30	0.07	**0.95**	1.07	3.98	4.68
Main w/o Gs	✓	✓	–	✓	✓	2.66	0.09	1.03	**0.95**	4.76	6.30
Main w/o Adv	✓	✓	✓	–	✓	2.27	**0.07**	0.96	1.06	3.81	**4.48**
Main w/o REx	✓	✓	✓	✓	–	2.55	0.08	0.96	1.06	4.54	5.60
Main, Full	✓	✓	✓	✓	✓	**2.25**	**0.07**	0.96	1.07	**3.80**	4.49

3) **VoxDisentangler** [24]. VoxDisentangler is a disentanglement-based speaker verification framework that explicitly separates speaker identity from environmental variability (e.g., channel and noise conditions) into distinct latent representations. Shortcut learning in ASV is not limited to demographic attributes: models can also exploit spurious correlations between speaker identity and nuisance factors such as channel, background noise, or recording conditions. Although VoxDisentangler is not designed to disentangle demographic attributes such as sex, we include it as a baseline to assess whether disentangling non-identity nuisance factors can indirectly mitigate sex-dependent performance disparities.

All methods are trained on the VoxCeleb2 development set and evaluated on VoxCeleb1 using the original (O), extended (E), and hard (H) trial protocols. To ensure a fair comparison, we follow identical training data splits, preprocessing steps, and evaluation metrics across all systems. Unless otherwise stated, all reported experiments use a batch size of 512.

IV. RESULTS

A. Results on VoxCeleb1-O/E/H.

Results are presented in Table II. GARBE results show that Fair-Gate improves sex-group fairness for Vox1-E/H while maintaining competitive utility. For VoxCeleb1-O, Fair-Gate yields a larger sex disparity compared to competing approaches. For the smallest and least challenging O protocol, subgroup disparity estimates are more sensitive to trial composition and the precise operating point. In contrast, for the extended and harder Vox1-E and Vox1-H protocols, Fair-Gate consistently reduces sex disparity while maintaining strong utility, suggesting that risk equalization and the complementary gating mechanisms are beneficial under more challenging evaluation conditions where sex-dependent shortcut cues are more likely to be exploited.

For Vox1-E, Fair-Gate achieves the best fairness score of $\mathrm{GARBE}(\tau_{1\%}) = 0.05$, substantially lower than for ECAPA (0.11), GRL (0.12), and VoxDisentangler (0.11). Improved fairness comes with improved utility: Fair-Gate reduces the EER to 1.11% (vs. 1.34% ECAPA, 1.25% GRL, 1.15% VoxDisentangler), and delivers a minDCF of 0.14 (vs. 0.17 for ECAPA, and the same results of 0.14 for GRL/VoxDisentangler).

For Vox1-H, Fair-Gate achieves the best EER and $\mathrm{GARBE}(\tau_{1\%})$ results, while the minDCF is comparable to that of the strongest baselines. Notably, the GRL baseline does not improve fairness over ECAPA: for Vox1-E, GARBE increases from 0.11 to 0.12 whereas, for Vox1-H, the result of 0.10 suggests that adversarial invariance alone is insufficient to equalise subgroup error rates for a common operating point. Overall, Fair-Gate offers the best fairness for Vox1-E/H demonstrating a stronger utility–fairness trade-off than the baselines.

B. Ablation study on VoxCeleb1-H

Ablation results on Vox1-H are shown in Table III. Overall, the selected Fair-Gate configuration with moderate REx ($r = 0.005$) provides the most favorable utility–fairness trade-off. It achieves the best EER (2.25%) while remaining highly competitive on $\mathrm{GARBE}(\tau_{1\%})$ and subgroup-specific error rates. The largest degradation is observed when removing either the routing-mass regularizer or the sex-branch supervision: both *w/o Cap* and *w/o Gs* worsen from 2.25% / 0.07 to 2.66% /

0.09 in EER / GARBE, and substantially increase subgroup FMRs, especially for the female subgroup. This suggests that the main subgroup disparity on Vox1-H is driven primarily by the false-match side, and that complementary routing control together with explicit sex-branch supervision is important for limiting subgroup-dependent shortcut reliance.

The remaining ablations indicate that the other components play more limited or supporting roles. Removing REx still degrades the trade-off, increasing EER from 2.25% to 2.55% and GARBE from 0.07 to 0.08, while also raising subgroup FMRs from 3.80% / 4.49% to 4.54% / 5.60%, which shows that risk equalization contributes meaningfully under the shared operating point. By contrast, removing the adversarial term changes fairness only marginally but slightly worsens utility (2.25% $\rightarrow$ 2.27%), suggesting that adversarial invariance is not the primary source of subgroup-gap reduction.

V. Conclusions

We present Fair-Gate, a fairness-oriented training framework for speaker verification that reduces discrepancies in sex-dependent error rates at common operating points. Fair-Gate combines risk variance equalization across proxy sex groups to discourage group-specific shortcuts during speaker classification with complementary local gating to explicitly route intermediate features into identity and sensitive pathways, restricting the leakage of sex-linked variation into speaker embeddings. Results derived using the VoxCeleb1 database shows that Fair-Gate improves the utility–fairness trade-off, with the clearest gains being achieved for the most challenging protocols. Future work should explore more reliable proxy-group construction, extend the framework to additional sensitive attributes, and evaluate robustness under cross-corpus shifts and broader deployment conditions.

Acknowledgement

This work was supported by the French Agence Nationale de la Recherche (ANR) via the SpeechPrivacy (ANR-23-CE23-0022) and GOOD-BIAS (ANR-25-CE39-6459) projects.

References

[1] W. Toussaint and A. Y. Ding, "Sveva fair: A framework for evaluating fairness in speaker verification," *arXiv preprint arXiv:2107.12049*, 2021.

[2] W. Hutiri, T. Patel, A. Y. Ding, and O. Scharenborg, "As biased as you measure: Methodological pitfalls of bias evaluations in speaker verification research," in *Proc. Interspeech*, 2024.

[3] G. Fenu, M. Marras, G. Medda, and G. Meloni, "Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition," in *Proc. Interspeech*, 2021, pp. 1892–1896.

[4] X. Chen, Z. Li, S. Setlur, and W. Xu, "Exploring racial and gender disparities in voice biometrics," *Scientific Reports*, vol. 12, no. 1, p. 3723, 2022.

[5] A. Hajavi and A. Etemad, "A study on bias and fairness in deep speaker recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[6] M. Estévez and L. Ferrer, "Study on the fairness of speaker verification systems across accent and gender groups," in *Proc. ICASSP*, 2023, pp. 1–5.

[7] O. Chouchane, C. Busch, C. Galdi, N. Evans, and M. Todisco, "A comparison of differential performance metrics for the evaluation of automatic speaker verification fairness," in *Proc. ODYSSEY*, 2024.

[8] W. Hutiri, L. Gorce, and A. Y. Ding, "Design guidelines for inclusive speaker verification evaluation datasets," in *Proc. Interspeech*, 2022, pp. 1293–1297.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[10] G. Bhattacharya, J. Alam, and P. Kenny, "Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training," in *Proc. ICASSP*, 2019, pp. 6041–6045.

[11] R. Peri, K. Somandepalli, and S. Narayanan, "A study of bias mitigation strategies for speaker recognition," *Computer Speech & Language*, vol. 79, p. 101481, 2023.

[12] 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.

[13] 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," in *Proc. ICASSP*, 2022.

[14] O. Chouchane, M. Panariello, C. Galdi, M. Todisco, and N. Evans, "Fairness and privacy in voice biometrics: A study of gender influences using wav2vec 2.0," in *Proc. BIOSIG*, 2023.

[15] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé *et al.*, "Introducing the voiceprivacy initiative," in *Proc. Interspeech*, 2020.

[16] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[17] Y. Qu, M. Panariello, M. Todisco, and N. Evans, "Reference-free adversarial sex obfuscation in speech," in *APSIPA 2025, 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2025.

[18] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[19] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International conference on machine learning*. PMLR, 2021, pp. 5815–5826.

[20] P.-G. Noe, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," in *Proc. Interspeech*, 2021.

[21] V. Prince, "Sex vs. gender," *International Journal of Transgenderism*, vol. 8, no. 4, pp. 29–32, 2005.

[22] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[24] K. Nam, H.-S. Heo, J.-w. Jung, and J. Chung, "Disentangled representation learning for environment-agnostic speaker recognition," in *Proc. Interspeech 2024*, 2024, pp. 2130–2134.

[25] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[26] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[27] A. Nagraniy, J. S. Chungy, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[28] A. F. Martin and C. S. Greenberg, "The nist 2010 speaker recognition evaluation." in *Interspeech*, 2010, p. 2726.

[29] J. J. Howard, E. J. Laird, R. E. Rubin, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Evaluating proposed fairness models for face recognition algorithms," in *Proc. ICPR*, 2022, pp. 431–447.