2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389804

MODEL-BASED FAIRNESS METRIC FOR SPEAKER VERIFICATION

Maliha Jahan, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak, Jesús Villalba

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Ensuring that technological advancements benefit all groups of people equally is crucial. The first step towards fairness is identifying existing inequalities. The naive comparison of group error rates may lead to wrong conclusions. We introduce a new method to determine whether a speaker verification system is fair toward several population subgroups. We propose to model miss and false alarm probabilities as a function of multiple factors, including the population group effects, e.g., male and female, and a series of confounding variables, e.g., speaker effects, language, nationality, etc. This model can estimate error rates related to a group effect without the influence of confounding effects. We experiment with a synthetic dataset where we control group and confounding effects. Our metric achieves significantly lower false positive and false negative rates w.r.t. baseline. We also experiment with VoxCeleb and NIST SRE21 datasets on different ASV systems and present our conclusions.

Index Terms— Fairness, Speaker verification

1. INTRODUCTION

The increasing demand for smart devices has highlighted the importance of advancing speech-processing technologies, such as automatic speech recognition (ASR) and speaker verification (ASV). As these technologies evolve, it is crucial to ensure no disparity in the performance among different speaker subgroups. Speaker subgroups refer to distinct groups of speakers based on factors such as gender, age, nationality, etc. Recent studies [1] [2] [3] have identified negative biases towards certain population subgroups in the performance of ASV systems trained on established datasets.

Hutiri et al. [1] present an in-depth analysis of bias present in the VoxCeleb Speaker Recognition Challenge benchmark. They proposed a new metric called *sub-group bias* to evaluate the bias using detection cost C_{Det} [4]. *sub-group bias* is the ratio between the detection cost of a subgroup and the overall detection cost. The authors uncovered that female speakers and speakers of non-US nationalities experience notable performance degradation. Another recent work by Estevez et al. [5] shows similar performance issues. They used C_{llr} [6], a calibration-sensitive metric, to measure the ASV system performance and Fairness Discrepancy Rate (FDR) [7] to measure fairness. Their results show performance gaps between different nationality and gender groups for systems trained with unbalanced backends.

However, the previous works [1] [5] on fairness in ASV systems fail to consider the potential influence of confounding factors, such as imbalances in the speakers' age or nationality distribution when gender fairness is being analyzed. Also, an oversight of speaker variability within the respective groups may result in a wrong conclusion regarding fairness.

We propose a new method to detect unfair bias in ASV systems while taking into account various confounding factors and speaker variabilities. Our work is primarily motivated by a study conducted by Liu et al. [8]. Their research introduced a model-based method for assessing fairness in ASR systems. They proposed mixed-effects Poisson regression [9] using speaker labels as random effects and any group label or confounding variables as fixed effects. The authors validated the efficacy of the proposed model through a series of comprehensive experiments using synthetic data. The contributions of our work are: (1) introducing a new model-based method to detect unfair biases in ASV systems; (2) generating a synthetic dataset to validate our method; (3) performing fairness analysis on two real datasets and presenting our findings.

2. MODEL-BASED METRICS

2.1. Bernoulli Regression to Measure ASV Fairness

Drawing inspiration from Liu et al.'s [8] fairness metric for ASR, we propose a model that can disentangle the effects of different factors on the error rates of ASV systems. Errors in speaker verification trials depend on several *factors*, such as the speaker's gender, language, accent, utterance duration, transmission channel, etc. Meanwhile, *groups* define the specific category of a factor the speaker belongs to. For instance, *male* and *female* could be the groups under the *gender* factor.

Suppose we want to assess the fairness of ASV regarding some factor of interest, e.g., the speakers' gender. Given a speaker verification trial s^{th}, we define the *groups* of a factor with a deterministic function $f(s)$, which could take three values for the *gender* factor, as explained later. We intend to test this factor's statistical significance across its different groups on the measured ASV metrics. In this case, the other factors affecting the error rate, like language or channel, are confounding factors whose effects we need to remove. Suppose all utterances had low SNRs for the male speakers and

This project was supported by NSF Award 2147350

high SNRs for the female speakers. If we do not remove the SNR effect, we will conclude that the system performs better for females than males. Likewise, if we want to assess fairness w.r.t. language, gender becomes a confounding factor.

A particularity of speaker verification is that metrics, such as equal error rate (EER) and detection cost function (DCF) [4], depend on two error probabilities, namely miss and false alarm rates. Misses are errors for target trials, while false alarms are errors for non-target trials. The trial characteristic factors impact differently to target and non-target trials. For example, channel mismatch between enrollment and test utterances may increase misses but decrease false alarms. Therefore, we need separate models with different parameters to estimate the factors' effects on misses and false alarms.

We define a binary variable t_s, which indicates whether the s trial is wrongly classified ($t_s = 1$) or not ($t_s = 0$). We assume that t_s is distributed as,

$$t_s \sim \text{Bernoulli}(q(s)) \quad \text{where} \tag{1}$$

$$q(s) = \begin{cases} h(\lambda_{\mathcal{T}}(s)) & \text{where s is a target trial} \\ h(\lambda_{\mathcal{N}}(s)) & \text{where s is a non-target trial} \end{cases} . \tag{2}$$

where h is a non-linear function that maps λ into the $[0, 1]$ interval. We can use $h(\lambda) = \text{sigmoid}(\lambda)$ or $h(\lambda) = \exp(-\exp(-\lambda))$, similar to [8]. Thus, the error probability is trial-dependent and is given by a regression function. The λ's are linear functions of the effects of the different factors,

$$\lambda_Z(s) = \mu_Z + \sum_{k=1}^{K} \mu_{Z,k}(f_k(s)) \quad \text{for } Z \in \mathcal{T}, \mathcal{N} \tag{3}$$

where $\mu_{\mathcal{T}}$ and $\mu_{\mathcal{N}}$ are universal trial-independent biases common to all target and non-target trials, respectively; $\mu_{\mathcal{T},k}$ and $\mu_{\mathcal{N},k}$ are the group effect, i.e., the error contribution, of the k^{th} factor; and $f_k(s)$ indicates the group that the s^{th} trial belongs to for factor k.

Groups are determined by the characteristics of both the enrollment and test utterances or speakers. When a factor has two possible categories, three groups are formed. Groups 0 and 1 are assigned to trials with utterances from the same category. Group 2 is assigned to 'cross-group' trials where each side belongs to different categories. For the gender example, the three groups, 0, 1, and 2, could be Male-Male, Female-Female, and Male-Female or Female-Male, respectively.

This model is trained by maximizing the log-likelihood w.r.t. its parameters. We enforced that the average effect over G groups for each factor is zero by adding to the loss function a penalty term, i.e., $m_{Z,k} = \sum_{g=1}^{G} \mu_{Z,k}(g) = 0$, for each factor $k = 1, \ldots, K$, and trial type $Z = \mathcal{T}, \mathcal{N}$. By doing this, the average group effects are observed by the factor-independent terms $\mu_{\mathcal{T}}$ and $\mu_{\mathcal{N}}$. In this manner, if we want to estimate the error probability when we remove the group effect of factor k, we just need to set $\mu_{Z,k} = 0$ and evaluate eq. (2).

2.2. Bernoulli Regression with Confunding Covariates

We can extend the Bernoulli regression model above to include additional explanatory covariates, which can capture the effects of confounding variables on miss and false alarm rates. We add an additional term to the λ coefficients

$$\lambda_Z(s) = \mu_Z + \sum_{k=1}^{K} \mu_{Z,k}(f_k(s)) + \boldsymbol{\theta}_Z^T \mathbf{x}(s), \tag{4}$$

for $Z \in \mathcal{T}, \mathcal{N}$, where $\mathbf{x}(s)$ is a vector of confounding variables not considered in the group effects of eq. (3) and $\boldsymbol{\theta}_Z$ is a vector of coefficients. $\mathbf{x}(s)$ can contain any relevant attributes such as utterance durations or signal-to-noise ratios (SNR).

2.3. Bernoulli Regression with Speaker Effect

Typically, each speaker appears in multiple enrollment and test utterances in ASV datasets. Trials from the same speakers will share some correlated properties–e.g., due to the accent or nationality of the speaker. In [10], Doddington classifies speakers into sheep, goats, lambs, and wolves depending on whether they are easy or difficult to recognize or whether they are easy to imitate or good speaker imitators. To model the correlation between trials involving the same speakers, we introduce new speaker-dependent latent variables into the model. For target trial s_i of speaker i, we have

$$\lambda_{\mathcal{T}}(s_i) = \mu_{\mathcal{T}} + \sum_{k=1}^{K} \mu_{\mathcal{T},k}(f_k(s_i)) + \boldsymbol{\theta}_{\mathcal{T}}^T \mathbf{x}(s) + r_{\mathcal{T},i}$$
$$\text{with } r_{\mathcal{T},i} \sim \mathcal{N}(0, \sigma_{\mathcal{T}}^2) ; \tag{5}$$

and for non-target trials s_{ij} of speaker i versus speaker j,

$$\lambda_{\mathcal{N}}(s_{ij}) = \mu_{\mathcal{N}} + \sum_{k=1}^{K} \mu_{\mathcal{N},k}(f_k(s_{ij})) + \boldsymbol{\theta}_{\mathcal{N}}^T \mathbf{x}(s) + r_{\mathcal{N},i} + r_{\mathcal{N},j}$$
$$\text{with } r_{\mathcal{N},i}, r_{\mathcal{N},j} \sim \mathcal{N}(0, \sigma_{\mathcal{N}}^2) . \tag{6}$$

Note that for targets, we have a single speaker random variable since both trials contain the same speaker, while for non-targets, we need one variable for each speaker in the trial. The speaker variables have a Gaussian prior with learnable variances $\sigma_{\mathcal{T}}^2$ and $\sigma_{\mathcal{N}}^2$. All trials involving speaker i and j share the same values of $r_{\mathcal{T},i}$, $r_{\mathcal{N},i}$, $r_{\mathcal{N},j}$. In this manner, errors in trials of a given speaker are no longer independent.

To train this model, we need to evaluate likelihoods by marginalizing the latent variables while considering that they are tied across trials from the same speakers. We approximated the required integrals by sampling.

2.4. Fairness Metric

To determine the statistical significance of a factor's effect, we use the confidence interval (C.I.) for the ratio R between the

equal error rates (EER) of the two groups that we are comparing. We used the bootstrap method [11] [12] to establish the 95% C.I. If the C.I. does not contain $R = 1$, we decide that the group effect is statistically significant. If we compute the EERs of each group regularly, we get our baseline method,

$$R_{\text{baseline}} = \frac{\text{EER}_1}{\text{EER}_0} \ . \tag{7}$$

However, those EERs will be affected by confounding effects.

We intend to use our model to get estimates of the EERs unaffected by confounding factors. The first step consists of applying a threshold to the ASV scores to accept or reject a trial as a target. Then, we compare this decision with the ground truth to get which trials have errors, i.e., we get the labels t_s. Following this, we train the target and non-target Bernoulli regression models.

Let us assume that we want to compare the effects of groups $g = 0$ and $g = 1$ of factor $k = k'$. We use our model to compute estimates of P_{Miss} and P_{FA} for each group, which only depend on factor k' but do not depend on any of the confounding factors $k \neq k'$, $\mathbf{x}$ or the speaker effects. As explained above, we just need to set the contributions of the confounding factors to zero and evaluate,

$$P_{\text{Miss},k',g} = h(\mu_{\mathcal{T}} + \mu_{\mathcal{T},k'}(g)) \tag{8}$$

$$P_{\text{FA},k',g} = h(\mu_{\mathcal{N}} + \mu_{\mathcal{N},k'}(g)) \ . \tag{9}$$

Then, we consider that EER is equivalent to minimum detection cost function (min. DCF) [13] at a target prior $P_{\mathcal{T}}$=0.5. Thus, we can write the confounding factor free EER as the average of miss and false alarm rate, and the EER ratio as

$$R_{k'} = \frac{P_{\text{Miss},k',1} + P_{\text{FA},k',1}}{P_{\text{Miss},k',0} + P_{\text{FA},k',0}} \ . \tag{10}$$

For example, k' could be the *gender* factor, and groups 0 and 1 could be the male-male and female-female trials, respectively. The effects of confounding factors like speaker effect, channel, language, etc., would be removed by our model.

We could also calculate fairness in terms of min. DCF or act. (actual) DCF. In this case, we first apply the min. DCF threshold or the act. DCF threshold ($-\text{logit}P_{\mathcal{T}}$) to the scores to get decisions and error trials. Then, we would train our model and estimate misses and false alarms with (8) and (9). Finally, we calculate the group effect ratio as,

$$R_{\text{DCF},k'}(P_{\mathcal{T}}) = \frac{P_{\mathcal{T}} P_{\text{Miss},k',1} + (1 - P_{\mathcal{T}}) P_{\text{FA},k',1}}{P_{\mathcal{T}} P_{\text{Miss},k',0} + (1 - P_{\mathcal{T}}) P_{\text{FA},k',0}} \ . \tag{11}$$

3. DATASETS

3.1. Synthetic Data

In real datasets, we lack a ground truth telling us if a group effect is statistically significant in ASV performance. This makes it challenging to evaluate our proposed model against the baseline using real data. To address this, we adopt the approach from [8] and generate synthetic scores with controlled group and confounding effects, providing a reliable ground truth for evaluating our model.

For each experiment, we generated 1000 sets of scores, each based on 500 speakers. We assumed a single factor with two groups, each with 250 speakers. We generated 5000 target and 5000 non-target scores representing the ASV log-likelihood ratios for each set. We generated scores following an additive model, where the score S for the n^{th} target ($\mathcal{T}$) and non-target ($\mathcal{N}$) trial with speakers i and j are

$$\begin{aligned} S(n,\mathcal{T}) = {} & S_{\text{base}}(n,\mathcal{T}) + S_{\text{grp}}(n,f(i),\mathcal{T}) + \\ & S_{\text{spk}}(i,\mathcal{T}) + S_{\text{conf}}(n,\mathcal{T}) x(n,f(i)) \end{aligned} \tag{12}$$

$$\begin{aligned} S(n,\mathcal{N}) = {} & S_{\text{base}}(n,\mathcal{N}) + S_{\text{grp}}(n,f(i),\mathcal{N}) + \\ & S_{\text{spk}}(i,\mathcal{N}) + S_{\text{spk}}(j,\mathcal{N}) + S_{\text{conf}}(n,\mathcal{N}) x(n,f(i)) \end{aligned} \tag{13}$$

where $f(i)$ is the group (g) of speaker i. We did not simulate cross-group trials, so $f(i) = f(j)$. This generation model is additive at the score level, while the error probability model is linear at the λ parameters of the distributions.

Each term in eqs. (12) and (13) are random variables sampled from Gaussian distributions. The factor-independent S_{base}, the group and speaker effect biases were sampled as,

$$S_{\text{base}}(n,Z) \sim \mathcal{N}(m_{\text{base}}(Z), \sigma^2_{\text{base}}) \tag{14}$$

$$S_{\text{grp}}(n,g,Z) \sim \mathcal{N}(m_{\text{grp}}(g,Z), \sigma^2_{\text{grp}}) \tag{15}$$

$$S_{\text{spk}}(i,Z) \sim \mathcal{N}(0, \sigma_{\text{spk}}(Z)^2) \quad \text{for } Z = \mathcal{T}, \mathcal{N} \ . \tag{16}$$

Finally, the binary variable $x(n,g)$ decides whether the confounding factor is present in the trial. The probability $p(g)$ of having the confounding factor differs for each group g. If the confounding factor is present, we add an extra random bias, which depends on the trial label,

$$x(n,g) \ \sim \text{Bernoulli}(p(g)) \tag{17}$$

$$S_{\text{conf}}(n,Z) \sim \mathcal{N}(m_{\text{conf}}(Z), \sigma^2_{\text{conf}}) \ . \tag{18}$$

In our experiments, we set $m_{\text{base}}(\mathcal{T}) = -m_{\text{base}}(\mathcal{N}) = 5$, $\sigma_{\text{base}} = 2.5$; $m_{\text{grp}}(0,\mathcal{T}) = m_{\text{grp}}(0,\mathcal{N}) = 0$, $m_{\text{grp}}(1,\mathcal{T}) = \{0, -0.5, -1, -2\}$, $m_{\text{grp}}(1,\mathcal{N}) = -m_{\text{grp}}(1,\mathcal{T})$, $\sigma_{\text{grp}} = 0.2$; $\sigma_{\text{spk}}(\mathcal{T}) = \sigma_{\text{spk}}(\mathcal{N}) = \{0, 0.5, 1, 2\}$; $p(1) = \{0.5, 0.7, 0.9\}$, $p(0) = 1 - p(1)$; and $m_{\text{conf}}(\mathcal{T}) = -2$, $m_{\text{conf}}(\mathcal{N}) = 2$, and $\sigma_{\text{conf}} = 0.2$. In summary, when we have a group or confounding effect we add a positive bias to non-target scores and a negative bias to targets. Meanwhile, when we have a speaker effect we add a bias that may be positive or negative independently of the sign of the trials, and this bias is the same for all trials from the same speaker and sign (target and non-target trials have different biases).

3.2. VoxCeleb1

We opted to use the VoxCeleb1 dataset [14], specifically, *VoxCeleb1-E* (Extended) and *VoxCeleb1-H* (Hard) test sets.

Table 1. Simulated experiment with equal group effect and different confounding factor probabilities for case and control groups. Score $S(n) = S_{\text{base}}(n) + S_{\text{conf}}(n)x(n, f(i))$

Con. Pr. (%)	**Baseline**		**Basic Model**		**Proposed**	
Case-Ctrl.	*Mean*	*FPR*	*Mean*	*FPR*	*Mean*	*FPR*
0 – 0	0.99	4.2%	0.97	0.6%	1.03	0.6%
50 – 50	1.00	5.4%	0.98	4.8%	1.02	2.6%
70 – 30	1.16	61.1%	1.39	95.3%	1.02	2.4%
90 – 10	1.35	99.8%	2.50	100.0%	1.11	5.3%

Table 2. Simulated experiment with equal group effect, random speaker effect, and confounding effect. Score $S(n) = S_{\text{base}}(n) + S_{\text{spk}}(i) + S_{\text{spk}}(j) + S_{\text{conf}}(n)x(n, f(i))$

Spk.	Conf. Pr. (%)	**Baseline**		**Proposed**	
Std.	*Case-Ctrl.*	*Mean*	*FPR*	*Mean*	*FPR*
0.5	0 – 0	1.01	6.5%	1.03	0.7%
1.0	0 – 0	1.01	13.0%	1.04	3.7%
2.0	0 – 0	1.01	31.5%	1.02	15.2%
1.0	50 – 50	1.00	18.0%	1.03	5.9%
1.0	70 – 30	1.15	60.8%	1.03	9.6%
1.0	90 – 10	1.32	96.3%	1.07	2.3%

VoxCeleb1 provides metadata on the speakers' genders, nationalities, and languages spoken. This allowed us to evaluate fairness in terms of gender and language. The *VoxCeleb1-E* test set comprises 581,480 trials randomly sampled from the complete VoxCeleb1 dataset. The *VoxCeleb1-H* set only contains trials between speakers of the same nationality and gender, making it more challenging. We did not use VoxCeleb-O because the number of speakers (40) is too small to draw statistically significant conclusions.

We calculated scores with three different x-vector models: ECAPA-TDNN small and large, and ResNet100. ECAPA-TDNN was introduced in [15]. ECAPA-TDNN small used three Res2Net layers of 512 dimensions, while the large version used four layers of 2048 dimensions. Our ResNet100 follows the structure from [16]. All models used channel-wise attentive pooling [15] and were trained on VoxCeleb2 dev. set with additive angular margin softmax loss (margin=0.2), followed by large margin fine-tuning (margin=0.4) and hard-prototype mining [17]. We used cosine scoring as the back-end. ECAPA-TDNN small, large, and ResNet100 obtain 1.16, 0.85, and 0.71% EER in VoxCeleb1-E, respectively, and 2.10, 1.66, and 1.30% EER on VoxCeleb1-H.

3.3. NIST SRE2021

We also used the NIST SRE21 dataset [18] as it provides rich metadata in terms of gender, audio source, and spoken language. The audio sources are either conversational telephone speech (CTS) at an 8 kHz sampling frequency or audio from video (AFV) at 16 kHz. There are three languages: English (ENG), Mandarin (CMN), and Cantonese (YUE).

We analyzed the fairness of a Res2Net50-based system from [19] with a back-bone with a scale of 8, channel-wise attentive pooling, and additive angular margin softmax classification layer. The system was trained on 4-second chunks with margin set to 0.3 and fine-tuned on 10-second chunks with margin=0.5. It was trained on the NIST SRE CTS Superset [20], NIST SRE16 dev+eval [21], and VoxCeleb 1+2 [14]. All data was processed at 16 kHz, with CTS data being linearly upsampled. A PLDA back-end [22] was used, which was trained on VoxCeleb+SRE data and adapted to the subset of CMN/YUE speakers on those datasets. The PLDA scores were calibrated into log-likelihood ratios using condition-dependent logistic regression taking into account source, language, and number of enrollment segments. This system obtained 4.9% EER, min. DCF=0.406 and act. DCF=0.415 on the NIST SRE21 audio eval set.

4. EXPERIMENTS AND RESULTS

4.1. Simulated Experiments

We performed a range of experiments by controlling the confounding variable, speaker effect, and group effect to evaluate our proposed method. For each experiment, we generated synthetic data and calculated the EER ratios from eq. (10) and (11). We used the bootstrap method, with 500 samples, to compute the 95% confidence interval (C.I.) of the EER ratios. If the C.I. includes 1.0, we conclude that there is no strong evidence for the group effect to be statistically significant.

A false positive (FP) and a false negative (FN) error happens when models predict equal group effects to be unequal– i.e., 1.0 not inside the C.I.–, and unequal group effects to be equal, respectively. We generated 1000 sets of scores and computed the EER ratio's 95% C.I. for each set. Then, we computed the FP and FN rates by counting the incorrectly classified sets. Since we calculated 95% C.I., the error rates should be around 5% if the method functions correctly [23]. We also calculated the mean EER ratio across sets of scores.

Table 1 displays the outcomes for the scores with equal group effects, a confounding variable, and no speaker effect. We have a case and a control (Ctrl.) group, each one with different probabilities of having the confounding effect (see eq. (17)). The table compares three models: the baseline, the basic model, which is a probabilistic model without confounding variables, making it close to the baseline, and our proposed model from eq. (4). For an equal group effect, the mean EER ratio should be one, and the FP rate close to 5%. However, we observe that the baseline and the basic model rapidly deviate from the ideal result as the difference between the *Ctrl.* and *Case* confounding probabilities increases. As the confounding effect is observed more times in the *Ctrl.*, it is mistaken as a group effect. On the contrary, the proposed model behaves correctly with $\text{FPR} \leq 5\%$.

The first block of Table 2 shows experiments with no confounding variable, equal group effects, and speaker effects for

Table 3. Simulated experiment with unequal group effect, random speaker effect, and confounding effect. Score $S(n) = S_{glob}(n) + S_{grp}(f(i,j)) + S_{spk}(i) + S_{spk}(j) + S_{conf}(n)x(n)$. $\lambda = \mu + \mu_g + r_i + r_j + \theta^T X$

Conf.	m_{grp} = *0.5*						m_{grp} = *1.0*						m_{grp} = *2.0*					
Prob. (%)	**Baseline**			**Proposed**			**Baseline**			**Proposed**			**Baseline**			**Proposed**		
Case-Ctrl.	*Mean*	*FNR*	*FPR*	*Mean*	*FNR*	*FPR*	*Mean*	*FNR*	*FPR*	*Mean*	*FNR*	*FPR*	*Mean*	*FNR*	*FPR*	*Mean*	*FNR*	*FPR*
50 – 50	1.35	2.0%	0.0%	1.61	6.3%	0.0%	1.77	0.0%	0.0%	2.46	0.0%	0.0%	2.84	0.0%	0.0%	5.28	0.0%	0.0%
30 – 70	0.85	27.2%	72.8%	1.64	9.1%	0.0%	1.13	42.0%	0.3%	2.52	0.0%	0.0%	1.89	0.0%	0.0%	5.51	0.0%	0.0%
10 – 90	0.51	0.0%	100%	1.70	38.5%	0.0%	0.70	0.1%	99.9%	2.65	0.6%	0.0%	1.25	4.7%	0.0%	5.98	0.0%	0.0%

different values of σ_{spk} (Spk. Std.) in eq. (16). The baseline and the proposed models exhibit higher FPR as we increase the std. However, the proposed model had less than half the FPR of the baseline. The second block of the table uses simulated scores with $\sigma_{\text{spk}} = 1$ and different confounding factor probabilities. Again, the proposed model is clearly superior to the baseline, which is fooled by the confounding effect.

In Table 3, we present the results of our experiments introducing group effects $S_{\text{grp}}(f(i))$ in addition to S_{spk} and S_{conf}. The group effect was added to the *Case* group, making it performs worse than *Ctrl*, resulting in EER ratios > 1. However, we added the confounding effects to make the *Ctrl* appear worse than the *Case*. We did this by setting the probability (*Conf. Prob.*) for the confounding effect larger for *Ctrl* than for *Case*, contrary to Table 1. This experiment has three outcomes for a given C.I.=$[a, b]$. $a > 1$ indicates a true positive as the model correctly claims statistical significance where *Case* is worse than *Ctrl*. When $1.0 \in [a, b]$, the model incorrectly infers that the group effect is not significant, giving us a false negative. If $b < 1$, the model claims statistical significance but incorrectly indicates *Ctrl* to be worse than *Case*, making it a false positive. The table presents results for several group effect score biases m_{grp} and several *Conf. Prob.* pairs. The baseline and the proposed models performed well for equal *Conf. Prob.s* in both groups. However, when we increase the Conf.-Prob. of *Ctrl* compared to *Case*, the baseline rapidly fails with large FARs and FPRs, while the proposed model is still very robust. The baseline only performed correctly for a large group bias as $m_{\text{grp}} = 2$.

4.2. VoxCeleb1

In the previous section, we demonstrated the baseline's susceptibility and our model's robustness to confounding factors. In this section, we apply our fairness model to our ASV systems' scores on real data. However, due to the absence of ground truth regarding the fairness of the ASV model in real data, our analysis is limited to comparing the outcomes from the baseline and proposed models.

Table 4 presents results on *VoxCeleb1-E* and *VoxCeleb1-H*, focusing on *gender* and *language* as the factors of interest. We only compared same-gender trials, i.e., female-female (*f-f*) vs. male-male (*m-m*), and same-language trials, i.e., Non-English-Non-English (*non-non*) vs. English-English (*eng-eng*). We disregarded cross-gender and cross-language trials. We compared the three ASV systems, described in Section 3.2, sorted in descending order w.r.t. EER value.

The table shows that both the baseline and the proposed model include 1.0 in the C.I. for *gender* in all three ASV systems. Thus, we claim no statistical significance on the factor *gender*. The baseline's mean ratios deviate widely from 1.0, giving the impression of a large performance gap between the groups. Meanwhile, the proposed model's mean ratios are consistently close to 1.0, barely changing across the three systems. Also, the C.I.s are much narrower for the proposed model than for the baseline, demonstrating greater precision.

Regarding *language*, the baseline only claimed statistical significance for ECAPA-Small in VoxCeleb-E and ResNet100 in VoxCeleb-H. Also, the baseline ratios and C.I.s were inconsistent across the ASV models. Meanwhile, the means and C.I.s of our model hardly changed across the different ASV models for *gender* and *language*. For VoxCeleb-E, our model's mean ratio seems to indicate a higher error for non-English trials than for English. However, the C.I. indicates that the difference is not strong enough to claim statistical significance. For VoxCeleb-H, the ratios are close to one, and the C.I. is tighter than for VoxCeleb-E.

The next experiment exhibits how the confounding effects impact the C.I. We only used *VoxCeleb-H* with the *ECAPA-Small* model for this experiment. Instead of grouping as *m-m* versus *f-f*, we randomly split each gender into two groups. For example, we take the *m-m* trials and split them into *m-m*$_0$ and m-m$_1$; and the same for females. We denote this as *pseudo-gender* groups. For each bootstrap sample, we sampled a different speaker grouping. Since both groups are from the same gender, there should be no statistical significance. Table 6 compares *m-m*$_0$ vs. *m-m*$_1$, and *f-f*$_0$ vs. *f-f*$_1$ for different fairness models: baseline, proposed model without confounding factors, proposed model with speaker effect, and proposed model with speaker and language confounding factors. Though all models correctly infer no statistical significance, the mean ratios converge to 1.0, and C.I.s get narrower as the confounding effects are added, thus, making the latter models more precise.

4.3. NIST SRE2021

The SRE21 experiments concentrated on three factors: *gender*, *source*, and *language*. The *gender* groups were male (*m-m*) and female (*f-f*) since there are no cross-gender trials in NIST evaluations. From the three languages in SRE21, we

Table 4. Fairness assessment in terms of *gender* and *language* on VoxCeleb 1 for different x-vector models.

x-Vector	**Gender Fairness** ($\mathrm{EER_{f-f}}/\mathrm{EER_{m-m}}$)				**Language Fairness** ($\mathrm{EER_{non-non}}/\mathrm{EER_{eng-eng}}$)			
	VoxCeleb-E		**VoxCeleb-H**		**VoxCeleb-E**		**VoxCeleb-H**	
	Baseline	**Model**	**Baseline**	**Model**	**Baseline**	**Model**	**Baseline**	**Model**
	Mean, C.I.	*Mean, C.I.*	*Mean, C.I.*	*Mean, C.I.*	*Mean, C.I.*	*Mean, C.I.*	*Mean, C.I.*	*Mean, C.I.*
ECAPA-Small	0.98, [0.77, 1.31]	1.06, [0.97, 1.18]	1.21, [0.83, 1.70]	1.06, [0.98, 1.16]	1.54, [1.06, 2.06]	1.25, [0.96, 1.52]	0.91, [0.51, 1.24]	1.10, [0.92, 1.30]
ECAPA-Large	1.41, [0.90, 2.07]	1.08, [0.95, 1.22]	0.83, [0.63, 1.14]	1.00, [0.91, 1.08]	0.91, [0.50, 1.24]	1.22, [0.95, 1.58]	1.37, [0.82, 1.96]	1.09, [0.92, 1.26]
ResNet-100	1.34, [0.82, 1.97]	1.01, [0.91, 1.10]	0.81, [0.62, 1.16]	1.05, [0.96, 1.16]	1.25, [0.76, 1.68]	1.35, [0.88, 1.78]	1.98, [1.06, 3.14]	1.05, [0.82, 1.30]

Table 5. Fairness assessment in terms of *gender*, *language*, and *source* on SRE21

Experiment	**Equal Error Rate**				**C-Primary**			
	Baseline		**Proposed Model**		**Baseline**		**Proposed Model**	
	Mean	*C.I.*	*Mean*	*C.I.*	*Mean*	*C.I.*	*Mean*	*C.I.*
Female vs. Male	0.72	[0.66, 0.77]	1.17	[1.10, 1.23]	0.38	[0.34, 0.42]	1.36	[1.24, 1.5]
AFV-AFV vs. CTS-CTS	0.65	[0.56, 0.73]	0.94	[0.92, 0.95]	0.85	[0.73, 0.98]	0.82	[0.75, 0.89]
AFV-AFV vs. AFV-CTS	0.40	[0.35, 0.43]	0.64	[0.63, 0.65]	0.82	[0.70, 0.93]	0.47	[0.46, 0.48]
YUE-YUE vs. CMN-CMN	0.91	[0.78, 1.04]	0.92	[0.83, 0.98]	0.78	[0.62, 0.97]	0.89	[0.75, 1.04]
YUE-YUE vs. ENG-ENG	1.04	[0.87, 1.21]	0.96	[0.90, 1.01]	0.95	[0.76, 1.16]	0.88	[0.73, 1.05]
YUE-YUE vs. CMN-YUE	0.66	[0.57, 0.74]	0.93	[0.86, 0.99]	0.48	[0.39, 0.58]	0.84	[0.70, 0.97]
YUE-YUE vs. CMN-ENG	0.85	[0.75, 0.99]	0.76	[0.70, 0.86]	1.07	[0.86, 1.30]	0.69	[0.58, 0.82]
YUE-YUE vs. YUE-ENG	0.64	[0.56, 0.72]	0.95	[0.90, 1.00]	0.47	[0.39, 0.58]	0.84	[0.72, 0.99]

Table 6. Fairness of *pseudo-gender* groups from VoxCeleb-H across fairness model versions.

Fairness Model	**Only Male**		**Only Female**	
	Mean	*C.I.*	*Mean*	*C.I.*
Baseline	1.03	[0.60, 1.73]	1.09	[0.78, 1.55]
$\lambda = \mu + \mu_g$	1.03	[0.64, 1.59]	1.02	[0.75, 1.40]
$\lambda = \mu + \mu_g + r_i + r_j$	1.03	[0.68, 1.45]	1.02	[0.75, 1.37]
$\lambda = \mu + \mu_g + \mu_{lang} + r_i + r_j$	0.98	[0.70, 1.37]	1.00	[0.83, 1.23]

Table 7. Effect of *source* (AFV-AFV vs CTS-CTS) with respect to *gender* in SRE21

Model	**Both Genders**		**Only Male**		**Only Female**	
	Mean	*C.I.*	*Mean*	*C.I.*	*Mean*	*C.I.*
Baseline	0.48	[0.36, 0.62]	0.58	[0.22, 1.14]	0.57	[0.41, 0.75]
Basic	0.47	[0.31, 0.77]	0.82	[0.70, 0.97]	0.86	[0.64, 0.95]
Proposed	0.91	[0.83, 0.97]	0.84	[0.73, 0.92]	0.96	[0.92, 0.98]

created six *language* groups: *YUE-YUE*, *CMN-CMN*, *ENG-ENG*, *YUE-CMN*, *CMN-ENG*, and *YUE-ENG*. Similarly, we formed three *source* groups (*AFV-AFV*, *CTS-CTS*, *AFV-CTS*).

Table 5 shows the means and C.I.s of the EER ratios as well as the actual C-Primary [18] for all group pairs. Based on our model, trials with female speakers are more prone to errors than male speakers. However, the baseline shows that male speakers' trials have higher errors. We believe some confounding factors made the male trials prone to more errors than the female ones. But when our model removed those factors, male trials showed better performance. Regarding *source*, both baseline and proposed models claim that *CTS-CTS* trials have higher errors than *AFV-AFV* trials, and the difference is even larger for *CTS-AFV* trials. However, our model shows that the difference between sources is lower than that indicated by the baseline. Regarding languages, our model claims statistical significance in 3 out of 5 EER and C-primary ratios. In general, mean ratios are more consistent across groups for the proposed model than the baseline.

For the experiment in Table 7, we used only *CMN-CMN* trials and compared the EERs w.r.t. *source* (*AFV-AFV* vs. *CTS-CTS*) while varying the *gender* groups. The *Both Genders* column is for the experiment where each *source* group has equal male and female trials. The *Only Male* and *Only Female* columns indicate taking *pseudo gender* groups similar to table 6. If the model works properly, it should reach the same conclusion on the statistical significance of *source* regardless of how the gender groups were formed. The table shows that our model consistently claims statistical significance with a mean ratio ~ 0.9, while the baseline gives inconsistent conclusions. The baseline and basic models have large differences between *Both Gender* and single-gender columns. Our model's CIs are notably narrower than the other two.

5. CONCLUSION

In this work, we proposed a model-based method to detect any unfair bias in ASV systems. Our method takes into account any confounding factors other than the factor of interest that may affect the trial errors. To validate our method, we generate a synthetic dataset and perform several simulated experiments. We compare our model's performance to the baseline, where we take the ratio of the regular EERs of each group. The experiment outcomes show that the proposed method yields significantly lower false positive and false negative rates compared to the baseline. We also experiment on VoxCeleb1 and NIST SRE21 datasets and demonstrate that our method produces consistent results.

6. REFERENCES

[1] Wiebke Toussaint Hutiri and Aaron Yi Ding, "Bias in Automated Speaker Recognition," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 6 2022, pp. 230–247, ACM.

[2] Sophie Si, Zhengxiong Li, and Wenyao Xu, "Exploring Demographic Effects on Speaker Verification," in *2021 IEEE Conference on Communications and Network Security (CNS)*. 10 2021, pp. 1–2, IEEE.

[3] Gianni Fenu, Hicham Lafhouli, and Mirko Marras, "Exploring Algorithmic Fairness in Deep Speaker Verification," in *Computational Science and Its Applications – ICCSA 2020*, Cagliari, 7 2020, pp. 77–93.

[4] NIST, "NIST 2019 Speaker Recognition Evaluation: CTS Challenge," Tech. Rep., NIST, 7 2019.

[5] Mariel Estevez and Luciana Ferrer, "Study on the Fairness of Speaker Verification Systems Across Accent and Gender Groups," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6 2023, pp. 1–5, IEEE.

[6] David A. van Leeuwen and Niko Brümmer, "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems," in *Speaker Classification*, vol. LNAI 4343, pp. 330–353. 2007.

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

[8] Zhe Liu, Irina-Elena Veliche, and Fuchun Peng, "Model-Based Approach for Measuring the Fairness in ASR," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 5 2022, pp. 6532–6536, IEEE.

[9] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370, 1972.

[10] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998, p. paper 0608.

[11] Michael Wood, "Statistical Inference using Bootstrap Confidence Intervals," *Significance*, vol. 1, no. 4, pp. 180–182, 12 2004.

[12] Thomas J. DiCiccio and Bradley Efron, "Bootstrap confidence intervals," *Statistical Science*, vol. 11, no. 3, 9 1996.

[13] Sachin S Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet, "The sri nist 2008 speaker recognition evaluation system," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4205–4208.

[14] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, pp. 101027, 3 2020.

[15] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*, ISCA, 10 2020, pp. 3830–3834, ISCA.

[16] Rostislav Makarov, Nikita Torgashov, Alexander Alenin, Ivan Yakovlev, and Anton Okhotnikov, "ID R&D System Description to VoxCeleb Speaker Recognition Challenge 2022," 2022.

[17] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "Cross-Lingual Speaker Verification with Domain-Balanced Hard Prototype Mining and Language-Dependent Score Normalization," in *Interspeech 2020*, ISCA, 10 2020, pp. 756–760, ISCA.

[18] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds, "The 2021 NIST Speaker Recognition Evaluation," 4 2022.

[19] Jesús Villalba, Bengt J Borgstrom, Saurabh Kataria, Magdalena Rybicka, Carlos D Castillo, Jaejin Cho, L. Paola García-Perera, Pedro A. Torres-Carrasquillo, and Najim Dehak, "Advances in cross-lingual and cross-source audio-visual speaker recognition: The jhu-mit system for nist sre21," 6 2022, pp. 213–220, ISCA.

[20] Seyed Omid Sadjadi, "NIST SRE CTS Superset: A large-scale dataset for telephony speaker recognition," 8 2021.

[21] National Institute of Standards and Technology, "NIST 2016 Speaker Recognition Evaluation Plan," 8 2016.

[22] Simon J.D. Prince and James H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, 10 2007, pp. 1–8, IEEE.

[23] Ana-Maria Simundic, "Confidence interval," *Biochemia Medica*, vol. 18, no. 2, pp. 154–161, 2008.