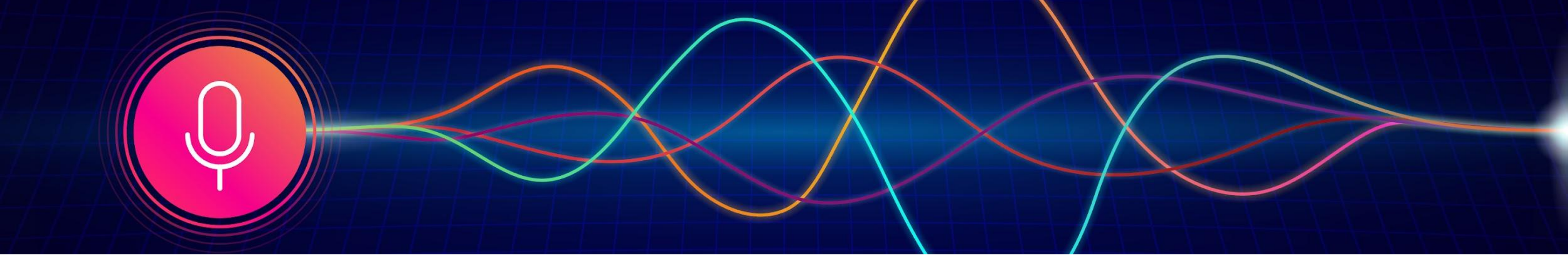


Detecting and mitigating bias in voice activated technologies

Presented by Wiebke (Toussaint) Hutiri, Delft University of Technology

to Microsoft Africa Research Institute

25 May 2022

Introduction

- In progress: PhD @ TU Delft (Netherlands) on Trustworthy Edge AI
- MSc in Comp. Sci. from UCT (South Africa)
- BSc Mech. Eng. from UCT (South Africa)

@wiebketous

What I'll talk about today

1. Contextualising Voice Activation
2. Bias in Automated Speaker Recognition
3. Inclusive Speaker Verification Evaluation Datasets
4. Fair EVA
5. Discussion

Contextualising Voice Activation

Speech: a great source of information!

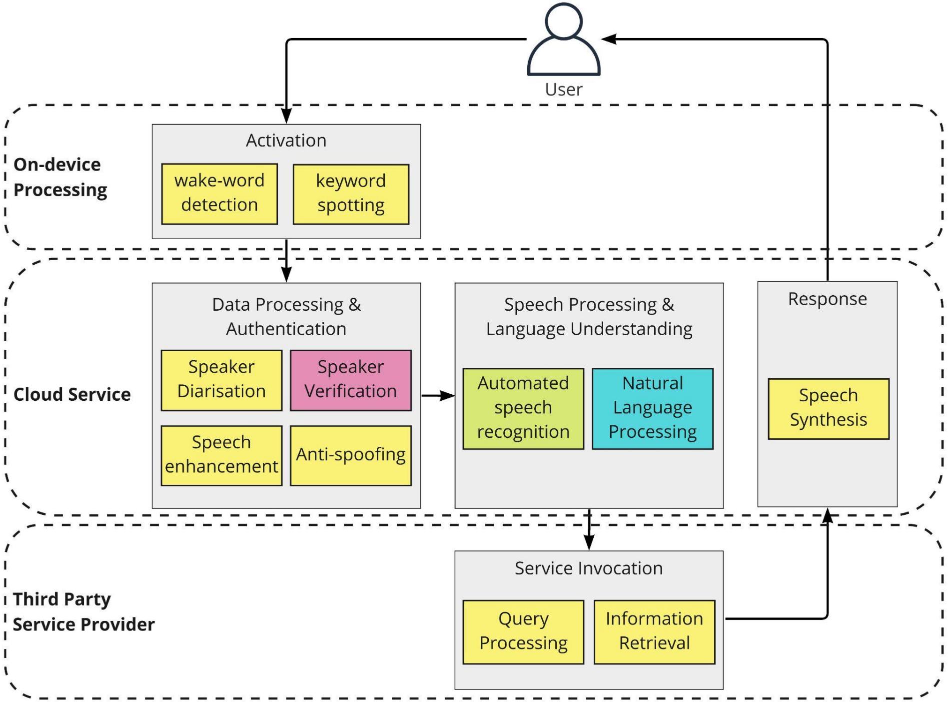
- Words
- Emotion
- Age
- Gender
- Regional/non-native accent
- Language
- **Identity → Speaker Recognition**

Contextualising Voice Activation

Speaker recognition:

- Speaker identification: *who spoke?*
- Speaker diarisation: *separate speakers*
- **Speaker verification:** *is the speaker who they claim to be?* → Voice Biometrics

Contextualising Voice Activation



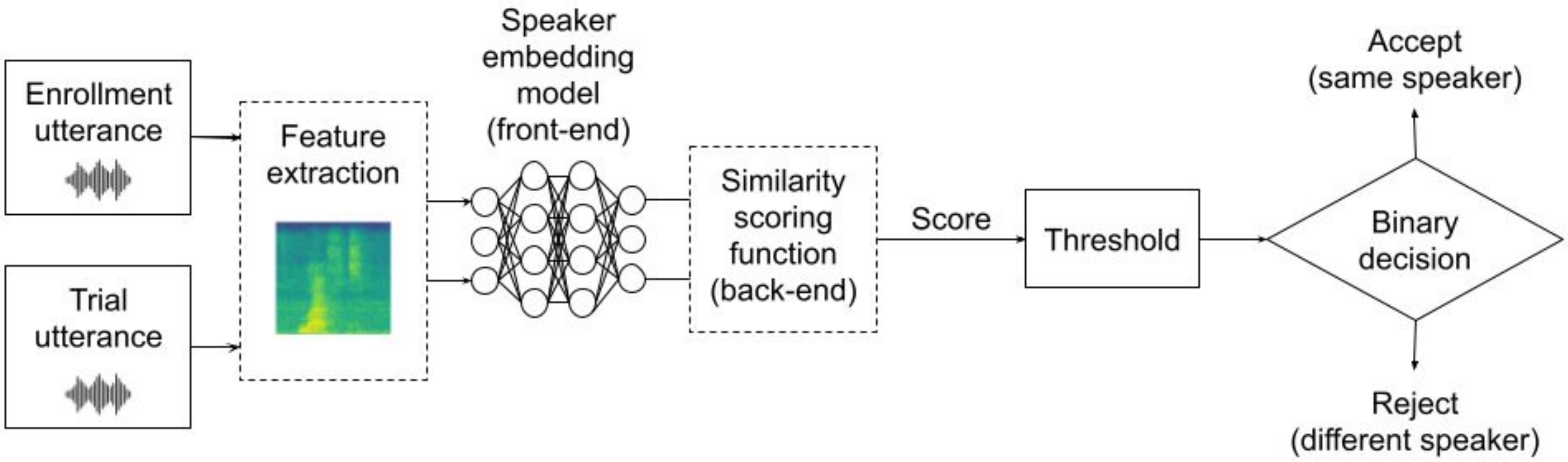
Bias in Automated Speaker Recognition*

Insights & Contribution

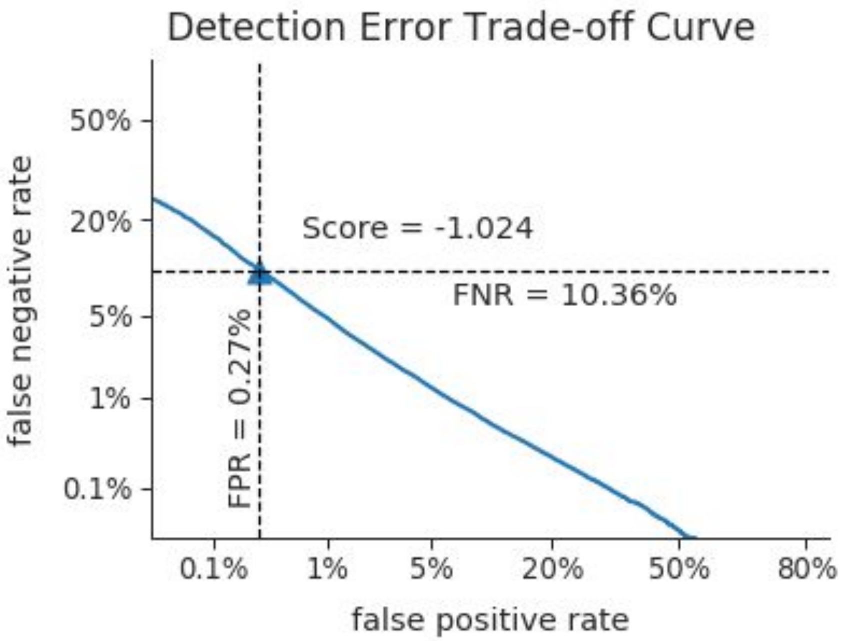
1. Present evaluation framework to **quantify performance disparities** in Speaker Verification (SV)
2. First evaluation of **bias in SV** → bias exists **at every stage** of the ML development pipeline
3. Recommend research directions to address bias in SV

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

Overview of Speaker Verification



Speaker Verification Evaluation

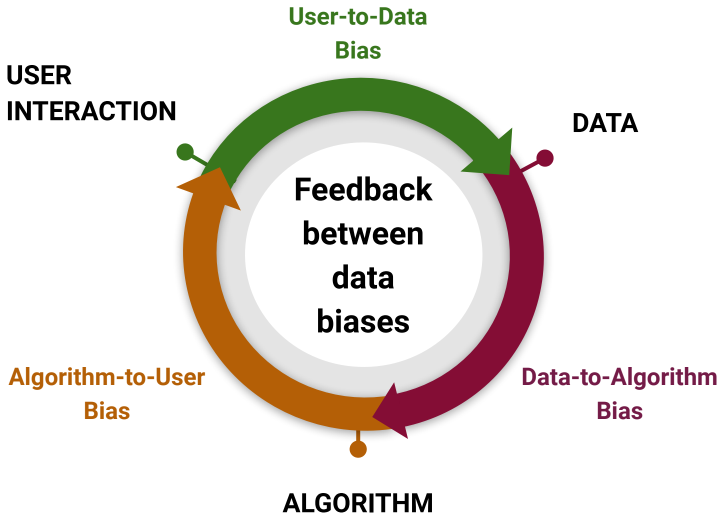


Fairness, Bias and Discrimination in ML

Fairness [1]**:**
Absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.

Bias [1]**:**
A source of unfairness, e.g. due to the data collection, sampling and measurement.

Discrimination [1]**:**
A source of unfairness due to human prejudice and stereotyping based on sensitive or protected attributes



1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning.

Research Approach

Empirical & analytical study of group bias in the VoxCeleb Speaker Recognition Challenge.

Experiment setup

- Models: two 34 layer ResNets trained on VoxCeleb2
- Evaluation Dataset: VoxCeleb 1
- Subgroups: speaker gender & nationality
- Bias evaluation measure:

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

7 Sources of Harm in the ML Life Cycle[2]

Data Generation

1. Historical bias
2. Representational bias
3. Measurement bias

Model Building & Implementation

4. Aggregation bias
5. Learning bias
6. Evaluation bias
7. Deployment bias

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

Bias in Automated Speaker Recognition

Historical Bias

Replicates biases, like stereotypes, that are present in the world as is or was.

VoxCeleb1 automated data generation pipeline:

1. VGGFace dataset → candidate speakers
   a. most searched names in Freebase knowledge graph & IMDB
2. HOG-based face detector → track faces
3. SYNC-Net → identify active speakers
4. VGG Face CNN → verify speaker's identity

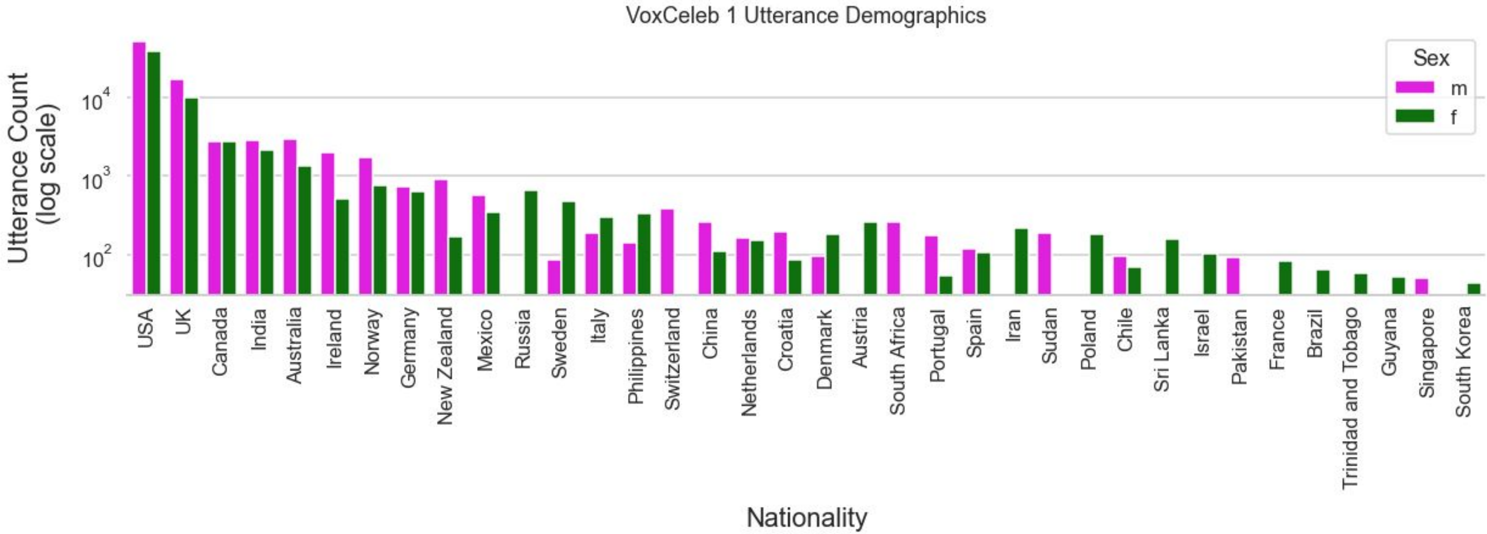
⇒ pipeline reinforces popularity bias in search results

⇒ bias in face recognition directly transferred to speaker verification

Bias in Automated Speaker Recognition

Representation Bias

Underrepresents a subset of the population in its sample, resulting in poor generalization for that subset.



Bias in Automated Speaker Recognition

Measurement Bias

Occurs in the process of designing features and labels to use in the prediction problem

Labelling choices in metadata

→ used for judgements about representation in dataset

→ inform subgroup design and thus bias evaluation

Nationality labels: speaker's citizenship from Wikipedia

Conflates accent and nationality, language not considered

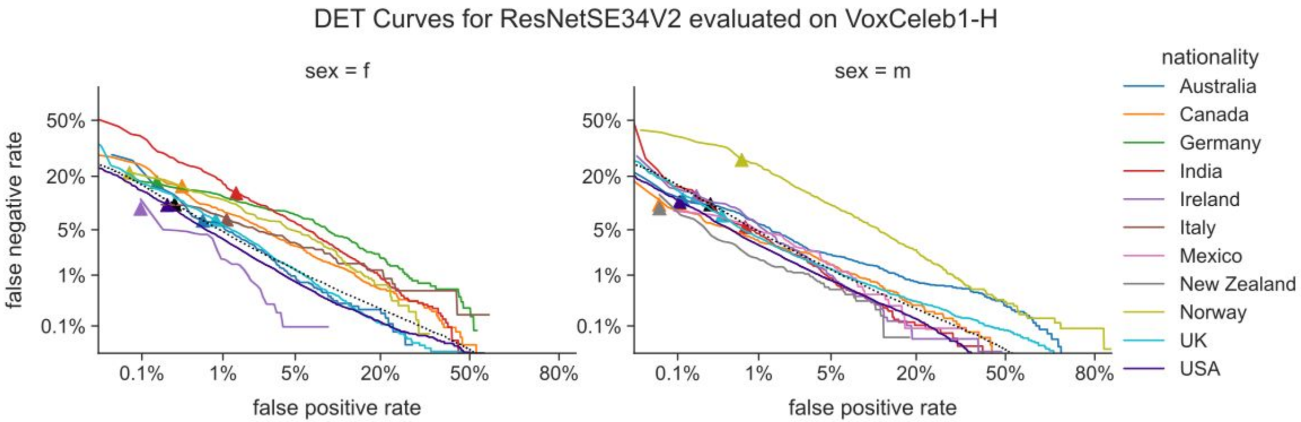
Nationality labels still have merit

Gender labels: labelling process unclear, only binary categories

Bias in Automated Speaker Recognition

Aggregation Bias

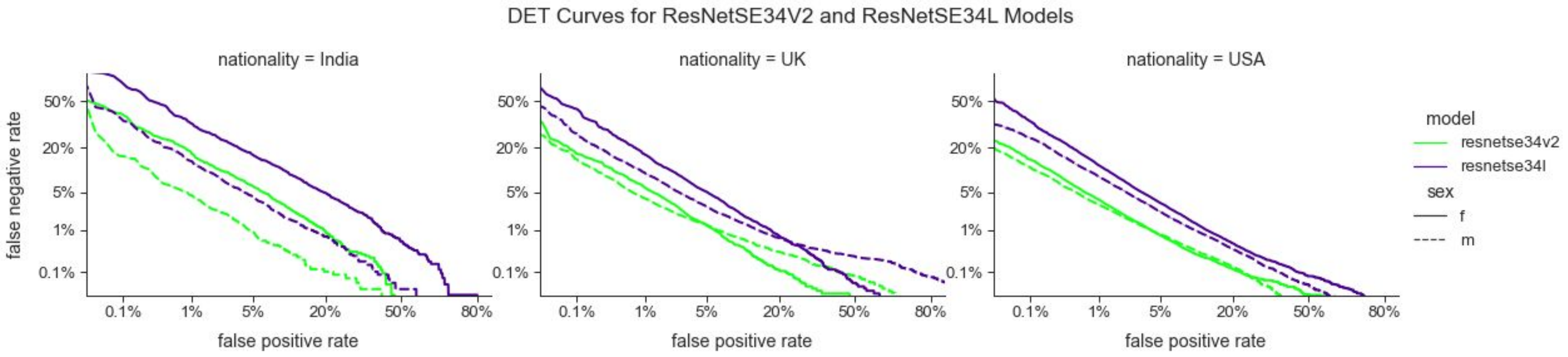
Arises when data contains underlying groups that should be treated separately, but that are instead subjected to uniform treatment.



Bias in Automated Speaker Recognition

Learning Bias

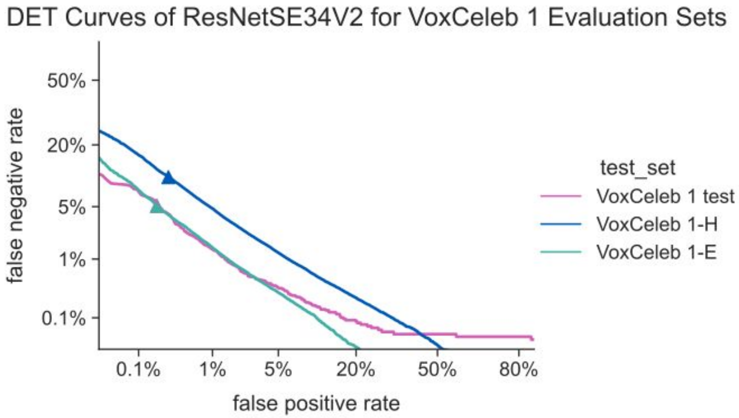
Concerns modeling choices and their effect on amplifying performance disparities across samples.



Bias in Automated Speaker Recognition

Evaluation Bias

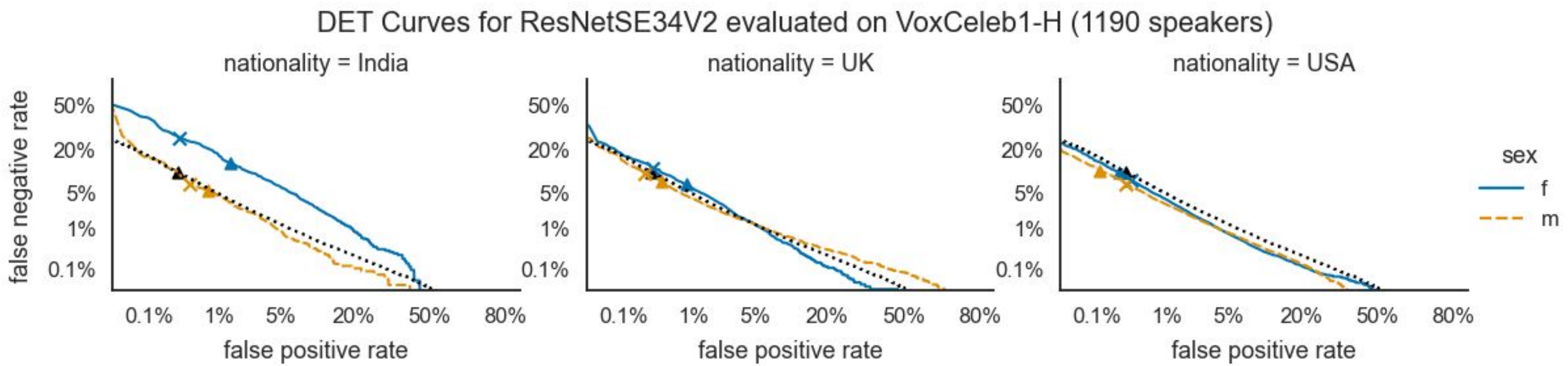
Is attributed to a benchmark population that is not representative of the user population, and to evaluation metrics that provide an oversimplified view of model performance.



Bias in Automated Speaker Recognition

Deployment Bias

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



Design Guidelines for Inclusive Speaker Verification Evaluation Datasets*

Insights & Contribution

1. **Difficulty** of utterance pairs impacts evaluation outcome
2. Difficulty **distribution varies** across speakers and groups
3. **Randomized** utterance pairings can result in significant performance variation if the utterance pair count / speaker is low
4. We propose an algorithm for **generating robust & inclusive** evaluation datasets from utterance pairs

★ Wiebke Toussaint Hutiri, Lauriane Gorce and Aaron Yi Ding. 2022. **Design Guidelines for Inclusive Speaker Verification Evaluation Datasets**. https://arxiv.org/abs/2204.02281

Schema for Grading Utterance Pairs

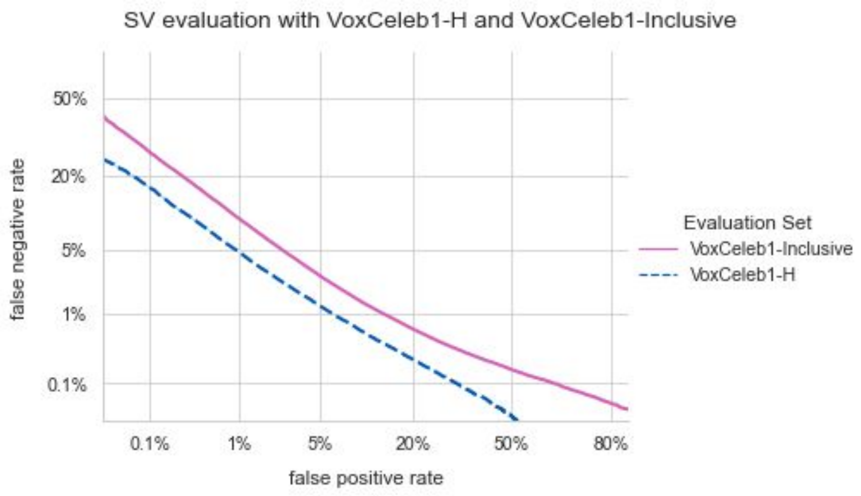
Utterance Pairs	Difficulty	Same Gender	Same Nationality	Same Channel	Same Noise
Same Speaker	cat 1 (trivial)	-	-	Yes	Yes
	cat 3 (medium)	-	-	No	n.k.
Different Speakers	cat 1 (trivial)	No	No	-	n.k.
	cat 2 (easy)	No	Yes	-	n.k.
	cat 3 (medium)	Yes	No	-	n.k.
	cat 4 (hard)	Yes	Yes	-	n.k.

Table 1: *Grading of utterance pairs (n.k. = not known)*

Effect of Utterance Pair Grading

Nationality	Speakers	Pairs	Pairs/ speaker	cat 1 (trivial)
USA	799	178122	222.9	12.9%
UK	215	53111	247.0	10.3%
Canada	54	10864	201.2	11.1%
India	26	10053	386.7	10.6%
Australia	37	8668	234.3	10.5%
Ireland	18	4960	275.6	8.5%
Norway	20	4906	245.3	10.0%
New Zealand	6	1811	301.8	10.1%
Germany	5	1256	251.2	17.0%
Mexico	5	1130	226.0	10.2%
Italy	5	571	114.2	17.0%

Table 3: *VoxCeleb1-H Same Speaker Utterance Pairs.*



Effect of Utterance Pair Count

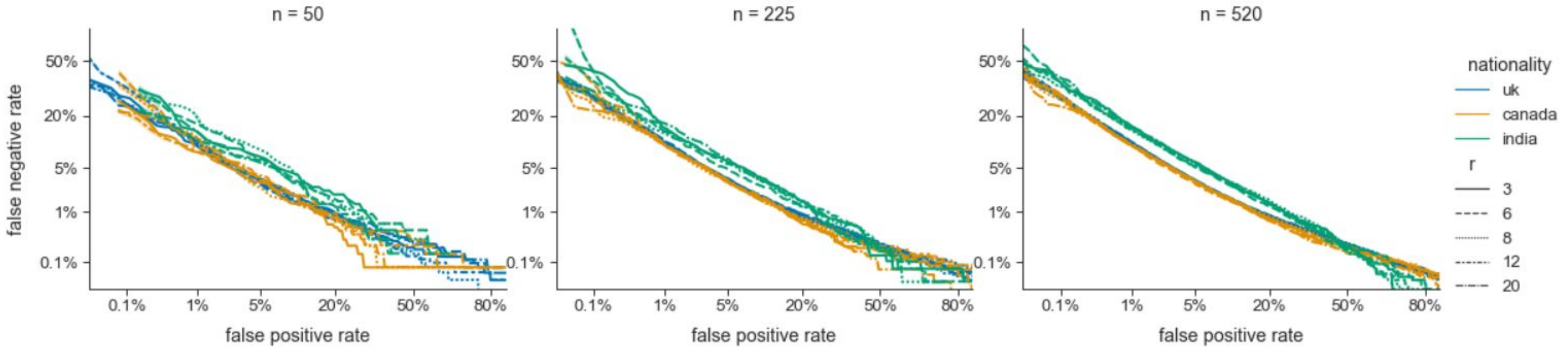


Figure 2: *DET curves show variability in evaluation outcomes for evaluation sets with 50, 225 and 520 utterance pairs (n) for Canadian, Indian and UK speakers. For each n five datasets were generated with different random seeds:* $r = 3, 6, 8, 12, 20$.

Dataset Design Guidelines

Inclusive evaluation datasets for robust speaker verification evaluation should have:

1. Equal # same speaker & different speaker utterance pairs for each speaker
2. At least 500 different speaker utterance pairs / speaker
3. Equal # utterance pairs / speaker
4. Equal distribution of difficulty gradings across utterance pairs / speaker
5. Utterance pairs with difficulty gradings representative of real-life usage scenarios
6. Several randomly generated utterance pairings

An Intro to Fair EVA

1. Voice technologies should work reliably for all users
2. Unchecked use of data and AI in their development raises concerns about bias and discrimination
3. We are building an audit tool, dataset and knowledge base to evaluate bias in voice biometrics.

moz://a Proud recipient of a Mozilla Tech Fund Award

Fair EVA Projects

Bias Tests for Voice Tech
Python library

Fair Evaluation Guidelines for speaker verification

Technology Audit
Of commercial voice biometrics products

Voice Biometrics 101
Interactive Multimedia for civil society

Database
Resource to investigate bias in voice tech

Find out more: https://www.faireva.org/

Discussion

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

★ Wiebke Toussaint Hutiri, Lauriane Gorce and Aaron Yi Ding. 2022. **Design Guidelines for Inclusive Speaker Verification Evaluation Datasets**. https://arxiv.org/abs/2204.02281

Contact w.toussaint@tudelft.nl

@wiebketous