

Exploring Algorithmic Fairness in Deep Speaker Verification

Gianni Fenu, Hicham Lafhouli, and Mirko Marras(✉)
{fenu,h.lafhouli,mirko.marras}@unica.it

Department of Mathematics and Computer Science, University of Cagliari,
Via Ospedale 72, 09124 Cagliari, Italy

Abstract. To allow individuals to complete voice-based tasks (e.g., send messages or make payments), modern automated systems are required to match the speaker's voice to a unique digital identity representation for verification. Despite the increasing accuracy achieved so far, it still remains under-explored how the decisions made by such systems may be influenced by the inherent characteristics of the individual under consideration. In this paper, we investigate how state-of-the-art speaker verification models are susceptible to unfairness towards legally-protected classes of individuals, characterized by a common sensitive attribute (i.e., gender, age, language). To this end, we first arranged a voice dataset, with the aim of including and identifying various demographic classes. Then, we conducted a performance analysis at different levels, from equal error rates to verification score distributions. Experiments show that individuals belonging to certain demographic groups systematically experience higher error rates, highlighting the need of fairer speaker recognition models and, by extension, of proper evaluation frameworks.

Keywords: Speaker recognition · Algorithmic fairness · Deep learning

1 Introduction

More and more systems and platforms are profiling individuals based on physical and behavioral characteristics. The resulting applicative scenarios range from accessing to smartphones to checking identities in online exams, from controlling autonomous vehicles to interact with robots, from assisting lawyers to delivering personalized services [6,7,23,28]. In these situations, where characterizing the current user is critical, one of the most prominent biometrics finding application across industries is based on human voices. The corresponding market is set to grow at a very high rate of 19.4% by 2021 [1]. Enabling voice commands into automated systems may support people to manage sensitive data (e.g., messages) or complete actions (e.g., shopping online) in a more natural way. However, handing over this responsibility to automated systems implies that they must be able to match the speaker's voice to a unique digital representation for identification or authentication, guaranteeing the security of the related applications [16].

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12252, pp. 77–93, 2020.
https://doi.org/10.1007/978-3-030-58811-3_6

Speaker recognition has been actively studied over the years, and has recently undergone a revolution thanks to deep-learned acoustic representations. The latter outperformed hand-crafted features based on Gaussian mixture models [29], joint factor analysis [10], or i-Vectors [21]. Modern systems extract fine-grained acoustic feature representations from pre-trained deep neural networks. State-of-the-art solutions include x-Vector [31] and ResNet [25] models, among others. Despite high overall accuracy, speaker verifiers may be influenced by the inherent characteristics of the individuals[1], and suffer from demographic bias that causes certain populations to experience higher error rates than others. This deficiency can even put certain groups more at risk than others against impostors attacks [24]. While such a bias could be attributed to the lack of diverse training data, it nevertheless brings into question the unfairness of speaker recognition models and its mitigation, also going beyond gender (e.g., age, nationality).

In this paper, we seek to analyze speaker recognition models, using a fairness-aware perspective, in order to understand how demographic characteristics affect model performance. We conducted an offline evaluation of different speaker verification models based on *ThinResNet* [25], currently considered among the state-of-the-art models for speaker recognition in the wild. We compared the performance against algorithmic biases, by assessing: (i) how the equal error rates vary when considering models trained for different languages, (ii) how algorithmic bias affects verification score distributions and the related false acceptance and false rejection rates, and (iii) how biases related to language, gender, and age of a user propagate in the performance achieved by the considered speaker verification models. These biases involving legally-protected classes of people might have operational implications (e.g., a language bias might affect speaker verification adoption in certain geographical areas, while a gender bias might limit men or women in using this technology). Our contribution is threefold:

- We propose a general framework for inspecting algorithmic bias in machine-learning tasks tailored to the context of deep speaker verification[2].
- We perform a fairness-aware analysis of speaker verification based on a new benchmark setup, with identities coming from 16 demographic groups.
- We assess the impact of demographic attributes (i.e., language, gender, and age) in the feature representation of deep speaker verification models.

The rest of this paper is structured as follows. Section 2 depicts the related work in speaker verification and fairness-aware machine learning. Then, Sect. 3 defines the components of a general speaker verification systems, data, protocols, and models, while Sect. 4 provides a performance analysis, with results and discussion. Finally, Sect. 5 states final remarks and insights for future research.

2 Related Work

The research presented in this paper relies on literature from both the speech community and the machine learning community.

[1] In this paper, we will use the terms "individuals" and "users" interchangeably.
[2] Code, data, and models are available at https://mirkomarras.github.io/fair-voice/.

2.1 Deep Speaker Verification

Traditional speaker recognition systems based on hand-crafted solutions relied on *Gaussian Mixture Models* (GMMs) [29] that are trained on low dimensional feature vectors, *Joint Factor Analysis* (JFA) [10] methods that model speaker and channel subspaces separately, or *i-Vectors* [21] that attempt to embed both subspaces into a single compact, low-dimensional space.

Modern systems leverage deep-learned acoustic representations, i.e., embeddings, extracted from one of the last layers of a neural network trained for standard or one-shot speaker classification [20,22]. The most prominent examples include *d-Vectors* [34], *c-Vectors* [8], *x-Vectors* [31], *VGGVox-Vectors* [26] and *ResNet-Vectors* [9]. Furthermore, deep learning frameworks with end-to-end loss functions have recently drawn attention to train speaker discriminative embeddings [19]. Their results proved that end-to-end systems with embeddings achieved better performance on short utterances, which are common in several contexts (e.g., robotics and proctoring), compared with traditional systems.

Speaker verification aims to confirm or refute the identify of a speaker based on an enrolled speech model. The user is asked to provide several samples of his speech, and the utterances are stored as a collection of acoustic feature vectors. Depending on the policy, the presented vocal input may be compared with all the vectors or with a single combined vector to make the verification decision.

2.2 Fairness in Machine Learning

Research on fairness in machine learning has been growing rapidly, as more and more automated decision systems are being deployed in highly sensitive areas that affect human lives and society [33]. The decisions have prediction problems at their core, and machine learning models have been used to maximize prediction performance in all these contexts. Generally, a decision is considered fair if it does not discriminate against people on the basis of their membership to a legally-protected group, such as sex or race [4]. For instance, the right to non-discrimination is embedded in the European normative framework, such as *Art. 21* of the *EU Charter of Fundamental Rights*, *Art. 14* of the *European Convention on Human Rights*, and *Artt. 18–25* of the *Treaty on the Functioning of the European Union* [15]. Ensuring that these models are less biased and adhere to the respective standards of fairness is difficult. There is overwhelming evidence showing that they can inherit or even perpetuate human biases in their decision, when trained on data that contain biased human decisions [27,30].

In practice, there are several definitions of algorithmic fairness that try to achieve this goal, mainly classified in individual fairness and group fairness. In addition, two categories of statistical fairness definitions are based on predicted classifications and a combination of predicted classifications or risk scores and actual outcomes. The first definition, i.e., demographic parity, is based on predicted classifications [12]. Demographic parity is fulfilled if people from different protected groups has on average equal classifications. Another definition implies that, if the classifier gets it wrong, it should be equally wrong for all protected

groups, since being more wrong for one group would result in harmful outcomes for this group compared to the other ones [17]. Hence, false negative and false positive rates should be equal across different protected groups.

Soft biometrics received increasing attention from the predictive perspective over the past decade [3,5]. Only relatively recently, bias- and fairness-oriented perspectives have emerged as an important research area, mainly focusing on face recognition systems [11]. For instance, the authors in [32] proposed a novel fair score normalization approach that is specifically designed to reduce the effect of bias in face recognition, and subsequently lead to a significant overall performance boost. Similarly, the authors in [2,14] showed the existence of demographic bias in the face representations returned by deep-learning-based face recognition models. These studies exposed concrete situations that may lead to a systematic discrimination of certain demographic groups. Moreover, other researchers analyzed the demographic bias in iris presentation attack detection algorithms, and showed that female users are significantly less protected by the attack detector in comparison to male users [13]. While bias and fairness in biometrics are receiving more and more attention, their impact on speaker verification is currently under-explored, motivating us to conduct this study.

3 Methodology

In this section, we formalize an experimental framework, including the speaker verification task, data, protocols, models, and their implementation details, underlying our research on bias and fairness in speaker verification.

3.1 Preliminaries

Let $A \subset \mathbb{R}^*$ denote the domain of audio waveforms with unknown length. We consider a traditional two-step processing pipeline with an intermediate visual acoustic representation $S \subset \mathbb{R}^{k \times *}$ (e.g., a spectrogram or a filterbank), and an explicit feature extraction step carried out by a speaker verification model θ, which produces fixed-length representations in $D \subset \mathbb{R}^e$. We denote the respective stages as $\mathcal{F} : A \rightarrow S$ and $\mathcal{D}_\theta : S \rightarrow D$. Given a *verification policy* p, a *decision threshold* τ, a speaker verification model θ, and N *enrolled utterances* per user, a speaker verification system can be defined as a function:

$$v_{p,\tau,\theta} : D \times D_u^N \rightarrow \{0, 1\} \tag{1}$$

which compares an input feature vector d from an unknown user with a set of enrolled feature vectors $d_u^1, ..., d_u^N$ from user u to confirm or refute the speaker's identity (1 and 0, respectively). We mainly consider a verification policy, which rely on a similarity function $\mathcal{S} : D \times D \rightarrow [0, 1]$, defined as follows:

$$v_{p,\tau,\theta} = any\left(\{\mathcal{S}(d, d_u^i) > \tau : i \in 1, \ldots, N\}\right) \tag{2}$$

where the identity of the current user is confirmed if any of the enrolled speech vectors has a similarity with the probe vector higher than the threshold. Our study considers $N = 1$ (one-shot setup), i.e., one vector as enrolment for a user.

3.2 Data

In this study, we leveraged speakers' data collected from Common Voice[3], one of the largest corpora including unconstrained speech of people, extracted from real-world scenarios, and featuring diverse acoustic environments. All waveforms were single-channel, 16-bit recordings sampled at 16 kHz. Such data was initially collected by the Mozilla Foundation in order to train an open source Speech-to-Text engine, and for this reason was not suitable for fairness analysis in speaker verification. However, to the best of our knowledge, except Common Voice, there is no other public dataset which comprises voice data coming from a range of languages and labelled with sensitive attributes, such as gender, accent, and age. Hence, after performing some modifications, which are detailed in the next subsection, we arranged Common Voice data to fit with our purposes.

The data we collected included individuals who declare sensitive attributes suitable for fairness analysis and span different languages (i.e., Chinese, French, German, English, and Kabyle). Even though Common Voice includes data from a wider range of languages, we selected the aforementioned languages since they are composed by enough utterances to conduct a statistically-relevant fairness analysis on verification models. Figure 1a shows the distribution of speakers across languages. Each speaker declared some sensitive attributes, i.e., his/her accent, age, and gender. Since these attributes are not declared by all the speakers, we filtered out the speakers who have not provided any sensitive attribute.

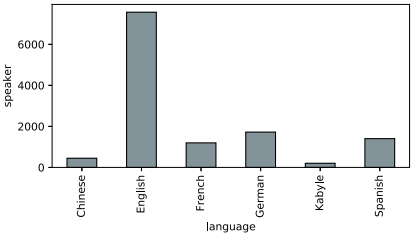
In numbers, we manipulated $1,046,078$ utterances coming from $12,057$ speakers. On average, each speaker contributed with around 80 utterances. Figure 1b plots the distribution of utterances per speaker. Furthermore, it can be observed that, while we did our best to keep the dataset balanced in terms of demographic groups, the dataset is still unbalanced with respect to the sensitive attributes (i.e., age and gender) and the corresponding demographic groups. The gender attribute[4] is labelled by the platform with *male* or *female*, while the age attribute is a categorical label identifying the age range, with a step of 10 years (e.g., teens, twenties). Figure 1c and d provide the individual distribution across gender and age. Please note that, while training and testing, we ensured that the representation of individuals with respect to their gender was balanced, and we kept comparable the individual distribution across ages for each gender-based group. This can represent a first pre-processing countermeasure against unfairness in speaker verification, as part of the contributions of this study.

3.3 Methodological Protocol

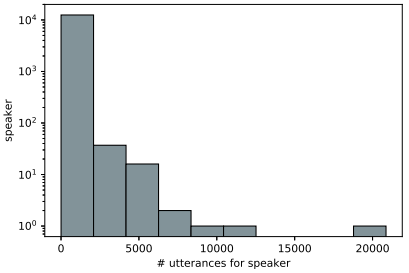
In this section, we detail the protocol we followed to perform the analysis presented in this paper, going from data collection to training and test procedures.

[3] https://voice.mozilla.org/it/datasets.

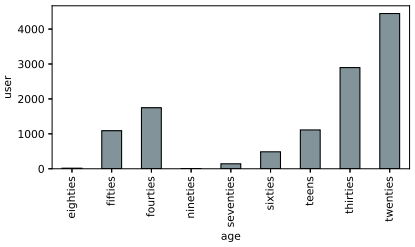
[4] While the gender is by no means a binary construct, to the best of our knowledge no dataset for speaker recognition with non-binary genders exists. What we are considering is a binary feature, as the current publicly available datasets offer.



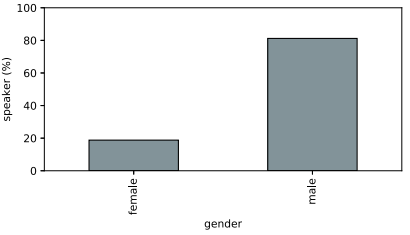
Speaker distribution across languages.



Utterance distribution across speakers.



Age distribution across speakers.



Gender distribution across speakers.

Fig. 1. Data Statistics. Representative distributions of individuals and utterances along languages, ages, and genders in the data leveraged by this study.

Utterances Download. The first step served to download the utterances from the Common Voice website. More precisely, we downloaded data coming from Chinese, French, German, English and Kabyle languages. To this end, for each language, we selected the corresponding language label from the dropdown list of languages, and we clicked on the download button to start collecting data.

Data Cleaning and Conversion. The second step implied to remove, from the datasets we downloaded, all the speakers who do not declare any sensitive attribute. Moreover, since all the utterances were originally coded in *MP3*, we leveraged FFMpeg[5] to convert each utterance into a *WAV* file. This would allow us to easily manipulate utterances with existing Python audio libraries. During this conversion process, all the utterances identified as corrupted were removed.

Data Organization. Then, we organized our data in five different folders, each corresponding to a certain language among Chinese, French, German, English, and Kabyle. Then, within each language folder, we created a folder for each user speaking that language, and stored his/her utterances inside. User's folders were identified with the format *id00000*, *id00001*, and so on. To take note of the demographic groups, a file lists language, id, gender, and age for each user.

[5] https://www.ffmpeg.org/.

Based on gender, we grouped users in male users (G_m) and female users (G_f). In the context of this paper, we also created two age-based user's groups, balancing their representation in the dataset. The first one included users who are less than 40 years old (A_y). The remaining users were included in the second group (A_o).

Train-Test Split. This step was dedicated to identify users and corresponding utterances to be used for training and testing speaker verification models. The goal was to create training and testing sets balanced in terms of gender and age representation, so that we could then test whether the model emphasized unfairness across demographic groups, even though it received properly-balanced training data. To this end, for training a model for a certain language, we identified the less represented gender, and then we randomly selected users from the other gender in order to meet the following properties: (i) the number of male and female users in the sampled data is equal, and (ii) the number of individuals per age range for each gender is equal. Chinese and Kabyle were excluded from the experiments since they had no enough users for properly training and testing the corresponding speaker verification model. To ensure reproducibility, we stored the list of sampled users into a training file for each considered language.

Regarding the trial testing verification pairs, for each language, we considered 40 speakers, including 10 users in $G_f \cap A_y$, 10 users in $G_f \cap A_o$, 10 users in $G_m \cap A_y$, and 10 users in $G_m \cap A_o$. For each speaker, we randomly chose his/her 10 utterances and, for each sampled utterance, we compared it with 4 randomly-sampled utterances of the same speaker (i.e., to create genuine comparisons) and 4 randomly-sampled utterances from other speakers (i.e., to create impostor comparisons). For impostor comparisons, we selected the probe speaker, so that we were able to test impostor matches with an individual with the same gender and age range, same age range but different gender, same gender but different age range, and different gender and age range of the current speaker. To ensure reproducibility, for each language, we stored the trial pairs into a test file.

Model Train. In this step, we dealt with the procedure followed for training the considered speaker verification models. More precisely, we trained a different instance of the deep speaker verification architecture for each considered language (i.e., one for English, one for Spanish, one for French, and one for German). Given the training file created for a certain language in the previous step and the set of training parameters (e.g., batch size, learning rate, and so on), we first loaded the file paths corresponding to all the utterances belonging to the users listed in the training file, and we assigned a label to each utterance based on the speaker who produced it. Then, we set up a batch generator which returned (spectrogram, label) pairs to be fed into the model while training. The speaker verification architecture was instantiated, and compiled according to the training parameters passed to the script. Finally, the instantiated architecture was trained with data returned by the batch generator. Implementation

details about the architecture and the training procedure are provided in the next sections[6].

Model Test. Finally, we followed a detailed procedure in order to test a pre-trained speaker verification model. More precisely, we first loaded the pre-trained model, and we removed the top layers used for aggregation and speaker classification while training, so that the top layer of the model was the fully-connected layer from which the speaker embeddings were extracted. Then, we loaded the list of trial verification pairs created in the previous steps for a given language. For each trial pair, we loaded the corresponding two audios, we computed their acoustic representations (e.g., spectrograms), and we fed them into the pre-trained model in order to extract the corresponding speaker embeddings. Then, we compared the two speaker embeddings by measuring their cosine similarity, and we saved such a value for each pair in a resulting CSV file. Given the sensitive attribute labels associated to users in the trial pairs, the related matching scores (i.e., cosine similarities), and the ground-truth labels (i.e., 1 if the two audios of a pair come from the same speaker, 0 otherwise), we computed the false acceptance rates and the false rejection rates for different demographic groups.

3.4 Experimented Speaker Verification Architecture

In this section, we describe the studied deep speaker verification architecture, namely Thin-ResNet, considered the state-of-the-art implementation for speaker verification in the wild [25]. This kind of architecture made it possible to leverage a convolutional-based architecture trainable end-to-end for the task of speaker recognition. The architecture expected to receive magnitude spectrograms extracted directly from raw audio waveforms, with no other pre-processing. Then, the deep neural network was used to extract utterance-level speaker embeddings. We treated spectrograms as single-channel images, and exploited the fact that deep networks can learn frequency-based filters, if needed for the speaker recognition task (e.g., filters can detect patterns in low-frequency regions).

Thin-ResNet is obtained from the ResNet-34 [18] architecture, known for high efficiency and good classification performance on image data. Residual-network (ResNet) architectures are based on standard multi-layer convolutional neural networks, but with added skip connections such that the layers add residuals to an identity mapping on the channel outputs. Original layers were modified to adapt to the spectrogram input, and we applied batch normalization before computing Rectified Linear Unit (ReLU) activations. Moreover, the fully-connected layer from the original ResNet-34 can be replaced by one of the following aggregation strategies, namely *NetVLAD*, *GhostVLAD*, or *Average Pooling Aggregation*.

[6] In the context of our work, where we are more interested in understanding algorithm characteristics beyond overall accuracy, the small further accuracy improvements that can probably be achieved through intensive hyper-parameter tuning would not substantially affect the main outcomes of our analyses.

For the latter, it was needed to replace the first fully-connected layer with two layers: a fully-connected layer of 9×1 (working on the frequency domain), and an aggregation layer (e.g., average pooling) working on $1 \times n$, where n depends on the length of the input audio (i.e., $n = 4$ for a 2-s segment). In this way, the network becomes invariant to temporal patterns but not to frequency patterns, that are the main source of information in speech, and made it possible to reduce network parameters.

3.5 Implementation Details

The different Thin-ResNet instances were implemented in Keras, with a Tensorflow backend. Both models and training code are made publicly available. During training, we randomly sampled segments from each utterance. We used 512-point (Fast Fourier Transforms) FFTs giving us spectograms of size 257×200 (frequency x temporal). Each model was trained using a fixed-size spectrogram corresponding to a 2-s interval. If the utterance was marked as shorter, we padded it with zeros to reach the targeted 2-s length. All audio waveforms were converted to single-channel, 16-bit streams at a 16 kHz sampling rate for consistency. Spectrograms were then generated in a sliding window fashion using a hamming window of width 25 ms and step 10 ms. Each spectrogram was normalized by subtracting the mean and dividing by the standard deviation of all frequency components in a single time step. No voice activity detection or automatic silence removal was applied. Each model was trained for speaker classification using standard *Softmax*, and served with batches of size 32. We used the Adam optimizer, with an initial learning rate of 0.001, and decreased the learning rate by a factor of 10 after every 10 epochs, until 30 epochs. On top of each Thin-ResNet instance, we used a *GhostVLAD* aggregation layer [35], with 10 clusters plus 2 ghost clusters. For testing, we considered speaker embeddings of size 512, extracted from the second fully-connected layer *fc2* of each Thin-ResNet. The implementation was built on top of the work proposed in [25].

4 Results and Discussion

In this section, we empirically evaluate the considered speaker verification models with respect to their performance on different demographic groups. We aim to answer three key research questions:

1. Are equal error rates influenced by the individual demographic membership?
2. How does the demographic group impact on verification score distributions?
3. Do models fail uniformly across users or concentrate errors on specific users?

4.1 RQ1: Effect on Equal Error Rates

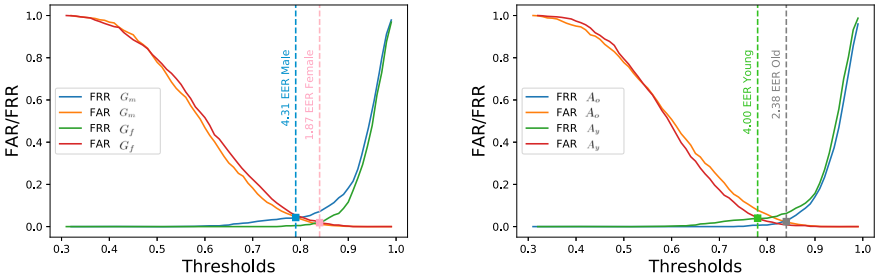
In this subsection, we run experiments on our data to assess whether the considered speaker verification models treated demographic groups differently. We

assumed that there was a corresponding demographic group for any combination of language (i.e., English, Spanish, French, German), gender (i.e., G_m, G_f), and age (i.e., A_y, A_o). To this end, we trained four speaker verification models, one model per considered language. Then, we measured the Equal Error Rates (EERs) achieved on the trial pairs defined in Sect. 3.3 for different demographic groups, depending on language, gender, and age. Please note that the EER refers to the configuration at which the false acceptance rate and the false rejection rate are equal to each other. The lower the better.

Table 1 showed the EERs obtained for each language, grouping individuals per gender and age. It can be observed that the speaker verification architecture introduced significant differences in performance among demographic groups, even though each demographic group was equally represented in the training data. There was no clear pattern that emerged across languages, but we observed an undesired behaviour varying in scale and attributes across languages. To have a more detailed picture, the speaker verification model trained on English utterances generally performed better for females than males. More precisely, we measured a gap of 4.25%-points between performance on users in G_f (no matter the age) and on users in $G_m \cap A_y$. The gap was reduced up to 0.75%-points when we compared users in G_f with users in $G_m \cap A_o$. Considering the speaker verification model trained on Spanish utterances, it can be observed that no discrimination across gender is introduced when we considered individuals in A_o (i.e., 7.38% against 7.37%). Conversely, statistically-significant differences were introduced between users in $G_m \cap A_y$ and in $G_f \cap A_y$, with a gap of 3.75%-point in favor of the first group. This tendency of the model to discriminate individuals in A_y across genders emerged on both English and Spanish models at different extent. Furthermore, the speaker verification model trained on French utterances showed slightly different performance. Overall, this model tended to perform better for males than females, and there was a clear intra-gender gap of around 0.80%-points for both males and females. The French speaker verification model confirmed also that there was a higher difference among individuals in A_y across genders (i.e., 3.75%-points), while the difference for individuals in A_o across gender was smaller (i.e., 1.75%-points). Finally, the speaker verification

Table 1. EER per demographic group. For each language, we conducted a statistical analysis on verification scores based on a Paired Student's t-test, at 5% level. Statistically-significant differences were detected in all cases, except for ($G_f \cap A_y$ and $G_f \cap A_o$) in English and ($G_f \cap A_o$ and $G_m \cap A_o$) in Spanish.

Gender group	Age group	Language			
		English	Spanish	French	German
G_f	A_y	2.00	5.50	8.38	10.25
G_f	A_o	2.00	7.38	7.13	4.75
G_m	A_y	6.25	1.75	4.63	7.50
G_m	A_o	2.75	7.37	5.38	2.00



Errors across genders. Errors across age ranges.

Fig. 2. Error Statistics. False Acceptance Rates (FARs) and False Rejection Rates (FRRs) across gender- and age-based demographics groups. (Color figure online)

model trained on German utterances exhibited more noisy patterns across genders and ages. Fixed the age, the model performed better for males than females, introducing a difference of around 2.75%-points across groups. When we fixed the gender, the best performance were measured for individuals in $G_f \cap A_o$ and in $G_m \cap A_o$.

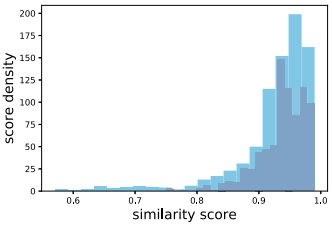
In the context of our study, it can be concluded that the considered models tended to perform better for male individuals than female individuals, when the age range was the same. Exception was made for the English model. Similarly, when the gender was the same, we observed a tendency of the speaker verification model to foster better performance for individuals in A_o than individuals in A_y.

4.2 RQ2: Influence on Verification Score Distributions

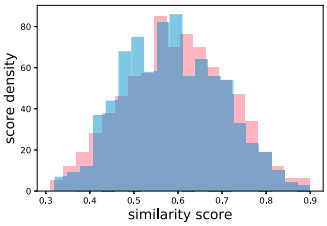
We next compare the considered speaker verification models trained on the four different languages to assess how they perform in terms of false acceptance rates and false rejection rates in comparison to each other. The goal here is to understand whether the speaker verification models introduced differences among demographic groups in terms of security (i.e., depending on false accepts) or usability (i.e., depending on false rejections). For conciseness, we report results on the English model, but the results on other languages showed similar patterns.

Figure 2 depicts the false acceptance rates and false rejection rates at various thresholds for individuals in G_m and G_f (left) and individuals in A_y and A_o (right)[7]. From Fig. 2a, it can be observed a difference on false accepts and false rejections across genders. More precisely, individuals in G_f (red line) tended to experience higher false accepts than individuals in G_m (orange line). Conversely, individuals in G_m (blue line) were more exposed to false rejects than individuals in G_f (green line). It follows that a common setup that implies a unique verification threshold across users would result in less security for female users and less usability for male users. It should be noted also that it would not

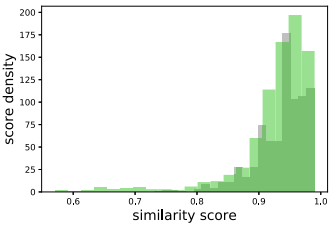
[7] Please note that the figures in this manuscript are best seen in color.



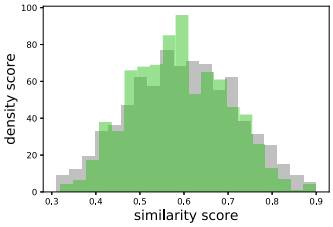
Genuine scores across genders.



Impostor scores across genders.



Genuine scores across age ranges.

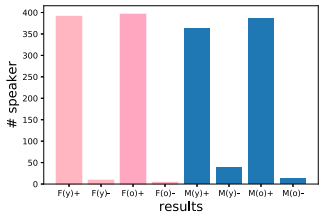


Impostor scores across age ranges.

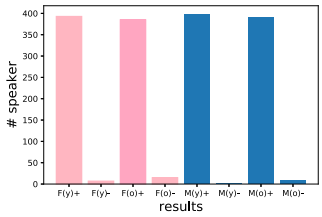
Fig. 3. Matching Score Distributions. Matching score distributions for genuine and impostor pairs between G_m (blue) and G_f (pink) individuals on the top, and between A_y (grey) and A_o (green) individuals on the bottom. (Color figure online)

be acceptable to setup different thresholds based on the gender due to privacy constraints. Moreover, it would not be scalable, since an increasing number of thresholds would be needed as more and more demographic groups are taken into account. Considering demographic groups based on the age range (Fig. 2b), similar observations could be made for individuals in A_y and A_o. Surprisingly, individuals in A_o (orange line) suffered from higher false accepts than individuals in A_y (red line). To the same extent, individuals in A_y (green line) got more false rejects than individuals in A_o (blue line). This resulted in less security for individuals in A_o and less usability for individuals in A_y.

To have a more detailed picture, we resorted our analysis to the distribution of matching scores for both genuine and impostor trial pairs. In Fig. 3, we reported such distributions along gender- and age-based demographic groups. Figure 3a highlighted that genuine matching scores over male individuals tended to fall into higher similarity values than those for female individuals. Conversely, Fig. 3b showed a more equal distribution of impostor matching scores along genders. It can be just observed that female individuals tended to have higher impostor matching scores, leading to more false accepts (i.e., less security). Similar patterns appeared in Fig. 3c and d across age ranges. Our analysis thus uncovered the need of creating speaker verification models that equally distribute matching scores, so that the same distribution pattern is measured on different demographic groups. We argue that this point is crucial to achieve higher fairness, since the verification threshold is unique among groups.



Genuine matching performance.



Impostor matching performance.

Fig. 4. Verification Errors. The verification performance in terms correct and incorrect matches across gender- and age-based demographic groups.

To provide more evidence on the above point, we considered a unique verification threshold computed on the user population as a whole (i.e., this practice is common on most of the biometric studies), and we counted in Fig. 4 the number of genuine and impostor matches that were correctly or incorrectly classified for each demographic group. Each demographic group result was identified by a gender label (F: G_f, M: G_m), an age-range label (y: A_y, o: A_o), and a classification result (+:correct, −:incorrect). Hence, for instance, the label "*F(y)+*" identified the number of matches correctly classified for individuals in $G_f \cap A_y$. From Fig. 4a, it can be observed the number of genuine matches correctly (true accepts) and incorrectly (false rejects) classified. Overall, there is a clear trend showing less rejection errors for female individuals (i.e., *F(y)−* and *F(o)−*) than male individuals (i.e., *M(y)−* and *M(o)−*). Regarding age ranges, individuals in A_y (i.e., *F(y)−* and *M(y)−*) experienced more rejection errors than individuals in A_o (i.e., *M(o)−* and *F(o)−*). Similar patterns were observed when we considered impostor matching performance. Figure 4b reported the number of impostor matches correctly (true reject) and incorrectly (false accepts) classified. These results confirmed that individuals in G_f and in A_o suffer from more false accepts.

Overall, it can be concluded that both genuine and impostor verification scores were strongly influenced by the demographic group, and the models tended to give less security or less usability to certain demographic groups.

4.3 RQ3: Errors per User Concentration

Finally, we investigated how speaker verification models performed on each of the considered users. To this end, we measured the total number of false accepts and false rejects for each user, and we reported the number of errors per user into plots to assess whether the speaker verification models made errors uniformly across users or concentrated errors to specific users and/or demographic groups.

Figure 5 showed the number of verification errors per user, with users sorted by decreasing number of errors. Considering the gender-based membership, Fig. 5a and b highlighted the fact that the errors were well-distributed within male users. Exception was made for a male user who resulted in 26 verification errors.

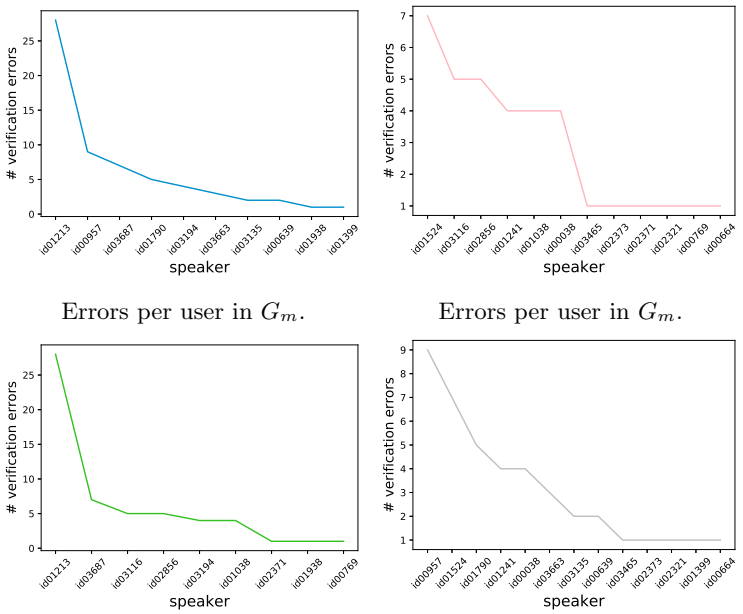


Fig. 5. Errors per User. Representative distribution of errors per user made by the speaker verification model, grouped by demographic group membership.

We further checked this result, and we found out that such a user recorded his audios in very challenging conditions, making it difficult the verification process, independently from the demographic group. This observation points to the fact that speaker verification is influenced by various co-variants (e.g.., spoken words, utterance length, noise) that, by extension, may complicate isolating the differences across demographic groups that derive from the user's membership to that group. Conversely, around half of the female users collected most of the errors among users from the same gender. Similar patterns were also observed when we considered the age-based membership in Fig. 5c and d. Overall, the number of errors per user was satisfactory, with less than 10% of errors per user.

In the context of our study, models tended to make errors uniformly across users in case of G_m and A_y. However, errors were more frequent for a subset of individuals in $G_f \cap A_y$, uncovering the need of fine-grained individual analysis.

5 Conclusions and Future Work

In this paper, we proposed a framework for analyzing deep speaker verification models in terms of their predictive accuracy, false positive rate, and false negative rate across demographic groups. Through a series of experiments, we showed

that, despite a balanced representation of the demographic groups in the underlying training data, the models can be quite different in terms of performance based on the demographic group, exhibiting undesired consequences. Our work provides a better understanding on how generalizable speaker verification models are to diverse demographic groups (i.e., language, gender, age), and fosters more fairness-oriented evaluations in speaker verification research.

In next steps, we plan to investigate bias and fairness on other models, beyond *ThinResNet*, and trace these biases back to some particular patterns of training data and acoustic representations. Moreover, we will consider other types of bias, and we will design proper countermeasures to the biases we uncovered.

References

1. Global voice recognition market 2017–2021 (2019). https://www.reportlinker.com/p04338419/Global-Voice-Recognition-Biometrics-Market.html. Accessed 30 Sept 2019
2. Alasadi, J., Al Hilli, A., Singh, V.K.: Toward fairness in face matching algorithms. In: Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia, pp. 19–25 (2019)
3. Anzalone, L., Barra, P., Barra, S., Narducci, F., Nappi, M.: Transfer learning for facial attributes prediction and clustering. In: Wang, G., El Saddik, A., Lai, X., Martinez Perez, G., Choo, K.-K.R. (eds.) iSCI 2019. CCIS, vol. 1122, pp. 105–117. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1301-5_9
4. Barocas, S., Hardt, M., Narayanan, A.: Fairness in machine learning. In: NIPS (2017)
5. Barra, P., Bisogni, C., Nappi, M., Freire-Obregón, D., Castrillón-Santana, M.: Gender classification on 2D human skeleton. In: 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART), pp. 1–4. IEEE (2019)
6. Boratto, L., Carta, S.: Modeling the preferences of a group of users detected by clustering: a group recommendation case-study. In: 4th International Conference on Web Intelligence, Mining and Semantics, WIMS, pp. 16:1–16:7. ACM (2014)
7. Boratto, L., Manca, M., Lugano, G., Gogola, M.: Characterizing user behavior in journey planning. Computing **102**(5), 1245–1258 (2020). https://doi.org/10.1007/s00607-019-00775-8
8. Chen, Y.h., Lopez-Moreno, I., Sainath, T.N., Visontai, M., Alvarez, R., Parada, C.: Locally-connected and convolutional neural networks for small footprint speaker recognition. In: Proceedings Interspeech 2015, pp. 1136–1140 (2015)
9. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: deep speaker recognition. In: Proceedings Interspeech 2018, pp. 1086–1090 (2018)
10. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
11. Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: a survey on an emerging challenge. arXiv:2003.02488 (2020)
12. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012)
13. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Demographic bias in presentation attack detection of iris recognition systems. arXiv:2003.03151 (2020)

14. Garcia, R.V., Wandzik, L., Grabner, L., Krueger, J.: The harms of demographic bias in deep face recognition research. In: 2019 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2019)
15. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Mag. **38**(3), 50–57 (2017)
16. Hansen, J.H., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process. Mag. **32**(6), 74–99 (2015)
17. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp. 3315–3323 (2016)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
19. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5115–5119. IEEE (2016)
20. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135. IEEE (2017)
21. Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: I-vector based speaker recognition on short utterances. In: Proceedings Interspeech 2011, pp. 2341–2344 (2011)
22. Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T.: Speaker identification and clustering using convolutional neural networks. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE (2016)
23. Mahfouz, A., Mahmoud, T.M., Eldin, A.S.: A survey on behavioral biometric authentication on smartphones. J. Inform. Secur. Appl. **37**, 28–37 (2017)
24. Marras, M., Korus, P., Memon, N., Fenu, G.: Adversarial optimization for dictionary attacks on speaker verification. In: Proceedings Interspeech 2019, pp. 2913–2917 (2019)
25. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: VoxCeleb: large-scale speaker verification in the wild. Comput. Speech Lang. **60**, 101027 (2020)
26. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: Proceedings Interspeech 2017, pp. 2616–2620 (2017)
27. Ramos, G., Boratto, L.: Reputation (in)dependence in ranking systems: demographics influence over output disparities. CoRR abs/2005.12371 (2020)
28. Reforgiato Recupero, D., Dessì, D., Concas, E.: A flexible and scalable architecture for human-robot interaction. In: Chatzigiannakis, I., De Ruyter, B., Mavrommati, I. (eds.) AmI 2019. LNCS, vol. 11912, pp. 311–317. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34255-5_21
29. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Process **10**(1–3), 19–41 (2000)
30. Selbst, A.D.: Disparate impact in big data policing. Ga. L. Rev. **52**, 109 (2017)
31. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5329–5333. IEEE (2018)
32. Terhörst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Post-comparison mitigation of demographic bias in face recognition using fair score normalization. arXiv preprint arXiv:2002.03592 (2020)
33. Tolan, S.: Fair and unbiased algorithmic decision making: current state and future challenges. arXiv preprint arXiv:1901.04730 (2019)

34. Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: International Conference on Acoustics, Speech and Signal Processing, pp. 4052–4056. IEEE (2014)
35. Zhong, Y., Arandjelović, R., Zisserman, A.: GhostVLAD for set-based face recognition. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11362, pp. 35–50. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20890-5_3