Pattern Recognition Letters 168 (2023) 131–137

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Causal reasoning for algorithmic fairness in voice controlled cyber-physical systems

Gianni Fenu, Mirko Marras*, Giacomo Medda, Giacomo Meloni

Department of Mathematics and Computer Science, University of Cagliari, Cagliari 09124, Italy

ARTICLE INFO

Article history:
Received 19 June 2022
Revised 6 February 2023
Accepted 12 March 2023
Available online 13 March 2023

Edited by: Maria De Marsico

Keywords:
Security
Authentication
Voice biometrics
Fairness
Speaker recognition

ABSTRACT

Automated speaker recognition is enabling personalized interactions with the voice-based interfaces and assistants part of the modern cyber-physical-social systems. Prior studies have unfortunately uncovered disparate impacts across demographic groups on the outcomes of speaker recognition systems and consequently proposed a range of countermeasures. Understanding why a speaker recognition system may lead to this disparate performance for different (groups of) individuals, going beyond mere data imbalance reasons and black-box countermeasures, is an essential yet under-explored perspective. In this paper, we propose an explanatory framework that aims to provide a better understanding of how speaker recognition models perform as the underlying voice characteristics on which they are tested change. With our framework, we evaluate two state-of-the-art speaker recognition models, comparing their fairness in terms of security, through a systematic analysis of the impact of more than twenty voice characteristics. Our findings include important takeaways to enable voice controlled cyber-physical-social systems for everyone. Source code and data are available at https://bit.ly/EA-PRLETTERS.

1. Introduction

Characterized by the mass integration of smart devices in our daily life, Cyber-Physical-Social Systems (CPSSs) are receiving an increasing attention [1]. This paradigm mainly promotes the integration of traditional cyber-physical systems with human and social factors that emerge during the system operation and management. A notable challenge in this domain is to ensure a seamless yet secure interaction with smart devices while respecting important human values. Prior work has identified personalization as one viable solution to address this challenge [2]. This assumption follows from the fact that one of the main contributors to the complexity of CPSS originates from human dynamics, which are guided by environmental and personal factors usually hard to predict. Personalization has been therefore suggested to better adapt CPSSs to individuals.

Intelligent voice-based interfaces and assistants are being leveraged for personalization into the existing CPSSs, from healthcare treatment [3] to entertainment services [4,5]. For instance, notable assistants, such as Amazon Alexa and Google Home, detect the active speaker based on the voiceprints (i.e., a unique voice signature) and provide personalized responses. Personalizing the handling of voice queries amid social voice environments is a key feature of next-generation CPSSs, considering that voice-based interfaces are becoming commonplace in an employee facing capacity, workspaces etc. [6].

Speaker recognition is a driver for personalization in voice-based interfaces and assistants, being used to confirm or refute the user's identity based on their voiceprints. With the wider availability of speech data and increasingly efficient computing resources, speaker recognition systems have achieved high accuracy by leveraging acoustic representations extracted from deep neural networks [7,8]. For certain demographic groups, however, these systems might under-perform due to differences across dialects (e.g., because of regional accent), intra-group heterogeneity (e.g., age, gender, or ethnicity), or speech pattern variability of each individual in the group (e.g., people with disabilities). As for now, research in speaker recognition mostly focused on quantifying differences of error rates across groups [9,10]. Potential sources of unfairness, beyond the under-representation of a demographic group [11], have been rarely investigated. Understanding *why* a speaker recognition system may lead to disparate performance for different (groups of) users is however essential to enable CPSSs for everyone.

Grounded on explanatory machine learning [12,13], our goal in this paper is to uncover the influence of voice characteristics on the disparate error rates emphasized by speaker recognition models. Our research towards addressing this goal is relevant for all the applications in which automated speaker recognition can

* Corresponding author.
E-mail addresses: fenu@unica.it (G. Fenu), mirko.marras@unica.it (M. Marras), giacomo.medda@unica.it (G. Medda), g.meloni31@studenti.unica.it (G. Meloni).

https://doi.org/10.1016/j.patrec.2023.03.014

be beneficial. For instance, uncovering the vocal characteristics of a certain (group of) speaker(s) that tends to experience higher false acceptance rates facilitates both humans' understanding about the causes of the disparate errors and judging about whether the disparate error rate is based on a learned demographic bias. Model interpretations about the most important characteristics that drive disparate error rates can allow to better think about and manage edge cases where those voice characteristics tend to appear more prominently. Furthermore, devising methods able to mitigate the effects of those voice characteristics can increase equity and social acceptance of speaker recognition models and, consequently, foster the responsible adoption of such technology in the real world.

With the above goal in mind, in this paper, we propose an experimental framework aimed at understanding how changing the underlying voice characteristics (either protected and non-protected) affects the speaker recognition models performance. Our framework is based on a statistical model that involves over twenty voice characteristics and studies their impact on the performance of two state-of-the-art speaker recognition models. This impact is referred to as *explanatory power*, as it reflects the capability of a voice characteristic to influence a given dependent variable (either security or usability). We then leverage our framework to address the following research questions: do sensitive attributes have any relationship with other speech covariates (RQ1), which speech covariates influence performance the most (RQ2), and to what extent the sensitive class affects performance (RQ3). To answer these research questions, the main novel contribution of this paper is threefold:

- We design a novel model-based explanatory framework to analyze the impact of voice characteristics on fairness estimates experienced by speaker recognition models.
- We evaluate two state-of-the-art speaker recognition models, comparing their fairness in terms of security on a large dataset, through our explanatory framework.
- Through a systematic analysis, we provide key observations and recommendations on the impact of voice characteristics on the performance of speaker recognition models.

Our findings highlight the interplay between (i) the disparate degrees of security that are propagated by speaker recognition systems across demographic groups and (ii) the characteristics of the vocal utterances that characterize each group. The consequent insights emphasize the need of mitigating the effects of the identified voice characteristics that tend to drive unfairness, to ensure the reliable and ethical deployment of this technology.

2. Related work

Speaker modeling. Speaker recognition is implemented via two main tasks: *identification* aims to detect the speaker's identity within a gallery of candidate speakers; *verification* aims to confirm the identity of the claimed speaker and operates in an open-set regime based on a gallery of enrolled speech samples. Speaker modeling has been recently dominated by deep neural networks [8] (DNNs) which significantly outperform classic solutions like GMM-UBM [14] or I-Vectors [15]. DNNs are typically pre-trained for the identification task, but are then adapted to open-set verification by discarding the classification head and extracting an intermediate representation, referred to as a *speaker embedding*. The embeddings of the query and enrolled samples are compared to confirm the speaker's identity.

Countless deep neural architectures have been proposed for speaker modeling. Some of the most prominent differences among the existing architectures involve the input acoustic representation, the backbone network, and the temporal pooling strategy. Directly using waveforms to learn a representation is possible [16], but it is much more common to use a hand-crafted 2D representation (e.g., spectrograms or filterbanks). The latter enables the adaptation of successful backbones from computer vision, e.g., VGG (Visual Geometry Group) [17] or ResNet (residual networks) [18,19]. Recurrent [20], pooling [19], or time delay neural networks [7] can be then used to deal with the time dimension typical of the vocal input.

Usually, trainable pooling layers achieve better results than simple pooling operators, (e.g., average pooling [19] or statistical pooling [7]). Some of the most successful learned designs include the family of VLAD (Vector of Locally Aggregated Descriptor) models. NetVLAD [21] assigns each frame-level descriptor to a cluster and computes residuals to encode the output features. Its variant GhostVLAD [21] excludes some of the original NetVLAD clusters from the final concatenation, such that undesirable speech sections are down-weighted.

Fairness in speaker recognition. Concerns that machine-learning models may discriminate against certain groups are commonplace. Several studies indeed provided evidence and approaches against biases, e.g., Mehrabi et al. [22], Goodman and Flaxman [23], Terhörst et al. [24], Nápoles and Koutsoviti Koumeri [25].

Unfairness issues have been just recently uncovered in the speaker recognition domain. Prior works measured and analyzed algorithmic unfairness on sensitive demographic groups (e.g., gender, age, nationality, accents) in terms of disparate error rates. The lack of training data representing the minority demographic groups was identified as the main reason behind such disparities: data balancing and pre-training strategies across groups were proposed as countermeasures in Fenu et al. [9,11], Zhang et al. [26], Meng et al. [27]. Subsequent progress in this field included unfairness treatments based on group-adapted encoders [28] or adversarial and multi-task learning techniques [29]. Evaluation frameworks aimed to investigate performance disparities across different demographic subgroups represent another line of research in voice only [30] and audio-visual biometrics [31].

None of the above studies questioned the origin of the disparities, beyond data imbalance. Our work is hence different from the literature, since it proposes an explanation framework for uncovering the influence of voice characteristics on disparities in error rates across demographic groups via causal reasoning [32–35]. It therefore allows us to *interpret* and *explain* which voice characteristics might lead the model to be *unfair* (why it is unfair), going beyond just claiming that a model is unfair. Our contribution should be seen as an additional layer towards creating responsible speaker recognition models, on top of accurate model architectures and assessments of performance disparities. In view of this, it would not be possible for us to make any comparison with other methods, since our work touches on unfairness from a different perspective and, hence, targets a different goal than the existing studies.

3. Methodology

Our methodology is based on two main phases, summarized in Fig. 1. We first created a range of speaker recognition models resulting in high overall accuracy performance. Then, we leveraged the pre-trained models to run an exploratory analysis on the impact of voice characteristics on performance.

3.1. Speaker recognition framework

In this section, we show our experimental setup and the details of the used data sets, speaker encoders, and so on.

Recognition task formulation. First, we mathematically define the addressed classification task. Formally, let $w \in [-1, 1]^*$ denote a *speech waveform* of variable length. We consider a feature extraction step, using a *speaker encoder* ($\mathcal{E}$), that produces a fixed-length

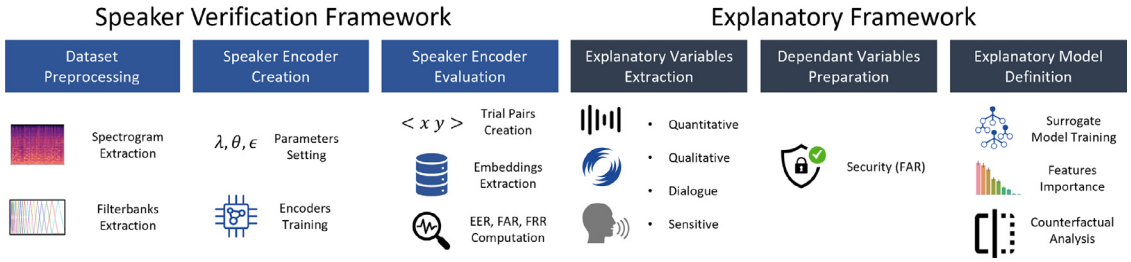

Fig. 1. Once the data set is pre-processed, we created two speaker encoders and ran an explanatory analysis to assess the impact of voice characteristics.

speaker embedding $\mathcal{E}(w) \in \mathbb{R}^e$, $e \in \mathbb{N}$. Given a *decision threshold* τ and a similarity function $\mathcal{S}$, a verification trial v can be defined as:

$$v_\tau(w_u, w_p) : \mathcal{S}(\mathcal{E}(w_u), \mathcal{E}(w_p)) > \tau \tag{1}$$

where, under the speaker encoder $\mathcal{E}$, an input speech waveform w_p from an unknown user p is compared with a speech waveform w_u from user u to confirm or refute the speaker's identity (1 and 0, respectively). Training a speaker recognition model then becomes an optimization problem aimed at maximizing the expectation on the following objective function:

$$\underset{(\mathcal{E},\tau)}{\text{argmax}} \underset{u,p}{\mathbb{E}} \begin{cases} v_\tau(w_p, w_u) & p = u \\ 1 - v_\tau(w_p, w_u) & p \neq u \end{cases} \tag{2}$$

Dataset preprocessing. In this study, we used a public dataset, namely `FairVoice` [10], since its speech waveforms are appropriately annotated with sensitive attributes and widely adopted for fair speaker recognition benchmarking [9–11]. This dataset includes speakers' data collected from Common Voice, one of the largest corpora including unconstrained speech extracted from real-world scenarios, and featuring diverse acoustic environments. All the waveforms were single-channel, 16-bit recordings sampled at 16 kHz. To the best of our knowledge, except for Common Voice, there is no other large public dataset which comprises voice data coming from a range of languages and labeled with a variety of sensitive attributes. Only the English and Spanish utterance sets contain enough samples to ensure statistical significance and were therefore considered in this study. The gender attribute was binary (male or female) and self-reported by the users on the platform. The age attribute was binarized in younger (age $\leq$ 40) and older (age $>$ 40) users.

Speakers who provided less than six utterances were discarded, leading to a total of 6321 English speakers and 1298 Spanish speakers. The dataset was then split into training and testing sets, according to the standard protocol proposed in Fenu et al. [11]. Specifically, we considered the balanced multi-language training set setting, where the speakers' representation was balanced by gender and age for both languages in the same training set, i.e., 155 speakers for each of the eight demographic groups obtained by combining gender, age, and language. For each speaker in the testing set, we generated 55 trial verification pairs: 5 positives (other utterances from the same speaker) and 50 negatives (other utterances from another speaker).

Speaker encoder creation. We used speaker encoders based on various Convolutional Neural Network (CNN) backbones and acoustic representations, including a ResNet model (ResNet-34 [17]) trained on spectrograms and a X-Vector model based on filter banks [36]. The ResNet model was adapted from computer vision to spectrogram inputs by replacing the last fully-connected (FC) layer with two layers: an FC one with support in the frequency domain and average pooling with support in the time domain. On the other hand, X-Vector includes five layers that operate on speech frames, with a time context centered at the current frame. A pooling layer aggregates frame-level outputs and computes mean and standard deviation. Two FC layers aggregate statistics across the time dimension. We used a GhostVLAD pooling [21].

Table 1
Performance of the considered speaker encoders under negative pairs created with another user from the same age range or the same gender (more challenging scenario).

	ResNet-34		X-Vector	
Negative pair type	*EER*	$FRR_{FAR1\%}$	*EER*	$FRR_{FAR1\%}$
Same age range	0.08	0.27	0.08	0.2
Same gender	0.11	0.43	0.11	0.3

We trained our speaker encoders from scratch using data from the training set. We randomly sampled segments from each utterance and standardized the inputs to 2-s clips (by cropping or padding). No voice activity detection or silence removal was applied. Spectrograms (filter banks) were generated in a sliding window fashion using a Hamming window of width 25 ms and step 10ms. We used 512-point (Fast Fourier Transforms) FFTs yielding spectrograms of size 257×200 and filter banks of size 24 × 300 (frequency×temporal). Each acoustic representation was normalized by subtracting the mean and dividing by the standard deviation of all frequency components in a single time step. The model was trained for classification using Softmax, and served with 32-sized batches. We used the Adam optimizer, with an initial learning rate of 0.001, decreased by a factor of 10 every 10 epochs, until convergence.

Speaker encoder evaluation. Subsequently, the pre-trained speaker encoders were deployed in a speaker verification system by stripping their classification heads. We then performed a detailed assessment of open-set speaker verification performance on the verification trail pairs created from the testing set. With speech waveforms of a test set, we generated a set of verification trial pairs (5 positives, 50 negatives), with individuals from the same language group. We tested the discriminability of 512-sized speaker embeddings on the obtained test pairs. Based on the collected cosine similarity scores, we found the ROC (Receiver Operating Characteristic) curve and derived standard metrics, including the equal error rate (EER). The resulting evaluations were used for the threshold calibration. Due to space constraints, we focused on the threshold corresponding to the EER. Evaluation metrics were computed for the entire population and each group. Table 1 summarizes the overall performance.

3.2. Explanatory framework

In this section, we describe the foundations of our explanatory framework that aims to study the impact of voice characteristics on the performance of different speaker encoders. The central question in the explanatory modeling research is the choice of explanatory variables. Two main approaches exist for choosing them, based on confirmatory or exploratory research. They can be

regarded as two complementary components of the same goal, i.e., finding relevant variables in the most efficient, reliable, and replicable way. The difference is that, in confirmatory research, the potential impact of different variables is hypothesized a-priori, based on existing theories. The confirmatory research approach is useful when researchers have a theory (or theories) supported by facts. The second approach is exploration-driven, which is used when there exists a lack of sufficient theory foundations. Exploratory research could likewise produce new hypotheses that could formally be evaluated later. Our study belongs to the second category, as we design a general framework. With it, we studied the impact of voice characteristics on speaker verification performance, in terms of false acceptance rates.

Explanatory variables extraction. The explanatory variables considered in our study include voice characteristics pertaining to a wide range of perspectives. Each user u was represented in terms of two main categories of characteristics: protected and non-protected. Non-protected characteristics (e.g., jitter, shimmer) were extracted from each speech waveform w belonging to user u and averaged across speech waveforms of that user u to obtain a vector x of size S, where S is the number of non-protected characteristics. As we will describe later, non-protected characteristics can be further divided into three sub-categories: *quantitative, qualitative*, and *dialogue*. Conversely, protected characteristics (e.g., gender and age sensitive attributes) were defined at user level and represented with the vector z of size T, where T is the number of protected attributes. Formally, a user u was represented as a vector $c_u = [z; j] \in \mathbb{R}^{S+T}$, where $[;]$ is the concatenation operator.

Indeed, speech can be influenced by protected attributes, such as age and body conformation. Even though the availability of protected explanatory variables in corpora adopted for speaker recognition is limited, our study considered the following three protected attributes included into `FairVoice`:

- **Gender** of the speaker, self-reported by users, represented as a binary label (male, female).
- **Age Range** of the speaker, with the label "younger" assigned to those with age ≤ 40, "older" otherwise.
- **Language** spoken by the speaker (English, Spanish).

Non-protected quantitative variables measure properties common to any audio signal and do not have any direct relation with personal user traits. Specifically, we considered:

- **Root mean square (RMS)** is the loudness of the audio signal, measured as the power of the wave averaged across its length; a low-volume audio sample could negatively affect recognition performance.
- **Decibels relative to full scale (dBFS)** represents the loudness of the audio signal in decibel (dB) units, under a logarithmic scale, relative to the maximum possible loudness.
- **Maximum Amplitude** that is reached by the sound wave.
- **Intensity** (Mean, Std. Dev., Skewness, Kurtosis) is the power of the audio signal per unit area perpendicularly to that area, measured in dB SPL (Sound Pressure Level).
- **Signal-to-Noise Ratio (SNR)** measures the noise of the audio signal in dB, where a lower value reveals a high noise in the audio signal.

Non-protected qualitative variables include all those characteristics of a audio signal that differ depending on the source that generated it. The vocal folds that produce the human voice are an organic structure. Hence, the oscillations of the voice could contain significant fluctuations. Characteristics like fundamental frequency or jitter are affected by the context of the dialogue. Specifically, we considered:

- **Harmonics-to-Noise Ratio/Harmonicity (HNR)** (Mean, Std. Dev., Skewness, Kurtosis) represents the degree of acoustic periodicity, with high values for signals where most of the energy is in the periodic part. This measure is influenced by personal traits and medical conditions [37].
- **Fundamental Frequency F0** (Mean, Std. Dev., Skew, Kurtosis) of a speech signal refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. The sound wave is divided into several windows, and F0 is extracted for each one as the average number of oscillations per second and expressed in Hertz. This property depends on gender, age, overall body size, and cultural aspects [38].
- **Formants F1, F2, F3, F4 Frequencies** (Mean, Std. Dev., Skewness, Kurtosis) are the first four lowest resonant frequencies of the vocal tract [38]. There is a significant positive correlation between vocal tract length and body size (either height or weight), but also clear differences in male and female vocal tract morphology [39]. After data cleaning, F1 skewness, F3 kurtosis and F4 kurtosis were maintained.
- **Formant Position** is the average standardized formant value for the first n (we use $n = 4$) formants [40].
- **Jitter** is the variation in signal frequency caused by irregular vocal fold vibration, included in all natural speech. This measure is influenced by several factors, such as loudness, language, gender, and personal habits, e.g., smoking or alcohol consumption [38]. In our study, jitter is measured with the *local* variation of Boersma and Weenink [37] implementation, defining it as the average absolute difference between consecutive periods divided by the average period.
- **Shimmer** is similar to jitter, but accounts for the variation in amplitude. This measure depends on personal traits. We adopted the *local dB* variation from Boersma and Weenink [37], defined as the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

When considering sound waves containing human voices, aspects related to what a person is saying and how the speech is made can influence the performance of a speaker recognition system. Non-protected dialogue variables include properties related to the way speech is generated by a speaker, such as:

- **Number of Syllables** could impact the recognition task.
- **Number of Pauses** in a speech could be relevant for the speaker recognition system.
- **Rate of Speech** is the number of syllables pronounced along the entire audio sample, related to the propensity of the user to speak in a certain amount of time.
- **Articulation Rate**, differently from the rate of speech, is the number of syllables pronounced only over the speaking duration; so it describes how fast the user speaks.
- **Speaking Duration Without Pauses (SDWP)** counts the total duration in seconds of the portions of the audio example where the user is speaking.

Dependent variables preparation. The dependent variable is the performance of the speaker recognition model at the user level. Performance is estimated using the False Acceptance Rate (FAR) experienced by the user, to align our evaluation protocol with other works analyzing the impact of voice data manipulation [41,42] focused more on security against imposture. FAR is the measure of the likelihood that the biometric security system will incorrectly accept an attempt by an unauthorized user. A system FAR is computed as the ratio of the number of false acceptances divided by the number of imposter verification attempts. False accepts are often the most serious security issue as it gives unauthorized users access to systems. Given verification pairs $\mathcal{D}$, the FAR for a user u

is computed as:

$$FAR(u) = \frac{|\{(u,p) \in \mathcal{D} \mid v_\tau(w_p, w_u) = 1 \cap u \neq p\}|}{|\{(u,p) \in \mathcal{D} \mid u \neq p\}|} \quad (3)$$

For simplicity, with FAR as dependent variable, in our experiments we consider secure (label 1) a value of $FAR = 0$ and insecure (label 0) a value of $FAR > 0$.

Explanatory model creation. Explanatory modeling refers to the application of statistical models to data for testing causal hypotheses about theoretical constructs. In this study, on the basis of the components described in the previous sections, the causal hypothesis is: *can explanatory variables, capturing voice characteristics, explain the variation of the dependent variable related to speaker verification performance*? To this end, we introduce a surrogate model as the explanatory model, which, for a certain speaker recognition system, is optimized to explain the dependent variable (FAR) by means of the explanatory variables extracted from the speakers' utterances. To analyze the dependency of the performance on the explanatory variables, we considered random forests and linear models as a surrogate model, leaving the usage of other families of explanatory models for future works. Formally, an exploratory statistical model $\mathcal{G}$ for a speaker verification system v can then be defined as:

$$\mathcal{G}_\theta : f(c_u) = \hat{y}^v \quad \text{s.t.} \quad h(\mathcal{G}) = \Psi^v \quad (4)$$

where $\hat{y}^v$ is the prediction of $\mathcal{G}$, h is a function that from $\mathcal{G}$ returns the importance weights $\Psi^v \in \mathbb{R}^{S+T}$, which are the hypothesized impactful parameters that vary in terms of the information captured from each characteristic in c_u. Training a surrogate model becomes then an optimization problem:

$$\underset{\theta}{\operatorname{argmin}} |f(c_u) - y^v| \quad (5)$$

where $\theta \in \mathbb{R}^*$ is a set of parameters, i.e., rules used internally by $\mathcal{G}$ to be optimized. Surrogate models were optimized via a grid search on all the audio samples included in the testing set. Specifically, the vector c of explanatory variables characterizing each user was fed as input of the surrogate model. The dependent variable was considered as the ground truth value to predict. Since we are interested in the explanation power of the surrogate models, no further split of this set is performed.

4. Experimental results

Our experiments analyzed whether protected attributes correlate with other speech covariates (RQ1), which speech covariates influence performance the most (RQ2), and whether changing the protected class affects performance (RQ3).

4.1. Relationship between explanatory variables (RQ1)

In a first analysis, we investigated whether protected explanatory variables have any relationship with other speech covariates we considered as explanatory variables (quantitative, qualitative, and dialogue). To this end, Fig. 2 shows the Pearson correlation among the considered explanatory variables. For conciseness, we present only the results for the Random Forest (RF) as a surrogate model, since it achieved a value close to 1 for both F1 score and AUC and can hence well explain the relationship of the dependent variable with the explanatory variables. We also played with linear models, but achieved both F1 Score and AUC lower than 0.65. For clarity, we also removed non-protected variables highly correlated with each other.

It can be observed that there was a high absolute correlation between gender and other voice characteristics, such as *F0 mean* and statistical moments measured on the distribution of the 4 formants (F1, F2, F3, F4). Similarly, *jitter local* and *shimmer local dB* had a positive correlation with gender as well, confirming that these characteristics are able to encode personal traits of each individual. Conversely, age range and language did not report any significant correlation with other speech covariates.

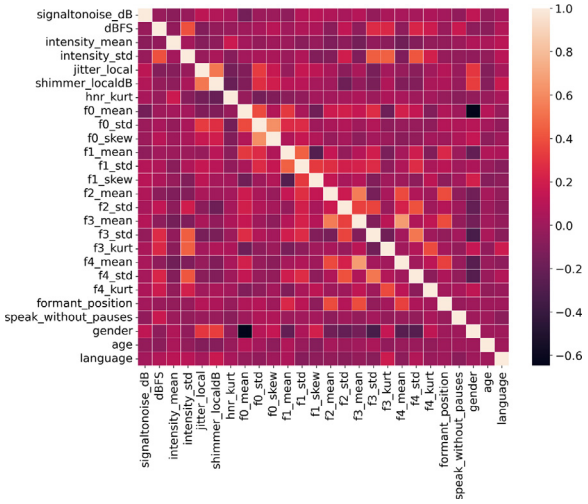

Fig. 2. [RQ1] Correlation between explanatory variables.

Observation 1. There are significant relationships between gender and other speech covariates. Age and language do not relate significantly with any covariates.

4.2. Influence of speech covariates on performance (RQ2)

In a second analysis, we analyzed which speech covariates influence speaker recognition performance the most. In particular, we examined the dependency of the FAR (security) from the considered explanatory variables by means of the surrogate models included in our explanatory framework. Before training the surrogate models, a variance inflation factor [12] was applied to remove multi-collinearity among explanatory variables (threshold equal to 5.0), which resulted in removing a range of less influential explanatory variables. The remaining ones were used to train the surrogate models. In order to uncover the influence of explanatory variables on recognition performance, we leverage techniques of permutation feature importance on the surrogate models, applied over 10 repetitions to ensure statistical significance.

Figure 3 collects the explanatory variable importance scores on ResNet-34 and X-Vector, respectively. It can be observed that the RF surrogate model considered the formants F1, F3, and F4 as well as the fundamental frequency F0 to be the most important variables for both speaker encoders. None of the protected explanatory variables were considered as important by the surrogate model, except for language in X-Vector. Overall, our results reveal that protected attributes do not directly affect the performance of the speaker recognition system. Other speech covariates, despite still being correlated with protected attributes, can be used to interpret and then counteract unfairness in terms of security on both speaker encoders.

Observation 2. Speech covariates pertaining to vocal frequency aspects explain the most the disparate security estimates across individuals.

4.3. Impact of protected class changes (RQ3)

Previous experiments revealed that there exists a relationship between protected attributes (especially gender) and other voice characteristics (RQ1) and that some key voice characteristics can

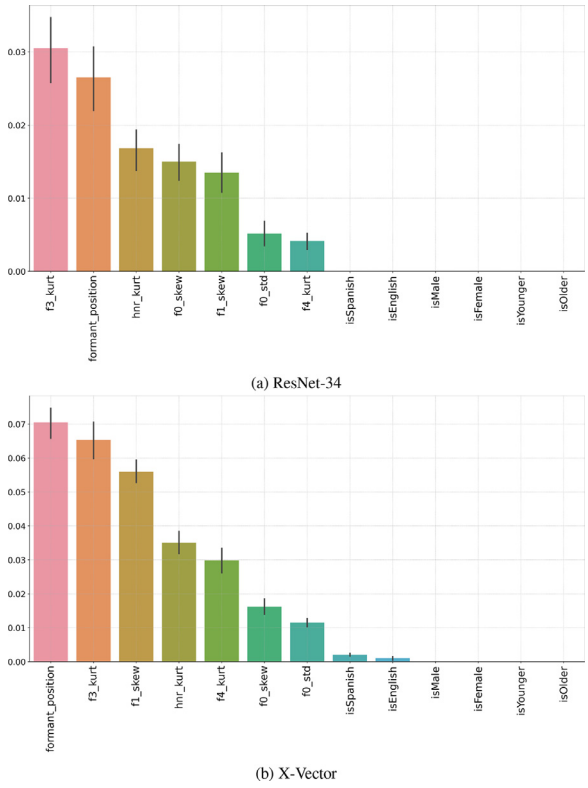

Fig. 3. [RQ2] Importance of voice characteristics on the predicted FAR.

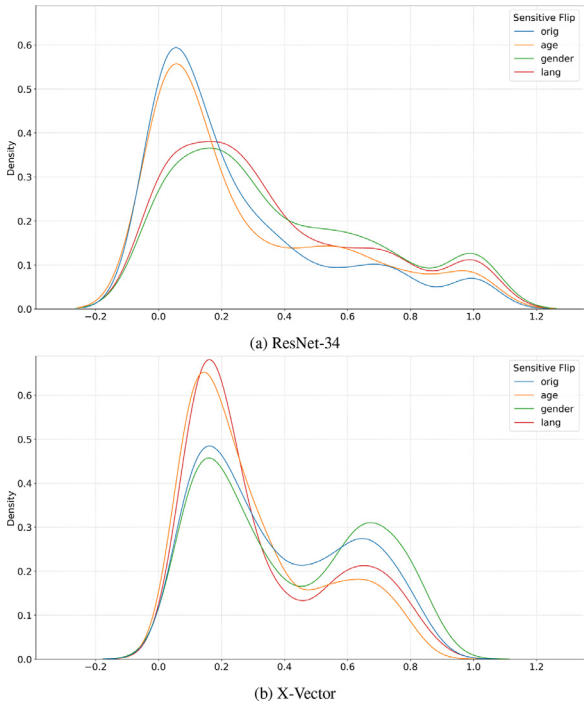

Fig. 4. [RQ3] Effect of flipped protected class on predicted FAR.

explain the error rates experienced by a speaker recognition system (especially FAR) to a good extent (RQ2). In our third and last analysis, we therefore leveraged the surrogate model to investigate what happens to the dependent variable when we flip the protected class of a user in his/her vector c, e.g., by modifying the gender (age; language) of a user from female (younger; English) to male (older; Spanish). Our goal is to compare the predictions of the surrogate model when the original vector and the vector with the flipped protected attribute are fed, respectively. Through our surrogate model, we provided "what if" feedback of the form "if an input data point was c_u' instead of c_u, then a speaker encoder outcome would be $\hat{y}$', and not $\hat{y}$".

Figure 4 reports the predicted FAR for original vectors and vectors with a protected class flipped on ResNet-34 and X-Vector respectively. The *orig* curve depicts the density distribution of the predicted FAR when the original vector of each speaker was used. The other curves represent the predicted FAR when the corresponding protected attribute is "flipped" for each user, i.e., the curve labeled *gender* was generated by flipping the gender class of each speaker and similarly for *language* and *age* range. It can be observed that flipping the gender and language classes resulted in a significant increase of predicted FAR level on ResNet-34. Conversely, flipping the language and age classes positively affected FAR predictions on X-Vector. Hence, the sensitive attributes, especially the language, are able to modify the predictions of RF. This observation is also in line with what we observed in the first two analyses.

Observation 3. The spoken language has the strongest impact on the security of the two considered speaker recognition systems.

5. Discussion and conclusions

In this paper, we aimed to explain the reasons behind disparate impacts in speaker recognition through the lens of voice characteristics, going beyond a mere attribution to the imbalance level across classes. To this end, we formulated and leveraged an explanatory framework that sheds light on the impact of protected and non-protected characteristics on the recognition performance of speaker encoders. Even though causal reasoning techniques reported an influence of protected attributes, our analysis of the importance of other speech covariates showed a significant relation of vocal properties with system security, agreeing with the findings of the other studies such as Fenu et al. [9,11].

Our findings in this study, paired with its limitations, will lead to future research directions. First, our observations prove that the causes of disparate performances go beyond mere memberships to certain demographic groups, but result from fine-grained voice characteristics (some of them related to the group membership). This opens a new perspective for analysis and mitigation of unfairness in speaker recognition where it might be no longer required to know the (hard to retrieve, especially due to privacy constraints) protected attribute labels. Other voice covariates emerged from our analysis can be used as a real proxy of such labels and as drivers for specific mitigation strategies (e.g., clustering users based on those characteristics and provide treatments to the disadvantaged clusters). Another line of research can also focus on making input waveform statistically indistinguishable from the perspective of the relevant voice characteristics, for instance through the use of autoencoders, in order to make speaker encoders robust to these characteristics. Nevertheless, the set of characteristics we considered can be extended, and other large data sets including more languages, protected attributes, and speech covariates can be leveraged. Furthermore, random forests were the solely explanatory model that resulted in good accuracy, but our framework can be extended to further surrogate models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data includes already existing publicly available dataset.

References

[1] Y.B. Abera, Y. Naudet, H. Panetto, A new paradigm and meta-model for cyber–physical-social systems, IFAC-PapersOnLine 53 (2) (2020) 10949–10954.
[2] B.A. Yilma, Y. Naudet, H. Panetto, Introduction to personalisation in cyber–physical-social systems, in: International Conferences on the Move to Meaningful Internet Systems, Springer, 2018, pp. 25–35.
[3] M. Jesús-Azabal, J. Rojo, E. Moguel, D. Flores-Martin, J. Berrocal, J. García-Alonso, J.M. Murillo, Voice assistant to remind pharmacologic treatment in elders, in: International Workshop on Gerontechnology, Springer, 2019, pp. 123–133.
[4] C.E. Rhee, J. Choi, Effects of personalization and social role in voice shopping: an experimental study on product recommendation by a conversational voice agent, Comput. Hum. Behav. 109 (2020) 106359.
[5] A. Ross, S. Banerjee, A. Chowdhury, Security in smart cities: a brief review of digital forensic schemes for biometric data, Pattern Recognit. Lett. 138 (2020) 346–354.
[6] S. Hussain, O. Ameri Sianaki, N. Ababneh, A survey on conversational agents/chatbots classification and design techniques, in: Workshops of the International Conference on Advanced Information Networking and Applications, Springer, 2019, pp. 946–956.
[7] D. Snyder, D. Garcia, G. Sell, D. Povey, S. Khudanpur, X-vectors: robust dnn embeddings for speaker recognition, in: Proc. ICASSP 2018, 2018.
[8] J.S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: deep speaker recognition, in: Proc. Interspeech 2018, 2018.
[9] G. Fenu, M. Marras, G. Medda, G. Meloni, Fair voice biometrics: impact of demographic imbalance on group fairness in speaker recognition, in: Proc. Interspeech 2021, ISCA, 2021, pp. 1892–1896.
[10] G. Fenu, H. Lafhouli, M. Marras, Exploring algorithmic fairness in deep speaker verification, in: Proc. the International Conference eon Computational Science and Its Applications (ICCSA), 2020, pp. 77–93.
[11] G. Fenu, G. Medda, M. Marras, G. Meloni, Improving fairness in speaker recognition, in: ESSE 2020: 2020 European Symposium on Software Engineering, Rome, Italy, November 6–8, 2020, ACM, 2020, pp. 129–136.
[12] Y. Deldjoo, A. Bellogín, T.D. Noia, Explaining recommender systems fairness and accuracy through the lens of data characteristics, Inf. Process. Manag. 58 (5) (2021) 102662.
[13] V. La Gatta, V. Moscato, M. Postiglione, G. SperlkA, Pastle: pivot-aided space transformation for local explanations, Pattern Recognit. Lett. 149 (2021) 67–74.
[14] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1–3) (2000) 19–41.
[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Speech Audio Process. 19 (4) (2011) 788–798.
[16] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with sincnet, in: Proc. SLT 2018, 2018, pp. 1021–1028.
[17] A. Nagrani, J.S. Chung, W. Xie, A. Zisserman, Voxceleb: large-scale speaker verification in the wild, Comput. Speech Lang. 60 (2020) 1–15.
[18] Z. Wang, K. Yao, X. Li, S. Fang, Multi-resolution multi-head attention in deep speaker embedding, in: Proc. ICASSP 2020, 2020, pp. 6464–6468.
[19] S. Yadav, A. Rai, Frequency and temporal convolutional attention for text-independent speaker recognition, in: Proc. ICASSP, 2020, pp. 6794–6798.
[20] L. Wan, Q. Wang, A. Papir, I. Lopez-Moreno, Generalized end-to-end loss for speaker verification, in: Proc. ICASSP, 2018, pp. 4879–4883.
[21] W. Xie, A. Nagrani, J.S. Chung, A. Zisserman, Utterance-level aggregation for speaker recognition in the wild, in: Proc. ICASSP 2019, 2019, pp. 5791–5795.
[22] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 115:1–115:35.
[23] B. Goodman, S.R. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation, AI Mag 38 (3) (2017) 50–57.
[24] P. Terhörst, J.N. Kolf, N. Damer, F. Kirchbuchner, A. Kuijper, Post-comparison mitigation of demographic bias in face recognition using fair score normalization, Pattern Recognit. Lett. 140 (2020) 332–338.
[25] G. Nápoles, L. Koutsoviti Koumeri, A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets, Pattern Recognit. Lett. 154 (2022) 29–36.
[26] Y. Zhang, Y. Zhang, B.M. Halpern, T. Patel, O. Scharenborg, Mitigating bias against non-native accents, in: H. Ko, J.H.L. Hansen (Eds.), Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18–22 September 2022, ISCA, 2022, pp. 3168–3172.
[27] Y. Meng, Y.-H. Chou, A.T. Liu, H.-y. Lee, Don't speak too fast: the impact of data bias on self-supervised speech models, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 3258–3262.
[28] H. Shen, Y. Yang, G. Sun, R. Langman, E. Han, J. Droppo, A. Stolcke, Improving fairness in speaker verification via group-adapted fusion network, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23–27 May 2022, IEEE, 2022, pp. 7077–7081.
[29] R. Peri, K. Somandepalli, S. Narayanan, A study of bias mitigation strategies for speaker recognition, Comput. Speech Lang. 79 (2023) 101481.
[30] W.T. Hutiri, A.Y. Ding, Bias in automated speaker recognition, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21–24, 2022, ACM, 2022, pp. 230–247.
[31] G. Fenu, M. Marras, Demographic fairness in multimodal biometrics: acomparative analysis on audio-visual speaker recognition systems, Procedia Comput. Sci. 198 (2022) 249–254.
[32] M.J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.
[33] J. Joo, K. Kärkkäinen, Gender slopes: counterfactual fairness for computer vision models by attribute manipulation, in: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, FATE/MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–5.
[34] S. Garg, V. Perot, N. Limtiaco, A. Taly, E.H. Chi, A. Beutel, Counterfactual fairness in text classification through robustness, in: Proc. AIES 2019, ACM, 2019, pp. 219–226.
[35] L. Sari, M. Hasegawa-Johnson, C.D. Yoo, Counterfactually fair automatic speech recognition, IEEE ACM Trans. Audio Speech Lang. Process. 29 (2021) 3515–3525.
[36] D. Snyder, D. Garcia, G. Sell, D. Povey, S. Khudanpur, X-vectors: robust DNN embeddings for speaker recognition, in: Proc. ICASSP 2018, 2018, pp. 5329–5333.
[37] P. Boersma, D. Weenink, Praat: doing phonetics by computer [computer program], 2022,
[38] T. Bäckström, O. Räsänen, A. Zewoudie, P.P. Zarazaga, L. Koivusalo, Introduction to speech processing, 2019,
[39] W.T. Fitch, J. Giedd, Morphology and development of the human vocal tract: a study using magnetic resonance imaging, J. Acoust. Soc. Am. 106 (3 Pt 1) (1999) 1511–1522.
[40] D.R. Feinberg, Parselmouth praat scripts in Python, 2022,
[41] T. Masuko, K. Tokuda, T. Kobayashi, Imposture using synthetic speech against speaker verification based on spectrum and pitch, in: Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16–20, 2000, ISCA, 2000, pp. 302–305.
[42] D. Matrouf, J. Bonastre, C. Fredouille, Effect of speech transformation on impostor acceptance, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14–19, 2006, IEEE, 2006, pp. 933–936.