Bias in Automated Speaker Recognition

Wiebke Toussaint Hutiri*
Aaron Yi Ding
w.toussaint@tudelft.nl
Delft University of Technology
Delft, The Netherlands

ABSTRACT

Automated speaker recognition uses data processing to identify speakers by their voice. Today, automated speaker recognition is deployed on billions of smart devices and in services such as call centres. Despite their wide-scale deployment and known sources of bias in related domains like face recognition and natural language processing, bias in automated speaker recognition has not been studied systematically. We present an in-depth empirical and analytical study of bias in the machine learning development workflow of speaker verification, a voice biometric and core task in automated speaker recognition. Drawing on an established framework for understanding sources of harm in machine learning, we show that bias exists at every development stage in the well-known VoxCeleb Speaker Recognition Challenge, including data generation, model building, and implementation. Most affected are female speakers and non-US nationalities, who experience significant performance degradation. Leveraging the insights from our findings, we make practical recommendations for mitigating bias in automated speaker recognition, and outline future research directions.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Security and privacy** → **Biometrics**; • **Computing methodologies** → *Speech recognition*; Machine learning.

KEYWORDS

speaker recognition, speaker verification, bias, fairness, audit, evaluation

ACM Reference Format:
Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3531146.3533089

1 INTRODUCTION

The human voice contains an uncanny amount of personal information. Decades of research have correlated behavioural, demographic, physiological, sociological and many other individual characteristics to a person's voice. Even if untrained human listeners cannot discern all the details, automated voice processing can: voice profiling can reveal sensitive personal attributes such as age, anatomy, health status, medical conditions, identity, intoxication, emotional state, stress, and truthfulness from speech [48]. Speaker recognition is a type of voice processing that automatically recognises the identity of a human speaker from personal information contained in their voice [8]. Today, speaker recognition permeates private and public life. Speaker recognition systems are deployed at scale in call centers, on billions of mobile phones and on voice-enabled consumer devices such as smart speakers. They grant access not only to personal devices in intimate moments, but also to essential public services for vulnerable user groups. For example, in Mexico speaker recognition is used to allow senior citizens to provide a telephonic proof-of-life to receive their pension [31].

In this paper we study bias in speaker recognition systems. Bias in machine learning (ML) is a source of unfairness [29] that can have harmful consequences, such as discrimination [55]. Bias and discrimination in the development of face recognition technologies [4, 42, 43], natural language processing [3] and automated speech recognition [1, 22, 52, 53] are well studied and documented. Bias in speaker recognition, a related domain, has received very limited attention. Yet, speaker recognition technologies are pervasive and process extremely sensitive personal data that is intricately intertwined with our individual identity. They are deployed in high-stakes applications, while the modality of their input data makes them susceptible to perpetrate discrimination. It is thus urgent to investigate bias in these systems, so that mitigating and regulatory actions can be taken to forestall potential negative consequences.

Drawing on Suresh and Guttag's *Framework for Understanding Sources of Harm* [51], we present the first detailed study on bias in speaker recognition. We approach this work as a combination of an analytical and empirical evaluation focused on the VoxCeleb Speaker Recognition Challenge [32], one of the most popular benchmarks in the domain with widely used datasets. Our study shows that existing benchmark datasets, learning mechanisms, evaluation practices, aggregation habits and post-processing choices in the speaker recognition domain produce systems that are biased against female and non-US speakers. Our contributions are:

(1) We present an evaluation framework for quantifying performance disparities in speaker verification - a speaker recognition task that serves as the biometrics of voice
(2) We apply this framework to conduct the first evaluation of bias in speaker verification. Our results show that bias exists at every stage of the ML development pipeline
(3) Informed by our evaluation, we recommend research directions to address bias in automated speaker recognition

[*] Corresponding author

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533089

Our paper is structured as follows. In Section 2 and 3 we review related work and provide a background on speaker recognition, its evaluation and supporting infrastructure for its development. We then present the empirical experiment setup and bias evaluation framework in Section 4. In Section 5 we present our findings of bias in data generation, and in Section 6 our findings of bias in model building and implementation. We discuss our findings, and make recommendations for mitigating bias in speaker recognition in Section 7. Finally, we conclude in Section 8.

2 RELATED WORK

In this section we provide a background on speaker recognition within its historical development context and present evidence of bias in the domain. We then introduce the theoretical framework on which we base our analytical and empirical bias evaluation.

2.1 Historical Development of Automated Speaker Recognition

Since its inception, research into speaker recognition has enabled voice-based applications for access control, transaction authentication, forensics, law enforcement, speech data management, personalisation and many others [44]. As a voice-based biometric, speaker verification is viewed to have several advantages: it is physically non-intrusive, system users have historically considered it to be non-threatening, microphone sensors are ubiquitously available in telephone and mobile systems or can be installed at low-cost if they are not, and in many remote applications speech is the only form of biometrics available [44]. Given the proliferation of speaker recognition systems in digital surveillance technologies, concerns over its pervasive, hidden and invasive nature are rising [24].

2.1.1 Parallels to Facial Recognition. The historical development of automated speaker recognition reflects that of facial recognition in many aspects. Similar to the development of facial recognition systems [43], research in early speaker recognition systems was supported by defense agencies, with envisioned applications in national security domains such as forensics [9]. The systems relied on datasets constructed from telephone corpuses and their development was greatly accelerated through coordinated, regular competitions and benchmarks.

2.1.2 From Classical Approaches to Deep Neural Networks. Two years after the deep learning breakthroughs in computer vision, Deep Neural Networks (DNNs) were first applied to speaker recognition systems [14]. Since 2016, DNNs have become the dominant technique for developing speaker recognition systems [23, 49, 50]. DNNs have distinguished themselves in important ways from traditional approaches for speaker recognition: their performance is superior on short speech utterances [50], they can be trained in an end-to-end fashion using only speaker labels, thus reducing laborious labelling efforts [14], and they can leverage many of the techniques that have demonstrated success in the image recognition domain. To enable the new era of deep speaker recognition, large scale datasets were needed to support research in this emerging area, and methods for generating them adapted approaches from face recognition. For example, a popular speaker recognition dataset, VoxCeleb [34], is derived from the voice signals in Youtube videos of celebrities contained in the well-known face recognition dataset VGG Face [40]. Another dataset, MOBIO [18], was developed jointly for mobile face recognition and speaker recognition.

2.2 Bias in Speaker and Speech Recognition

2.2.1 Early Evidence of Bias in Speaker Recognition. It is well established that speaker characteristics such as age, accent and gender affect the performance of speaker recognition [11]. In acknowledgement of this, past works in speech science, like research promoted through the 2013 Speaker Recognition Evaluation in Mobile Environments challenge, have reported speaker recognition performance separately for male and female speakers [19]. The submissions to the challenge made it clear that bias is a cause of concerns: of 12 submissions, all submitted systems performed worse for females than for males on the evaluation set. On average the error rate for females was 49.35% greater than for males. Despite these performance differences being acknowledged, they went unquestioned and were attributed solely to an unbalanced training set that contained a male:female speaker ratio of 2:1. In later works the discrepancy between female and male speakers is still evident and reported, but remains unquestioned and unaddressed [39]. Historically, a common approach to avoid gender-based bias has been to develop separate models for female and male speakers [21]. While this may be insufficient to eradicate bias, generating separate feature sets for female and male speakers can reduce it [27]. Beyond considering binary gender, evaluating demographic performance gaps based on other speaker attributes is less common, and intersectional speaker subgroups have not been considered.

2.2.2 Nuanced Evaluation No Longer Common Practice. Since the adoption of Deep Neural Networks (DNNs) for speaker recognition, practices of evaluating system performance for speaker subgroups seem to have disappeared. Several system properties beyond performance have been considered in recent years, such as robustness [2] and privacy [35]. However, research in robustness and privacy in speaker recognition does not address the glaring gap that remains in the domain: system performance appears biased against speaker groups based on their demographic attributes. Only one recent study investigates bias in end-to-end deep learning models based on speaker age and gender [7], reconfirming the importance of balanced training sets.

2.2.3 Bias in Automated Speech Recognition. In automated speech recognition, which is concerned with the linguistic content of voice data, not with speaker identity, recent studies have provided evidence that commercial automated caption systems have a higher word error rate for speakers of colour [52]. Similar racial disparities exist in commercial speech-to-text systems, which are strongly influenced by pronunciation and dialect [22]. Considering their shared technical backbone with facial recognition systems, and shared data input with automated speech recognition systems, we expect that bias and harms identified in these domains will also exist in speaker recognition systems. Mounting evidence of bias in facial and speech recognition, the abundance of historic evidence of bias and the vacuum of public information about bias in speaker recognition, strengthen the motivation for our work.

2.3 Sources of Harm in the ML Life Cycle

We draw on Suresh and Guttag's [51] *Framework for Understanding Sources of Harm* through the ML life cycle to ground our investigation into bias in automated speaker recognition. Suresh and Guttag divide the ML life cycle into two streams and identify seven sources of bias related harms across the two streams: 1) the data generation stream can contain historical, representational and measurement bias; and 2) the model building and implementation stream can contain learning, aggregation, evaluation and deployment bias. *Historical bias* replicates bias, like stereotypes, that are present in the world as is or was. *Representation bias* underrepresents a subset of the population in the sample, resulting in poor generalization for that subset. *Measurement bias* occurs in the process of designing features and labels to use in the prediction problem. *Aggregation bias* arises when data contains underlying groups that should be treated separately, but that are instead subjected to uniform treatment. *Learning bias* concerns modeling choices and their effect on amplifying performance disparities across samples. *Evaluation bias* is attributed to a benchmark population that is not representative of the user population, and to evaluation metrics that provide an oversimplified view of model performance. Finally, *deployment bias* arises when the application context and usage environment do not match the problem space as it was conceptualised during model development.

Next we introduce automated speaker recognition, and then show analytically and empirically how these seven types of bias manifest in the speaker recognition development ecosystem.

3 BACKGROUND

Speaker recognition refers to the collection of data processing tasks that identify a speaker by their voice [8]. Core tasks in speaker recognition are *speaker identification*, which determines a speaker's identity from a subset of speakers, *speaker verification*, which validates if a speaker's identity matches the identity of a stored speech utterance, and *speaker diarisation*, which is concerned with partitioning speech to distinguish between different speakers [2]. While technical implementation details differ in the three areas, their communities overlap, they share datasets and participate in the same competitions. We focus our investigation in this paper on speaker verification, which underlies voice biometrics. However, as the tasks have evolved together, many of the biases that we uncover in speaker verification also apply to speaker identification and diarisation. In this section we provide a high level overview of speaker verification and its evaluation, as well as its supporting ecosystem of competitions and benchmarks that have advanced the field. We refer the reader to [2] for a detailed technical survey on state-of-the-art speaker recognition, and to [21] for a review on the classical speaker recognition literature prior to the advent of Deep Neural Networks (DNNs).

3.1 Speaker Verification Overview

A speaker verification system determines whether a candidate speaker matches the identity of a registered speaker by comparing a candidate speaker's speech signal (i.e. *trial utterance*) to the speech signal of a registered speaker (i.e. *enrollment utterance*). Speaker verification is classified based on its training data as text-dependent if speech signals are fixed phrases or text-independent if not, prompted if speech was produced by reading text or spontaneous if not [9]. Spontaneous text-independent speech is the type of speech that occurs naturally when a speaker interacts with a voice assistant or a call centre agent, and presents the most general speaker verification task.

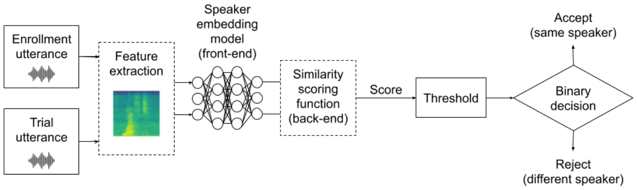

Figure 1: Speaker verification data processing pipeline

As shown in Figure 1, many speaker verification systems consists of two stages, a front-end that generates a speaker embedding model for enrollment and trial utterances, and a back-end that computes a similarity score for the two resultant embeddings. Alternatively, end-to-end speaker verification directly learns a similarity score from training utterances [14]. Modern speaker verification systems use DNNs to learn the front-end embedding, or to train the end-to-end system [2]. As the final step of the speaker verification process, the score output is compared to a threshold. Speaker identity is accepted if the score lies above the threshold, and rejected if it lies below the threshold.

3.2 Speaker Verification Evaluation

To evaluate speaker verification systems, scores are generated for many pairs of enrollment and trial utterances. The utterance pairs are labelled as being from the same or from different speakers. Two typical score distributions generated from many same and different speaker utterance pairs are shown in Figure 6 in the Appendix. After calibrating the speaker verification system to a threshold (e.g. *equal error rate* or *detection cost*), utterance pairs with a score below the threshold are classified as different speakers and the trail utterance is rejected. Utterance pairs with a score above the threshold are classified as the same speaker, and accepted. As the two distributions overlap, classification is not perfect. At a particular threshold value there will be false positives, i.e. utterance pairs of different speakers with a score above the threshold, and false negatives, i.e. utterance pairs of the same speakers with a score below the threshold.

Speaker verification performance is determined by its false positive rate (FPR) and false negative rate (FNR) at the threshold value to which the system has been calibrated [9]. It is accepted that the two error rates present a trade-off, and that selecting an appropriate threshold is an application-specific design decision [38]. The threshold value is determined by balancing the FPR and FNR error rates for a particular objective, such as obtaining an *equal error rate* (EER) for FPR and FNR, or minimising a cost function. The *detection cost function* (DCF) is a weighted sum of FPR and FNR across threshold values, with weights determined by the application requirements. To compare performance across models, systems are frequently tuned to the threshold value at the minimum of the DCF, and the corresponding *detection cost* C_{Det} value is reported as a metric.

Various detection cost functions have been proposed over time, such as the following, proposed in the NIST SRE 2019 Evaluation Plan [37]:

$$C_{Det}(\theta) = C_{FN} \times P_{Target} \times P_{FN}(\theta) + C_{FP} \times \left(1 - P_{Target}\right) \times P_{FP}(\theta)$$
$$P_{Target} = 0.05,\ C_{FN} = 1,\ C_{FP} = 1 \tag{1}$$

Speech science literature recommends that *detection error trade-off* (DET) curves [9] are used to visualise the trade-off between FPR and FNR, and to consider system performance across various thresholds. DET curves visualise the FPR and FNR at different operating thresholds on the x- and y-axis of a normal deviate scale [26] (see Figure 7 in the Appendix). They can be used to analyse the inter-model performance (across models), and are also recommended for analysing intra-model performance (across speaker subgroups in a model).

3.3 Competitions and Benchmarks

Speaker recognition challenges have played an important role in evaluating and benchmarking advances in speaker verification. They were first initiated within the Information Technology Laboratory of the US National Institute of Standards and Technology (NIST) to conduct evaluation driven research on automated speaker recognition [9]. The NIST Speaker Recognition Evaluation (SRE) challenges and their associated evaluation plans have been important drivers of speaker verification evaluation. In addition, new challenges have emerged over time to address the requirements of emerging applications and tasks. Table 1 summarises recent challenges, their organisers and the metrics used for evaluation. Most challenges have adopted the minimum of the detection cost function, *min* C_{Det}, recommended by the NIST SREs as their primary metric. As the NIST SREs have modified this function over time, different challenges use different versions of the metric. In the remainder of this paper we evaluate bias in the VoxCeleb Speaker Recognition Challenge (SRC).

4 EXPERIMENT SETUP

Launched in 2019, the objective of the VoxCeleb SRC is to "probe how well current [speaker recognition] methods can recognise speakers from speech obtained 'in the wild'" [10]. The challenge has four tracks: open speaker diarisation, open and closed fully supervised, and closed self-supervised speaker verification. It serves as a well-known benchmark, and has received several hundred submissions over the past three years. The popularity of the challenge and its datasets make it a suitable candidate for our evaluation, representative of the current ecosystem. We evaluate group bias in the speaker verification track of the VoxCeleb SRC.

4.1 Baseline Models

The challenge has released two pre-trained baseline models [15] trained on the VoxCeleb 2 training set [33] with close to 1 million speech utterances of 5994 speakers. 61% of speakers are male and 29% of speakers have a US nationality, which is the most represented nationality. More detailed metadata is not readily available. The baseline models are based on a 34-layer ResNet trunk architecture. *ResNetSE34V2* [15] is a larger model, with an architecture optimised for predictive performance. *ResNetSE34L* [5] is a smaller model that contains less than a fifth of the parameters of *ResNetSE34V2* and has smaller input dimensions. This reduces the computation time and the memory footprint of the model, two important considerations for on-device deployment in applications like smartphones and smart speakers. The model developers have optimised it for fast execution. We downloaded and used both baseline models as black-box predictors in our evaluation. The technical details of the baseline models are summarised in Table 3 in the Appendix.

4.2 Evaluation Dataset

We evaluate the baseline models on three established evaluation sets that can be constructed from the utterances in the VoxCeleb 1 dataset [33]. VoxCeleb 1 was released in 2017 with the goal of creating a large scale, text-independent speaker recognition dataset that mimics unconstrained, real-world speech conditions, in order to explore the use of DNNs for speaker recognition tasks [34]. The dataset contains 153 516 short clips of audio-visual utterances of 1251 celebrities in challenging acoustic environments (e.g. background chatter, laughter, speech overlap) extracted from YouTube videos. The dataset also includes metadata for speakers' gender and nationality, and is disjoint from VoxCeleb 2 which is used for training. Three different evaluation sets have been designed for testing speaker verification with VoxCeleb 1. We consider all three evaluation sets in our analysis. The evaluation sets are discussed in detail in §6.3.

4.3 Speaker Subgroups and Bias Evaluation Measures

We selected subgroups based on attributes and categories captured in the VoxCeleb metadata: gender and nationality. We then established bias by evaluating performance disparities between these subgroups using existing evaluation measures in speaker verification. Reusing attributes and category labels, though practical for facilitating our study, perpetuates existing bias. We reflect on the consequences of this in our analysis of measurement bias in §5.3.

Our first technique for establishing bias is to plot the DET curves for all subgroups, and to compare the subgroups' DET curves to the overall curve for all subgroups. As speaker verification systems must operate on the DET curve, this presents the theoretical performance boundary of the model across subgroups. Secondly, we consider bias at the threshold to which the system has been calibrated, which ultimately presents the operating point of the system. Here we consider an unbiased system as one that has equal false positive and true positive (or false negative) rates across subgroups, in line with the definition of equalized odds [12]. We compare each subgroup's performance to the overall system performance to facilitate comparison across a large number of subgroups, and thus deviate slightly from the formal definition of equalized odds. We use $C_{Det}(\theta)$ as defined in Equation 1 to determine the calibration threshold and quantify the relative bias towards each subgroup with the ratio of the subgroup cost $C_{Det}(\theta)^{SG}$ to the overall cost $C_{Det}(\theta)^{overall}$ at the threshold value where $C_{Det}(\theta)$ is minimized for the overall system:

Name	Organiser	Years	Metrics
NIST SRE [9]	US National Inst. of Standards & Tech.	1996 - 2021	Detection Cost Function
SRE in Mobile Env's [19]	Idiap Research Institute	2013	DET curve, EER, *half total error rate*
Speakers in the Wild SRC [28]	at Interspeech 2016	2016	$min\ C_{Det}{}^{*}$ (SRE2016), R_{prec}, C_{llr}
VoxCeleb SRC [32]	Oxford Visual Geometry Group	2019 - 2021	$min\ C_{Det}{}^{*}$ (SRE2018), EER
Far-Field SVC [41]	at Interspeech 2020	2020	$min\ C_{Det}{}^{*}$, EER
Short Duration SVC [58]	at Interspeech 2021	2020 - 2021	$norm\ min\ C_{Det}{}^{*}$ (SRE08)
SUPERB benchmark [57]	CMU, JHU, MIT, NTU, Facebook AI	2021	EER*

Table 1: Evaluation metrics for Speaker Verification and Recognition Challenges (SVC and SRC) (* primary metric)

$$subgroup\ bias = \frac{C_{Det}\left(\theta_{@\ overall\ min}\right)^{SG}}{C_{Det}\left(\theta_{@\ overall\ min}\right)^{overall}} \qquad (2)$$

If the *subgroup bias* is greater than 1, the subgroup performance is worse than the overall performance, and the speaker verification model is prejudiced against that subgroup. Conversely, if the *subgroup bias* is less than 1, the model favours the subgroup. If the ratio is exactly 1, the model is unbiased for that subgroup.

4.4 Black-box Bias Evaluation Framework

We designed a framework[1] that replicates a real evaluation scenario to evaluate bias in the VoxCeleb SRC benchmark. Figure 2 shows an overview of our approach. We start with pairs of single-speaker speech utterances in the evaluation dataset as input, and use the baseline models, *ResNetSE32V2* and *ResNetSE34L*, as black-box predictors. The baseline models output scores for all utterance pairs in the evaluation set. We set the threshold to the value that minimizes the overall system cost of the DCF and accept or reject speakers in utterance pairs based on that. Our predicted binary acceptance is then compared to the true labels of the utterance pairs to determine false positive and false negative predictions. Using the metadata for speakers, we allocate each utterance pair to a subgroup based on the attributes of the enrollment utterance. From these inputs we evaluate bias by establishing the FPR, FNR and thus $C_{Det}\ (\theta)^{SG}$ at the threshold value for each subgroup. We also plot DET curves from the outputs scores for each subgroup. The evaluation is repeated for each of the three VoxCeleb 1 evaluation sets. Using this evaluation framework, we now identify sources of bias in data generation (Section 5) and model building and implementation (Section 6).

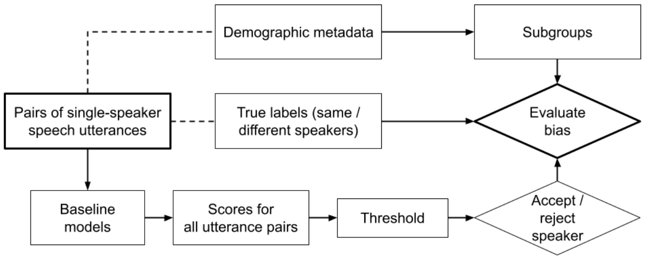
Demographic metadata
Subgroups
Pairs of single-speaker speech utterances
True labels (same / different speakers)
Evaluate bias
Baseline models
Scores for all utterance pairs
Threshold
Accept / reject speaker

Figure 2: Framework for black-box bias evaluation of speaker verification models

5 BIAS IN DATA GENERATION

In this section we identify sources of bias in the VoxCeleb SRC that arise during data generation (see Suresh and Guttag's *Framework for Understanding Sources of Harm* described in §2.3). The stage involves data generation, population definition and sampling, measurement and pre-processing, with the goal of creating training, test and benchmark datasets. The types of bias that arise in these processes are *historical bias*, *representation bias* and *measurement bias.*

5.1 Historical Bias

Historical bias replicates biases, like stereotypes, that are present in the world as is or was.

The VoxCeleb 1 dataset was constructed with a fully automated data processing pipeline from open-source audio-visual media [34]. The candidate speakers for the dataset were sourced from the VGG Face dataset [40], which is based on the intersection of the most searched names in the Freebase knowledge graph and Internet Movie Database (IMDB). After searching and downloading video clips for identified celebrities, further processing was done to track faces, identify active speakers and verify the speaker's identity using the HOG-based face detector [20], Sync-Net [6] and VGG Face CNN [47] respectively. If the face of a speaker was correctly identified, the clip was included in the dataset.

Bias in facial recognition technologies is well known [4, 42, 43], and historic bias pervades the automated data generation process of VoxCeleb. The VoxCeleb 1 inclusion criteria subject the dataset to the same bias that has been exposed in facial recognition verification technology and reinforce popularity bias based on search results [29]. Moreover, the data processing pipeline directly translates bias in facial recognition systems into the speaker verification domain, as failures in the former will result in speaker exclusion from VoxCeleb 1.

5.2 Representation Bias

Representation bias underrepresents a subset of the population in its sample, resulting in poor generalization for that subset.

The VoxCeleb 1 dataset is skewed towards males and US nationals, as can be seen in Figure 8 in the Appendix. Performance for this group is the most reliable and aligns the closest with the average performance. For subgroups with the smallest amount of speakers, such as Italian, German and Irish females, DET curves in Figure 9 in the Appendix show that performance is unreliable. In the context of benchmark evaluations, such skewed representation not only provides an inadequate understanding of the real capabilities

[1] The code for the evaluation has been released as an open-source python library: https://github.com/wiebket/bt4vt/releases/tag/v0.1

of speaker verification for a diverse population of people, but it also shapes the development of the technology towards the group of people that are most represented. Representation bias affects the quality of our bias evaluation for underrepresented subgroups. However, there are sufficient subgroups that have a reasonable representation of speakers (USA, Canadian, UK, Indian and Australian males and females) to support our efforts of gathering evidence of bias.

Recent work on age recognition with the VoxCeleb datasets [13] shows that speakers between ages 20 and 50 are most represented in the dataset, indicating that representation bias is also evident across speaker age. Nationality, gender and age only account for some of the attributes of a speaker's voice that affect automated speaker recognition [48]. We discuss additional attributes that are likely to affect performance in the following section on *measurement bias*. Being a celebrity dataset that is not representative of the broad public, it is likely that VoxCeleb 1 contains representation bias that affect many other sensitive speaker attributes. Representation bias contributes to aggregation bias (§6.1), evaluation bias (6.3) and deployment bias (§6.4) in speaker verification, and is discussed in further detail in those sections.

5.3 Measurement Bias

Measurement bias occurs in the process of designing features and labels to use in the prediction problem.

In our analysis of measurement bias we focus on labelling choices made in the VoxCeleb 1 metadata, which our study inherits in our subgroup design choices. While these labels are not used for making predictions, they are used to make judgements about speaker representation in the dataset. They also inform subgroup design, which plays a fundamental role in our group-based bias analysis.

The VoxCeleb 1 dataset creators inferred nationality labels from speakers' countries of citizenship, as obtained from Wikipedia. Their underlying motivation for doing this was to assign a label that is indicative of a speaker's accent [33]. Conflating nationality and accent is problematic, as people with the same citizenship can speak the same language with different accents. Likewise, many countries have citizens speaking different languages (e.g. India has 7 languages with more than 50 million first language speakers each [56]). Using nationality as a subgroup label has merits[2], even if conflating nationality, accent and language raises concerns. The nationality-based performance differences that we observe suggest that language, accent, ethnicity and dialect may also produce disparate performance.

The metadata considers only binary gender categories, namely male and female. From the dataset description it is unclear what method was followed to label speakers by gender. Many concerns about gender labelling in face analysis technologies have been pointed out in prior research [45], and similar concerns hold true in speaker recognition. Simply replacing a binary gender classification with more categories is not a recommended alternative. Even if it were possible to produce accurate labels, they might help to mitigate bias in speaker verification only while offering a new surface for harm, for example through voice-based gender classification enabled targeting.

6 BIAS IN MODEL BUILDING AND IMPLEMENTATION

Having analysed bias in data generation, we now present evidence of bias in the model building and implementation stage of the VoxCeleb SRC benchmark. In the ML pipeline this stage involves model definition and training, evaluation and real-world deployment. The types of bias that arise in these processes are *aggregation bias*, *learning bias*, *evaluation bias* and *deployment bias*. We found evidence of each type of bias in our evaluation.

6.1 Aggregation Bias

Aggregation bias arises when data contains underlying groups that should be treated separately, but that are instead subjected to uniform treatment.

We evaluate aggregation bias by plotting disaggregated DET performance curves for speaker subgroups based on nationality and gender. In Figure 3 we show the DET curves for female (left) and male (right) speakers across 11 nationalities for the *ResNetSE34V2* model evaluated on the *VoxCeleb 1-H* evaluation set. The dotted black DET curve shows the overall performance across all subgroups. DET curves above the dotted line have a high likelihood of performing worse than average, while DET curves below the dotted line will generally perform better than average. It is easy to see that the DET curves of female speakers lie mostly above the average DET curve, while those of male speakers lie below it. The model is thus likely to perform worse than average for females, and better for males. Figure 9 in the Appendix shows DET subplots for each nationality, highlighting disparate performance across nationalities.

The triangular markers show the FPR and FNR at the threshold $\theta_{@\ overall\ min}$ where the overall system DCF is minimized. The markers for male and female speaker subgroups are dispersed, indicating that the aggregate system calibration results in significant operating performance variability across subgroups. Table 4 in the Appendix shows the *subgroup bias* for all subgroups. With the exception of US female speakers, all females have a *subgroup bias* greater than 1, and thus perform worse than average.

The DET curves and *subgroup bias* demonstrate disparate performance based on speakers' gender and nationality. They also show that the model is fit to the dominant population in the training data, US speakers. The trends in aggregation bias that we observe for *ResNetSE34V2* are evident in all three evaluation sets, as well as *ResNetSE34L*. They indicate that speaker verification models do not identify all speaker subgroups equally well, and validate that performance disparities between male and female speakers identified in the past [27] still exist in DNN speaker verification models today.

6.2 Learning Bias

Learning bias concerns modeling choices and their effect on amplifying performance disparities across samples.

The *ResNetSE34V2* and *ResNetSE34L* models are built with different architectures and input features. The two architectures have been

[2] Discrimination based on national origin can have legal consequences, for instance, Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based on national origin in the United States

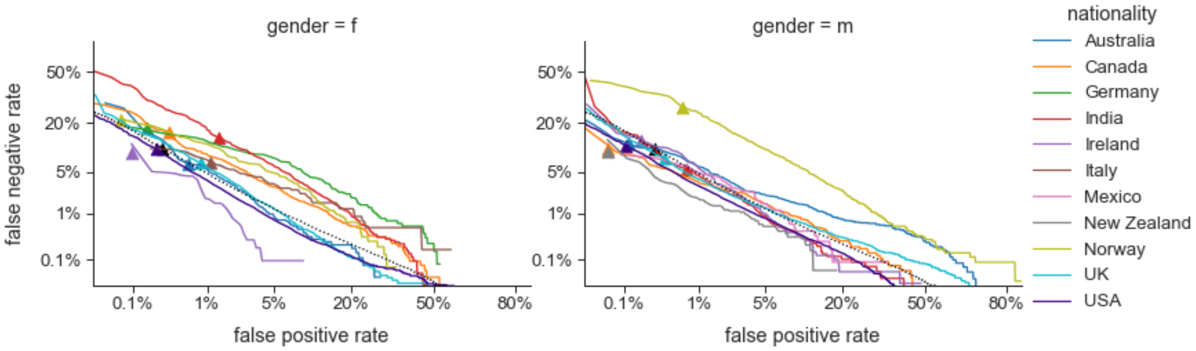

Figure 3: Aggregation bias in *ResNetSE34V2* with the *VoxCeleb 1-H* evaluation set: 1) the aggregate model (dotted black line) is fit to the dominant population (US speakers) and 2) the operating performance (triangular markers) across subgroups has high variability when the system is tuned to the overall threshold.

designed for different goals respectively: to optimise performance and to reduce inference time. Optimisation here refers to the design goal, not the optimiser of the model. The reduced number of parameters, smaller model size and reduced number of computations of *ResNetSE34L* are desirable attributes for on-device deployment.

In Figure 4 we plot the *subgroup bias* for both models for all subgroups. On the dotted line the models perform equally well for a subgroup. The greater the distance between a marker and the line, the greater the performance disparity between the models for that subgroup. As described in §4.3, subgroup performance is worse than average if the *subgroup bias* is greater than 1, and better than average if it is less than 1. We make three observations: Firstly, the *subgroup bias* for both models for male subgroups is close to or less than 1, indicating that at the threshold value males experience better than average performance for both models. Secondly, we observe that *subgroup bias* for male US speakers is equal for both models, indicating that performance disparities remain consistent for the over-represented group. Thirdly, we observe that neither of the two models reduces performance disparities definitively: *ResNetSE34V2* has a lower *subgroup bias* for 7 subgroups, *ResNetSE34L* for 10 subgroups.

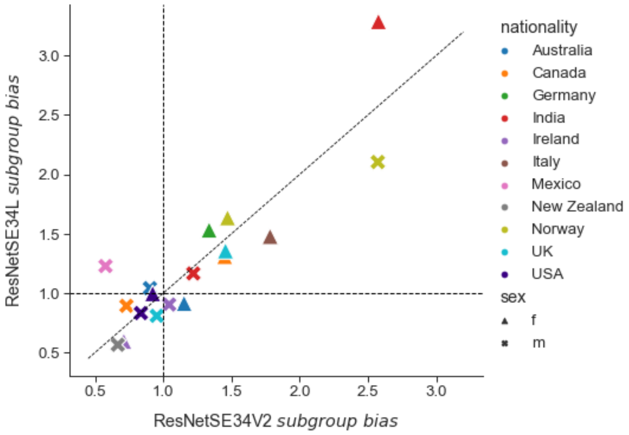

Figure 4: Learning bias based on model architecture. *Subgroup bias* for the performance optimized *ResNetSE34V2* is shown on the x-axis, and for the speed optimized *ResNetSE34L* on the y-axis. On the diagonal *subgroup bias* for the two models is equal.

In addition to examining *subgroup bias* we have plotted the DET curves for both models across subgroups in Figure 10 in the Appendix. We observe that the smaller *ResNetSE34L* increases the distance between DET curves for males and females with nationalities from the UK, USA and Ireland, indicating that the model increases performance disparities between male and female speakers of these nationalities. For Australian, Indian and Canadian speakers the distance between DET curves for males and females remains unchanged, while for Norwegian nationalities they lie closer together. Together these results point to learning bias, highlighting that modeling choices such as the architecture, the number of model parameters and the input feature dimensions can amplify performance disparities in speaker verification. The disparities tend to negatively affect female speakers and nationalities with few speakers. Our results reinforce other studies that have shown that bias can arise when reducing model size during pruning [16, 54], but are insufficient to point out the exact modeling choices that affect learning bias. This remains an area of future work.

6.3 Evaluation Bias

Evaluation bias is attributed to a benchmark population that is not representative of the user population, and to evaluation metrics that provide an oversimplified view of model performance.

6.3.1 *Evaluation Datasets.*

Representative benchmark datasets are particularly important during ML development, as benchmarks have disproportionate power to scale bias across applications if models overfit to the data in the benchmark [51]. Three evaluation sets can be constructed from the VoxCeleb 1 dataset to benchmark speaker verification models. *VoxCeleb 1 test* contains utterance pairs of 40 speakers whose name starts with *E*. *VoxCeleb 1-E* includes the *entire* dataset, with utterance pairs sampled randomly. *VoxCeleb 1-H* is considered a *hard* test set, that contains only utterance pairs where speakers have the same gender and nationality. Speakers have only been included in *VoxCeleb 1-H* if there are at least 5 unique speakers with the same gender and nationality. All three evaluation sets contain a balanced count of utterance pairs from same speakers and different speakers. We have calculated the speaker and utterance level demographics for each evaluation set from the dataset's metadata, and summarise the attributes of the evaluation sets in Table 2.

Several observations can be made based on the summary in Table 2: the VoxCeleb 1 dataset suffers from representation bias (see Section 5.2) and all three evaluation sets over-represent male speakers and US nationals. Furthermore, the sample size of *VoxCeleb 1 test* is too small to use it for a defensible evaluation. Its inclusion criterion based on speakers' names introduces additional representation bias into the evaluation set, as names strongly correlate with language, culture and ethnicity.

In addition to these obvious observations, the summary also reveals subtler discrepancies. Speaker verification evaluation is done on utterance pairs. Demographic representation is thus important on the speaker level to ensure that the evaluation set includes a variety of speakers, and on the utterance level to ensure that sufficient speech samples are included for each individual speaker. A significant mismatch in demographic representation between the speaker and utterance level is undesirable. If the representation of a subgroup is higher on the speaker level than the utterance level, this misrepresents the demographics that matter during evaluation and may indicate underrepresentation of individual speakers. Conversely, if the representation of a subgroup is lower on the speaker level, this increases the utterance count per speaker, suggesting overrepresentation of individual speakers. When considering utterances instead of speakers, the representation of females in relation to males decreases from 61% to 42% for *VoxCeleb 1 test*, from 82% to 72% for *VoxCeleb 1-E* and from 79% to 70% for *VoxCeleb 1-H*. The evaluation sets thus not only contain fewer female speakers, they also contain fewer utterances for each female speaker, which reduces the quality of evaluation for female speakers.

We evaluate *ResNetSE34V2* with the three evaluation sets and plot the resulting DET curves in Figure 5. The DET curve of *VoxCeleb 1 test* is irregular, confirming that this evaluation set is too small for a valid evaluation. In a FPR range between 0.1% and 5%, which is a reasonable operating range for speaker verification, model performance is similar on *VoxCeleb 1 test* and *VoxCeleb 1-E*. The curve of *VoxCeleb 1-H* lies significantly above the other two evaluation sets, indicating that the model performs worse on this evaluation set. Our empirical results illustrate that model performance is highly susceptible to the evaluation set, and show how evaluation bias can affect speaker verification models during evaluation.

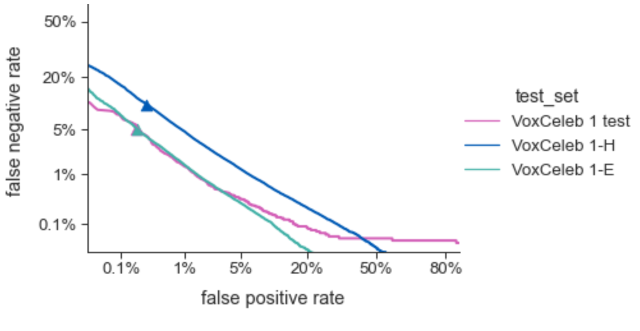

Figure 5: Evaluation bias in the three VoxCeleb 1 evaluation sets with *ResNetSE34V2*

6.3.2 *Evaluation Metrics.* The two dominant metrics used in speaker verification benchmarks, including the VoxCeleb SRC, are the equal error rate (EER) and the minimum value of the detection cost function $C_{Det}(\theta_{@\ overall\ min})$ (see Table 1). Both error metrics give rise to evaluation bias. The EER presents an oversimplified view of model performance, as it cannot weight false positives and false negatives differently. Yet, most speaker verification applications strongly favour either a low FPR or a low FNR [9]. The NIST SREs do not promote the use of the EER for speaker verification evaluation for this reason [9], which makes it particularly concerning that new challenges like the SUPERB benchmark evaluate only the EER [57]. $C_{Det}(\theta_{@\ overall\ min})$ can weight FPR and FNR, but has its own shortcomings. Firstly, the detection cost function has been updated over the years, and different versions of the metric are in use. This is impractical for consistent evaluation of applications across time. Secondly, the cost function is only useful if the FPR and FNR weighting reflect the requirements of the application. Determining appropriate weights is a normative design decision, and has received very limited attention in the research community. In benchmarks weights are typically not adjusted, which oversimplifies real-life evaluation scenarios. Finally, $C_{Det}(\theta_{@\ overall\ min})$ presents a limited view of a model's performance at a single threshold value. While DET curves can provide a holistic view on the performance of speaker verification models across thresholds, many recent research papers do not show them, and those that do only show aggregate curves.

The aggregate form of current evaluation practices based on and optimised for average performance hides the nature of harm that arises from evaluation bias. Ultimately, what matters when a speaker verification system is deployed, are the FPR and FNR. False positives pose a security risk, as they grant unauthorized speakers access to the system. False negatives pose a risk of exclusion, as they deny authorized speakers access to the system. We consider the FPR and FNR for subgroups at $C_{Det}(\theta_{@\ overall\ min})$ in relation to the average FPR and FNR in Table 5 in the Appendix. US male speakers have a FPR and FNR ratio of 1, indicating that this subgroup will experience error rates in line with the average. On the other end of the spectrum Indian female speakers have a FPR and FNR that are 13 and 1.3 times greater than average, indicating that this subgroup is exposed to a significant security risk, and a greater risk of exclusion.

6.4 Deployment Bias

Deployment bias arises when the application context and usage environment do not match the problem space as it was conceptualised during model development.

6.4.1 *Application Context.* Advancements in speaker verification research have been funded by governments to advance intelligence, defense and justice objectives [9]. The underlying use cases of speaker verification in these domains have been biometric identification and authentication. From this lens, the speaker verification problem space has been conceptualized to minimize false positives, which result in security breeches. Research on evaluation and consequently also model development has thus focused on attaining low FPRs. This dominant, but limited view promotes deployment bias in new use cases, which require evaluation practices and evaluation datasets tailored to their context.

	VoxCeleb 1 test	VoxCeleb 1-E	VoxCeleb 1-H
unique speakers	40	1 251	1 190
unique utterance pairs	37 720	579 818	550 894
speaker pairing details	-	random sample	same gender, nationality
speaker pair inclusion criteria	name starts with ’E’	all	>=5 same gender, nationality speakers
female / male speakers (%)	38 / 62	45 / 55	44 / 56
female / male utterances (%)	29.5 / 70.5	41.8 / 58.2	41.1 / 58.9
count of nationalities	9	36	11
top 1 nationality (% speakers / utterances)	US (62.5 / 59.6)	US (63.9 / 61.4)	US (67.1 / 64.7)
top 2 nationality (% speakers / utterances)	UK (12.5 / 13.9)	UK (17.2 / 18.3)	UK (18.1 / 19.3)
top 3 nationality (% speakers / utterances)	Ireland (7.5 / 6.7)	Canada (4.3 / 3.8)	Canada (4.5 / 3.9)

Table 2: VoxCeleb 1 evaluation sets show that the benchmark’s population is not representative across gender and nationality

Today, speaker verification is used in a wide range of audio-based applications, ranging from voice assistants on smart speakers and mobile phones to call centers. A low FPR is necessary to ensure system security. In voice assistants, false positives also affect user privacy, as positive classifications trigger voice data to be sent to service providers for downstream processing [46]. When used in forensic applications, false positives can amplify existing bias in decision-making systems, for example in the criminal justice system [17]. Even if the FPR is low, the speaker verification system will have a high FNR as trade-off, and the consequences of this must be considered. The FNR affects usability and can lead to a denial of service from voice-based user interfaces. The more critical the service, the higher the risk of harm associated with the FNR. Consider, for example, the previously mentioned speaker verification system used as proof-of-life of pensioners [31]. As long as the system is able to identify a pensioner correctly, it relieves the elderly from needing to travel to administrative offices, thus saving them time, money and physical strain. If the system has disparate FNR between demographic subgroups, some populations will be subjected to a greater burden of travel.

Evaluation practices aside, many speaker verification applications will suffer from deployment bias when evaluated on the utterance pairs in the VoxCeleb 1 evaluation datasets. Voice assistants in homes, cars, offices and public spaces are geographically bound, and speakers using them will frequently share a nationality, language and accent. These user and usage contexts should be reflected in the evaluation sets. The VoxCeleb 1 evaluation sets with randomly generated utterance pairs (i.e. *VoxCeleb 1 test* and *-E*) are inadequate to capture speaker verification performance in these application scenarios. Even *VoxCeleb 1-H*, which derives its abbreviation *-H* from being considered the *hard* evaluation set, is inadequate to evaluate speaker verification performance in very common voice assistant scenarios, such as distinguishing family members. Furthermore, the naming convention of the evaluation sets promotes a limited perspective on speaker verification application contexts: naming *VoxCeleb 1-H* the *hard* evaluation set creates a false impression that the randomly generated utterance pairs of *VoxCeleb 1-E* are the typical evaluation scenario.

6.4.2 Post-processing. The operating threshold of a speaker verification system is calibrated after model training (see §3.2). This post-processing step amplifies aggregation bias (discussed in §6.1) and deployment bias due to the application context (discussed above). The operating threshold is set in a calibration process that tunes a speaker verification system to a particular evaluation set. If the evaluation set does not take the usage environment and the characteristics of speakers in the environment into consideration, this can give rise to further deployment bias due to post-processing. As discussed above, the VoxCeleb 1 evaluation sets encompass a very limited perspective on application scenarios, and thresholds tuned to these evaluation sets will suffer from deployment bias due to post-processing in many contexts.

The speaker verification system threshold is typically calibrated for the overall evaluation set. This gives rise to a form of aggregation bias that arises during post-processing and deployment. Instead of calibrating the threshold to the overall evaluation set, it could be tuned for each subgroup individually. Using the detection cost function as example, this means setting the threshold for a subgroup to the value θ where $C_{Det}\,(\theta)$ is minimized for the subgroup (i.e. $C_{Det}\,\left(\theta_{@\ SG\ min}\right)^{SG}$). If the detection cost at the subgroup’s minimum is smaller than at the overall minimum, then the subgroup benefits from being tuned to its own threshold. By calculating the ratio of the subgroup’s overall detection cost and the subgroup’s minimum detection cost, we can get an intuition of the extent of bias. If the ratio is greater than 1, the subgroup will benefit from being tuned to its own threshold. The greater the ratio, the greater the bias and the more the subgroup will benefit from being tuned to its own minimum. Table 4 in the Appendix shows the ratios for all subgroups. It is clear that all subgroups would perform better if tuned to their own threshold. However, female speakers with a mean ratio of 1.37 will experience greater benefit from threshold tuning than male speakers with a mean ratio of 1.09. Visually, the effect of calibrating subgroups to their own threshold can be seen in Figure 11 in the Appendix.

7 DISCUSSION

In this paper we have presented an in-depth study of bias in speaker verification, the data processing technique underlying voice biometrics, and a core task in automated speaker recognition. We have provided empirical and analytical evidence of sources of bias at every stage of the speaker verification ML development workflow. Our study highlights that speaker verification performance degradation due to demographic attributes of speakers is significant, and can be attributed to aggregation, learning, evaluation, deployment, historical, representation and measurement bias. Our findings echo concerns similar to those raised in the evaluation for facial recognition technologies [43]. While our findings are specific to speaker verification, they can, for the most part, be extended to automated speaker recognition more broadly. Below we present recommendations for mitigating bias in automated speaker recognition and discuss limitations of our work.

7.1 Recommendations

7.1.1 Inclusive Evaluation Datasets for Real Usage Scenarios. We have shown that speaker verification evaluation is extremely sensitive to the evaluation set. The three evaluation sets specified for the VoxCeleb 1 dataset induce evaluation bias, and are insufficient for evaluating many real-world application scenarios. Representative evaluation datasets that are inclusive on a speaker and utterance level are thus needed. On an utterance level, an appropriate evaluation set should contain sufficient utterance pairs for all speakers, and pairs should reflect the reality of the application context. This requires guidelines for constructing application-specific utterance pairs for evaluation. As discussed in §5.3, our approach for constructing subgroups replicates measurement bias in the labelling choices of VoxCeleb. Future work should consider speaker groups based on vocal characteristics such as pitch, speaking rate, and vocal effort, and consider speaker diversity across languages and accents. Moreover, research on diversity and inclusion in subgroup selection [30] presents a starting point that can inform the design of more inclusive speaker verification evaluation datasets.

7.1.2 Evaluation Metrics that Consider Consequences of Errors. Considering the consequences of errors across application contexts is necessary to reduce deployment bias in speaker verification. Speaker verification evaluation and testing should carefully consider the choice and parameters of error metrics to present robust evaluations and comparison across models for specific application contexts. To this end, guidelines are needed for designing application specific error metrics, and for evaluating bias with these metrics. Such guidelines should determine acceptable FPR and FNR ranges, and guide normative decisions pertaining to the selection of weights of cost functions. Alternative evaluation metrics, such as those used for privacy-preserving speaker verification [25, 36], should also be studied for evaluating bias. To assess aggregation bias in speaker verification, disaggreated evaluation across speaker subgroups is needed. DET curves, which have history in speaker verification evaluation, should be used for visualizing model performance across speaker subgroups. Additionally error metrics should also be computed and compared across subgroups to mitigate evaluation bias.

7.1.3 Learning and Engineering Approaches for Mitigating Bias. Bias in speaker recognition is a new area of study, and interventions are needed to address *learning*, *deployment*, *aggregation* and *measurement bias*. We make some suggestions for interventions that can mitigate these types of bias. Speaker verification will improve for all subgroups if they are tuned to their own threshold rather than the overall threshold. Developing engineering approaches to dynamically select the optimal threshold for subgroups or individual speakers will improve the performance of speaker verification in deployed applications. Subgroup membership is typically not known at run time, making this a challenging task with potential trade-offs against privacy. Further research is also required to study how optimisation for on-device settings, such as model compression, pruning and small-footprint architectures, affect *learning bias*. Previous work in audio keyword spotting has shown that performance disparities across speaker subgroups can be attributed to model input features and the data sample rate at which the voice signal was recorded [53]. Studying and mitigating sources of *measurement bias* due to data processing and input features thus remain an important area for future work.

7.2 Limitations

Our work presents the first study of bias in speaker verification development and does not study bias in commercial products, which we position as an area for future work. Our aim was to study typical development and evaluation practices in the speaker verification community, not to compare speaker verification algorithms. We thus designed a case study with a confined scope, using publicly available benchmark models as black box predictors. Our findings should be interpreted with this in mind, and not be seen as a generic evaluation for all speaker verification models. We constructed demographic subgroups based on those included in the *VoxCeleb1-H* evaluation set. Some subgroups thus have insufficient sample sizes, which affects the quality of our empirical evaluation for these subgroups. However, as discussed in detail in §6.3, small subgroups are in themselves a source of representation bias that needs to be addressed. We observed that the performance difference that we identified between male and female speakers, and across nationalities, persist across small and large subgroups.

8 CONCLUSION

Automated speaker recognition is deployed on billions of smart devices and in services such as call centres. In this paper we study bias in speaker verification, the biometrics of voice, which is a core task in automated speaker recognition. We present an in-depth empirical and analytical study of bias in a benchmark speaker verification challenge, and show that bias exists at every stage of the machine learning development workflow. Most affected by bias are female speakers and non-US nationalities, who experience significant performance degradation due to aggregation, learning, evaluation, deployment, historic and representation bias. Our findings lay a strong foundation for future work on bias and fairness in automated speaker recognition.

ACKNOWLEDGMENTS

This research was partially supported by projects funded by EU Horizon 2020 research and innovation programme under GA No. 101021808 and GA No. 952215.

REFERENCES

[1] Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? *INTERSPEECH* (2005), 2205–2208. https://www.isca-speech.org/archive/interspeech_2005/addadecker05_interspeech.html

[2] Zhongxin Bai and Xiao Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99. https://doi.org/10.1016/j.neunet.2021.03.004

[3] Tolga Bolukbasi, Kai-wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker ? Debiasing Word Embeddings. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems.* 4356 – 4364.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research: Conference on Fairness, Accountability, and Transparency*, Vol. 81. 1889–1896.

[5] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong Jin Lee, and Icksang Han. 2020. In defence of metric learning for speaker recognition. *Proceedings of the Annual*

Conference of the International Speech Communication Association, INTERSPEECH 2020-Octob (2020), 2977–2981. https://doi.org/10.21437/Interspeech.2020-1064

[6] Joon Son Chung and Andrew Zisserman. 2017. Out of time: Automated lip sync in the wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10117 LNCS, i (2017), 251–263. https://doi.org/10.1007/978-3-319-54427-4{_}19

[7] Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. 2021. Fair Voice Biometrics : Impact of Demographic Imbalance on Group Fairness in Speaker Recognition. (2021), 1892–1896.

[8] Sadaoki Furui. 1994. An Overview of Speaker Recognition Technology. In *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification.* 1 – 9.

[9] Craig S. Greenberg, Lisa P. Mason, Seyed Omid Sadjadi, and Douglas A. Reynolds. 2020. Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech and Language* 60 (2020). https://doi.org/10.1016/j.csl.2019.101032

[10] Oxford Visual Geometry Group. 2021. The VoxCeleb Speaker Recognition Challenge 2021. https://www.robots.ox.ac.uk/~vgg/data/voxceleb/competition2021.html

[11] John H.L. Hansen and Taufiq Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine* 32, 6 (2015), 74–99. https://doi.org/10.1109/MSP.2015.2462851

[12] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* (2016), 3323–3331.

[13] Khaled Hechmi, Trung Ngo Trong, Ville Hautamaki, and Tomi Kinnunen. 2021. VoxCeleb Enrichment for Age and Gender Recognition. (2021). http://arxiv.org/abs/2109.13510

[14] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-End Text-Dependent Speaker Verification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 5115–5119.

[15] Hee Soo Heo, Bong Jin Lee, Jaesung Huh, and Joon Son Chung. 2020. Clova baseline system for the VoxCeleb speaker recognition challenge 2020. *arXiv* (2020), 1–3.

[16] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising Bias in Compressed Models. https://arxiv.org/abs/2010.03058

[17] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[18] Elie Khoury, Laurent El Shafey, Christopher McCool, Manuel Günther, and Sébastien Marcel. 2014. Bi-modal biometric authentication on mobile phones in challenging conditions. *Image and Vision Computing* 32, 12 (2014), 1147–1160. https://doi.org/10.1016/j.imavis.2013.10.001

[19] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, Z. Boulkcnafet, L. M. Mazaira Fernandez, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Gunther, J. Zganec-Gros, R. Zazo Candil, F. Simoes, M. Bengherabi, A. Alvarez Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz, D. Van Leeuwen, J. Gonzalez-Dominguez, M. Uliani Neto, E. Boutellaa, P. Gomez Vilda, A. Varona, D. Petrovska-Delacretaz, P. Matejka, J. Gonzalez-Rodriguez, T. Pereira, F. Harizi, L. J. Rodriguez-Fuentes, L. El Shafey, M. Angeloni, G. Bordel, G. Chollet, and S. Marcel. 2013. The 2013 speaker recognition evaluation in mobile environment. *Proceedings - 2013 International Conference on Biometrics, ICB 2013* (2013). https://doi.org/10.1109/ICB.2013.6613025

[20] Davis E. King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[21] Tomi Kinnunen and Haizhou Li. 2009. An Overview of Text-Independent Speaker Recognition : from Features to Supervectors. *Speech Communication* 52, 1 (2009), 12. https://doi.org/10.1016/j.specom.2009.08.009

[22] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *PNAS* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117/-/DCSupplemental.y

[23] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. 2017. Deep speaker: An end-to-end neural speaker embedding system. *arXiv* (2017).

[24] Beryl Lipton and Quintin Cooper. 2021. The Catalog of Carceral Surveillance: Voice Recognition and Surveillance. https://www.eff.org/deeplinks/2021/09/catalog-carceral-surveillance-voice-recognition-and-surveillance

[25] Mohamed Maouche, Brij Mohan, Lal Srivastava, Nathalie Vauquier, Marc Tommasi, Emmanuel Vincent, Mohamed Maouche, Brij Mohan, Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, Mohamed Maouche, Brij Mohan, Lal Srivastava, Nathalie Vauquier, Emmanuel Vincent, and De Lorraine. 2020. A comparative study of speech anonymization metrics. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.* Shanghai, China.

[26] A Martin, G Doddington, T Kamm, M Ordowski, and M Przybocki. 1997. *The DET Curve in Assessment of Detection Task Performance.* Technical Report. National Institute of Standards and Technology (NIST), Gaithersburg MD. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.4489

[27] Luis Miguel Mazaira-Fernandez, Agustín Álvarez-Marquina, and Pedro Gómez-Vilda. 2015. Improving speaker recognition by biometric voice deconstruction. *Frontiers in Bioengineering and Biotechnology* 3, September (2015), 1–19. https://doi.org/10.3389/fbioe.2015.00126

[28] M McLaren, L Ferrer, D Castan, and A Lawson. 2016. The Speakers in the Wild (SITW) speaker recognition database.. In *Interspeech.* pdfs.semanticscholar.org. https://pdfs.semanticscholar.org/3fe3/58a66359ee2660ec0d13e727eb8f3f0007c2.pdf

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv* (2019).

[30] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), 117–123. https://doi.org/10.1145/3375627.3375832

[31] Marta Morrás. 2021. BBVA Mexico allows its pensioner customers to provide proof of life from home thanks to Veridas voice biometrics. https://veridas.com/en/bbva-mexico-allows-pensioner-customers-provide-proof-of-life-from-home/

[32] Arsha Nagrani, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman. 2020. VoxSRC 2020: The Second VoxCeleb Speaker Recognition Challenge. (2020). http://arxiv.org/abs/2012.06867

[33] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech and Language* 60 (2020), 101027. https://doi.org/10.1016/j.csl.2019.101027

[34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *arXiv* (2017), 2616–2620.

[35] Andreas Nautsch, Abelino Jim, Mohamed Amine, Aymen Mtibaa, Mohammed Ahmed, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-barrero, and Dijana Petrovska-delacr. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech and Language* 58 (2019), 441–480. https://doi.org/10.1016/j.csl.2019.06.001

[36] Andreas Nautsch, Jose Patino, Natalia Tomashenko, Junichi Yamagishi, Paul Gauthier Noé, Jean François Bonastre, Massimiliano Todisco, and Nicholas Evans. 2020. The privacy ZEBRA: Zero evidence biometric recognition assessment. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-Octob (2020), 1698–1702. https://doi.org/10.21437/Interspeech.2020-1815

[37] NIST. 2019. NIST 2019 Speaker Recognition Evaluation Plan. 1 (2019), 1–7.

[38] NIST. 2020. *NIST 2020 CTS Speaker Recognition Challenge Evaluation Plan.* Technical Report. 1–8 pages.

[39] Soo Jin Park, Caroline Sigouin, Jody Kreiman, Patricia Keating, Jinxi Guo, Gary Yeung, Fang-Yu Kuo, and Abeer Alwan. 2016. Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition.. In *Interspeech 2016.* ISCA, San Francisco, CA, USA. https://doi.org/10.21437/Interspeech.2016-523

[40] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference.*

[41] Xiaoyi Qin, Ming Li, Hui Bu, Wei Rao, Rohan Kumar Das, Shrikanth Narayanan, and Haizhou Li. 2020. The INTERSPEECH 2020 far-field speaker verification challenge. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-Octob (2020), 3456–3460. https://doi.org/10.21437/Interspeech.2020-1249

[42] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), 429–435. https://doi.org/10.1145/3306618.3314244

[43] Inioluwa Deborah Raji and Genevieve Fried. 2021. About Face: A Survey of Facial Recognition Evaluation. (2021). http://arxiv.org/abs/2102.00813

[44] Douglas A. Reynolds. 2002. An Overview of Automatic Speaker Recognition Technology. *IEEE* (2002).

[45] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359246

[46] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. 2020. Unacceptable, where is my privacy? Exploring Accidental Triggers of Smart Speakers. (8 2020). http://arxiv.org/abs/2008.00508

[47] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015), 1–14.

[48] Rita Singh. 2019. *Profiling Humans from their Voice.* https://doi.org/10.1007/978-981-13-8403-5

[49] D Snyder, D Garcia-Romero, D Povey, and S Khudanpur. 2017. Deep Neural Network Embeddings for Text-Independent Speaker Verification. *Interspeech*

(2017). https://www.isca-speech.org/archive/Interspeech_2017/pdfs/0620.PDF

[50] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5329–5333.

[51] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization*.

[52] Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2017-Augus (2017), 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[53] Wiebke Toussaint, Akhil Mathur, Aaron Yi Ding, and Fahim Kawsar. 2021. Characterising the Role of Pre-Processing Parameters in Audio-based Embedded Machine Learning. In *The 3rd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things (AIChal- lengeIoT 21)*. Association for Computing Machinery, Coimbra, Portugal, 439–445. https://doi.org/10.1145/3485730.3493448

[54] Wiebke Toussaint, Akhil Mathur, Fahim Kawsar, and Aaron Yi Ding. 2022. Tiny, always-on and fragile: Bias propagation through design choices in on-device machine learning workflows. (2022), 19 pages. http://arxiv.org/abs/2201.07677

[55] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Bias Preservation in Machine Learning : The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review, Forthcoming* (2021), 1–51. https://ssrn.com/abstract=3792772

[56] Wikipedia contributors. 2022. List of languages by number of native speakers in India. https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India [Online; accessed 6-May-2022].

[57] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark. (2021). http://arxiv.org/abs/2105.01051

[58] Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukas Burget. 2020. *Short-duration Speaker Verification (SdSV) Challenge 2021: the Challenge Evaluation Plan*. Technical Report. 1–13 pages. http://arxiv.org/abs/1912.06311

A APPENDIX

A.1 Speaker Verification Evaluation

Example Speaker Verification Output Score Distributions

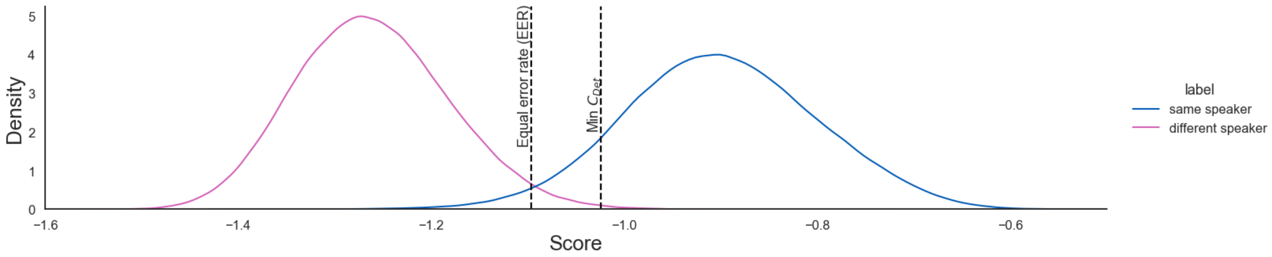

Figure 6: Distribution of speaker verification scores: blue are same speaker trials, pink are different speaker trials. The dotted lines are possible threshold values. Scores to the left of a threshold are rejected, scores to the right are accepted.

Example Detection Error Trade-off Curves

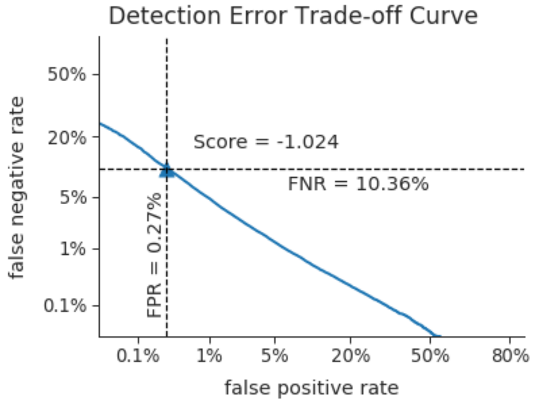

Figure 7: Detection Error Trade-off (DET) curve of a speaker verification system: the blue line shows false positive and false negative error rates at different score values. For example, at the blue triangle the score = -1.024, FPR = 0.27% and FNR = 10.36%

Summary of Technical Details of VoxCeleb SRC Baseline Models

Model	*ResNetSE34V2*	*ResNetSE34L*
Published in:	[15]	[5]
Alternative name in publication:	performance optimised model, H/ASP	Fast ResNet-34
Additional training procedures:	data augmentation (noise & room impulse response)	-
Parameters:	8 million	1.4 million
Frame-level aggregation:	attentive statistical pooling	self-attentive pooling
Loss function:	angular portotypical softmax loss	angular portotypical loss
Input features:	64 dim log Mel filterbanks	40 dim Mel filterbanks
Window (width x step):	25ms x 10ms	25ms x 10ms
Optimized for:	predictive performance	fast execution

Table 3: Attributes of two VoxCeleb SRC baseline models

A.2 Representation Bias

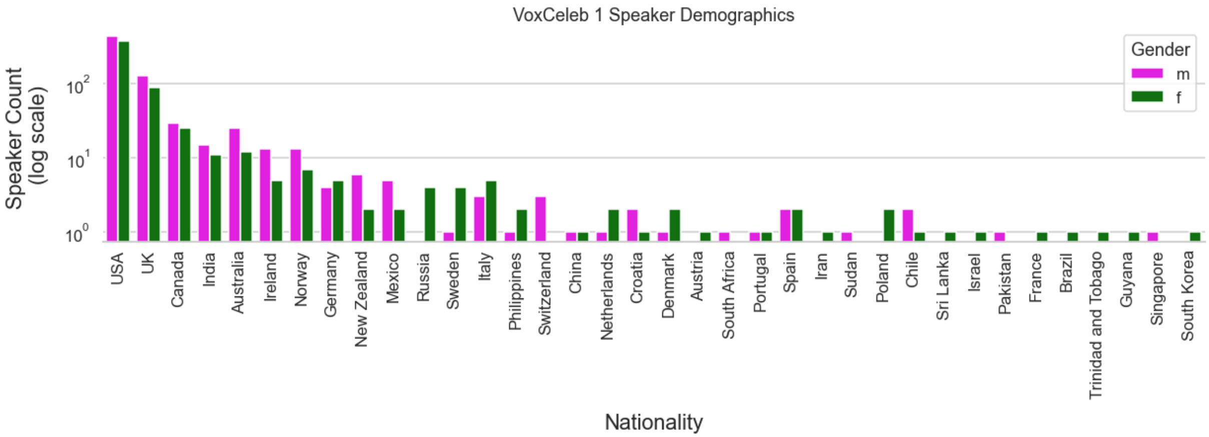

Figure 8: VoxCeleb 1 Speaker Demographics

A.3 Evaluation Bias

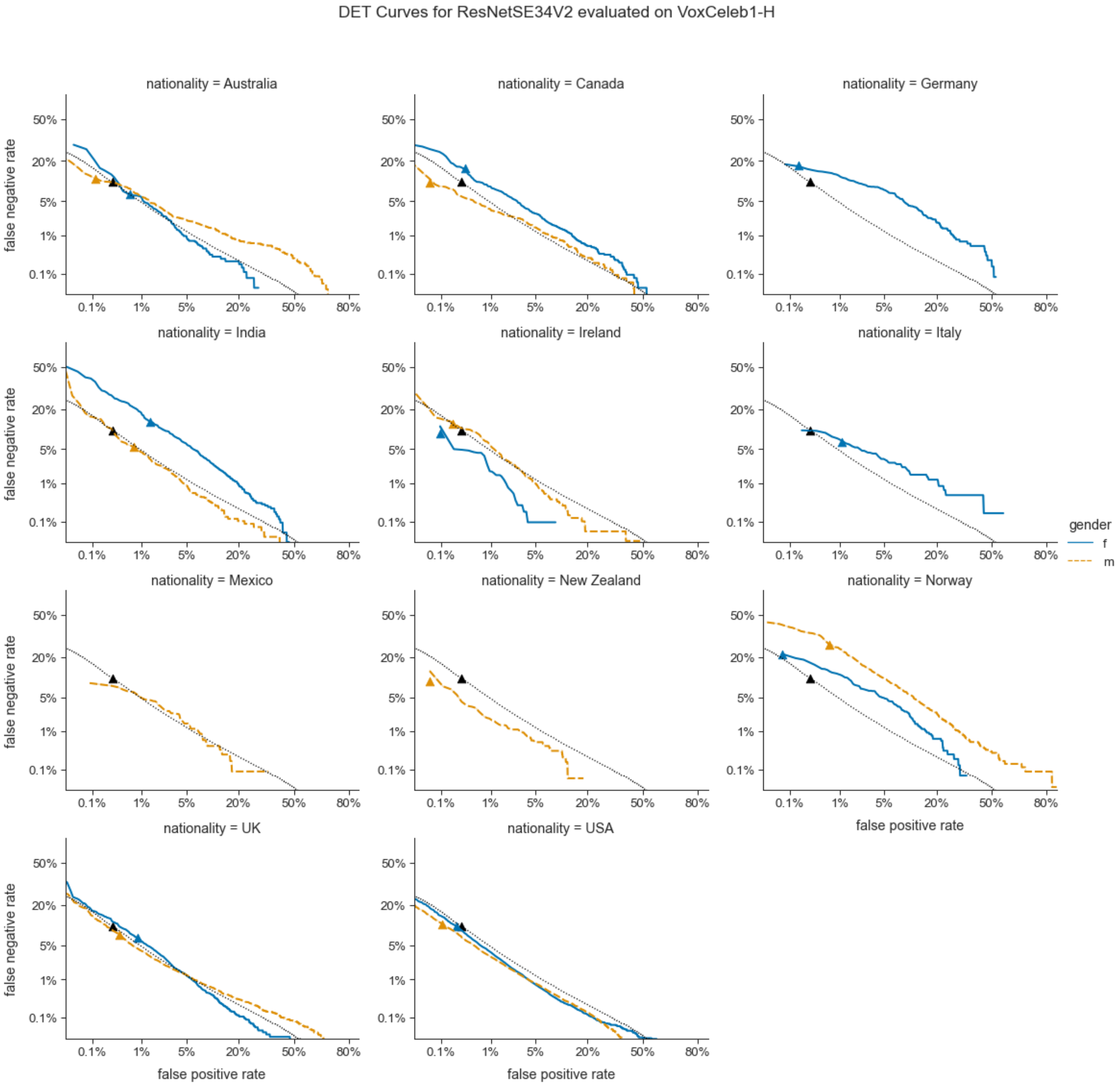

Figure 9: *ResNetSE34V2* DET curves for speaker subgroups evaluated on the *VoxCeleb 1-H* evaluation set. The dotted black lines indicate the aggregate overall DET curve across all subgroups.

A.4 Learning Bias

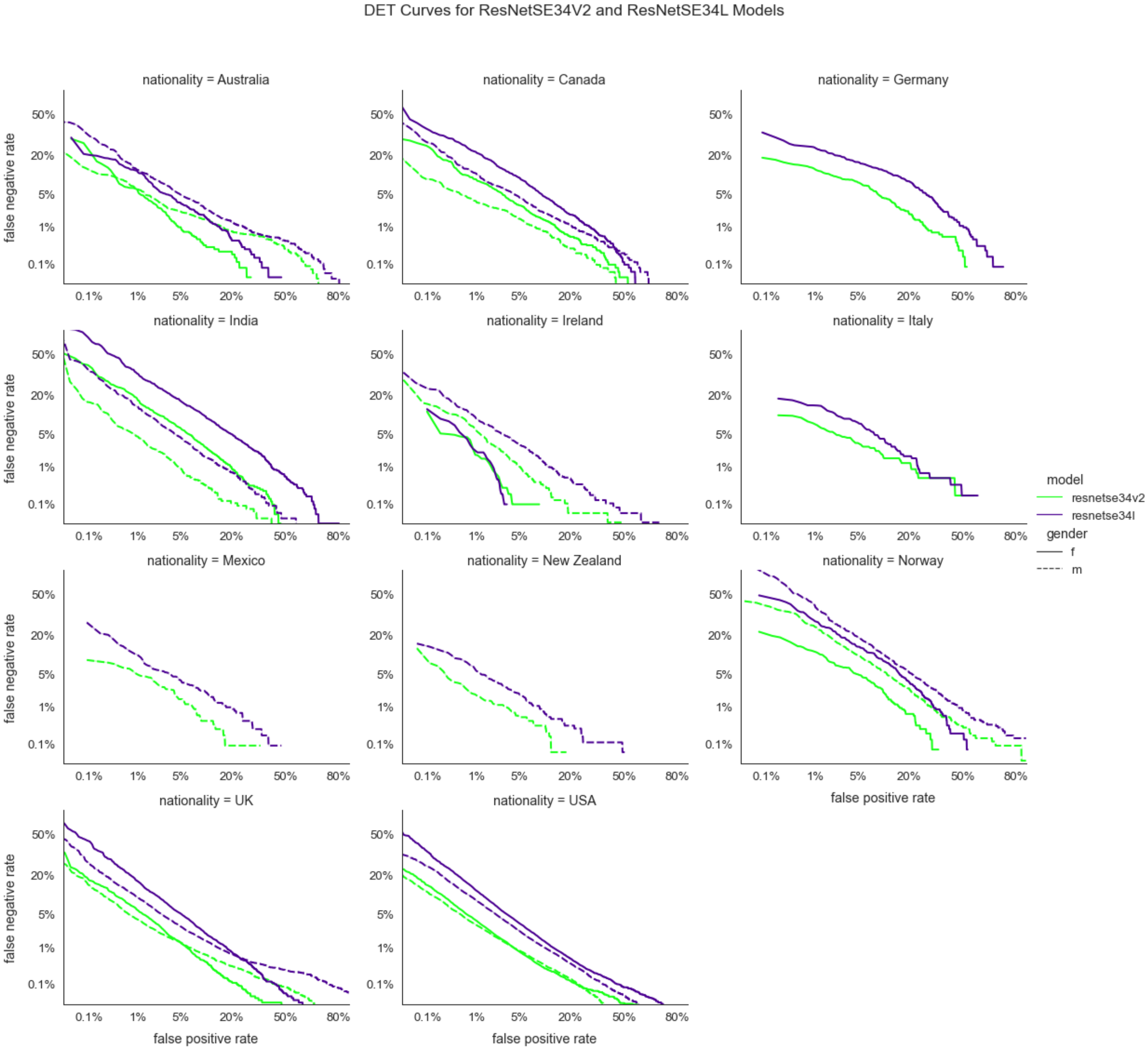

Figure 10: Learning bias based on model architecture

A.5 Aggregation and Post-processing Deployment Bias

Subgroup (SG)	Unique speakers	$C_{Det}(\theta_{@\ overall\ min})^{SG}$	*subgroup bias*	$C_{Det}(\theta_{@\ SG\ min})^{SG}$	*threshold bias*
mexico_m	5	0.090	0.5768	0.090	1.0000
newzealand_m	6	0.104	0.6668	0.086	1.2093
ireland_f	5	0.110	0.7109	0.070	1.5714
canada_m	29	0.114	0.7304	0.104	1.0962
usa_m	431	0.130	0.8357	0.122	1.0656
australia_m	25	0.140	0.9020	0.136	1.0294
usa_f	368	0.142	0.9224	0.140	1.0143
uk_m	127	0.148	0.9523	0.140	1.0571
ireland_m	13	0.162	1.0432	0.160	1.0125
australia_f	12	0.178	1.1523	0.154	1.1558
india_m	15	0.190	1.2200	0.144	1.3194
germany_f	5	0.208	1.3359	0.184	1.1304
canada_f	25	0.224	1.4501	0.202	1.1089
uk_f	88	0.226	1.4558	0.172	1.3140
norway_f	7	0.228	1.4711	0.210	1.0857
italy_f	5	0.276	1.7827	0.104	2.6538
norway_m	13	0.398	2.5720	0.396	1.0051
india_f	11	0.400	2.5766	0.318	1.2579

Table 4: Detection costs, *subgroup bias* and post-processing aggregation bias (see Equation 3) for subgroups at overall and subgroup minimum thresholds with $C_{Det}(\theta_{@\ overall\ min})^{overall} = 0.154$. Subgroups above the horizontal black line have a *subgroup bias* less than 1 and perform better than average when tuned to $C_{Det}(\theta_{@\ overall\ min})$. Female subgroups are on average subjected to more bias than male subgroups.

$$threshold\ bias = \frac{C_{Det}\left(\theta_{@\ overall\ min}\right)^{SG}}{C_{Det}\left(\theta_{@\ SG\ min}\right)^{SG}} \tag{3}$$

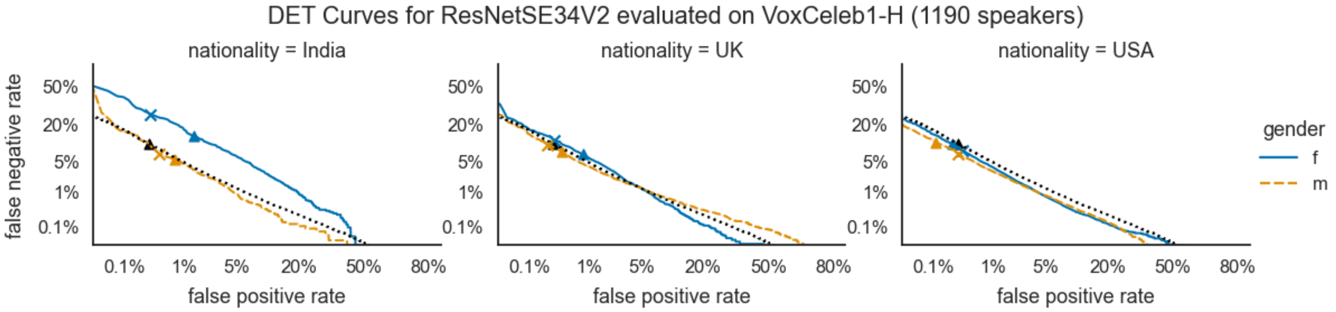

Figure 11: DET curves and thresholds for male and female speakers of Indian, UK and USA nationalities for ResNetSE34V2 evaluated on the VoxCeleb1-H test set.

Figure 11 shows DET curves and thresholds for ResNetSE34V2 for male and female speakers of Indian, UK and USA nationalities evaluated on *VoxCeleb 1-H*. We use the following conventions: triangle markers show the FPR and FNR at the overall minimum threshold $C_{Det}\left(\theta_{@\ overall\ min}\right)^{SG}$, cross markers show the FPR and FNR at the subgroup minimum threshold $C_{Det}\left(\theta_{@\ SG\ min}\right)$, and dotted black lines and markers are used for the overall DET curve and threshold. The DET curve of female Indian speakers lies far above the overall aggregate, indicating that irrespective of the threshold, the model will always perform worse than aggregate for this subgroup. In the operating region around the tuned thresholds, the model also performs worse for female speakers from both the UK and the USA. Being tuned to $C_{Det}\left(\theta_{@\ overall\ min}\right)$ does not affect the FNR and improves the FPR of USA female and male speakers. For other speaker subgroups, especially UK females and Indian females and males, either the FPR or the FNR deteriorates significantly when tuned to the overall minimum. For all subgroups the threshold at the subgroup minimum, $C_{Det}\left(\theta_{@\ SG\ min}\right)$, shifts the FPR and FNR closer to those of the minimum overall threshold, suggesting that performance will improve when optimising thresholds for subgroups individually.

A.6 Application Context Deployment Bias

Subgroup	Unique speakers	FPR ratio overall	FNR ratio overall
mexico_m	5	0.0000	0.8173
canada_m	29	0.5171	0.9396
newzealand_m	6	0.5218	0.8487
norway_f	7	0.6306	1.9682
ireland_f	5	0.9037	0.8408
usa_m	431	1.0000	1.0000
australia_m	25	1.1055	1.0745
germany_f	5	1.5023	1.6162
ireland_m	13	1.6864	1.1675
usa_f	368	2.0542	0.9287
canada_f	25	3.1483	1.4749
uk_m	127	3.5339	0.6986
australia_f	12	5.6031	0.6008
norway_m	13	6.0866	2.5233
india_m	15	6.6852	0.4975
uk_f	88	7.8514	0.6168
italy_f	5	10.3484	0.6202
india_f	11	13.0387	1.2497

Table 5: FPR and FNR ratios for subgroups at $C_{Det}(\theta_{@\ overall\ min})$. The ratio is calculated by dividing the subgroup FPR and FNR by the overall FPR and FNR respectively. It thus presents a relative view on how much better or worse the subgroup error rates are in relation to the overall error rates.