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
Detecting and mitigating bias in voice activated technologies

Contextualising Voice Activation

User

On-device Processing
- Activation
  - wake-word detection
  - keyword spotting

Cloud Service
- Data Processing & Authentication
  - Speaker Diarisation
  - Speaker Verification
  - Speech enhancement
  - Anti-spoofing
- Speech Processing & Language Understanding
  - Automated speech recognition
  - Natural Language Processing

Third Party Service Provider
- Service Invocation
  - Query Processing
  - Information Retrieval
- Response
  - Speech Synthesis
Bias in Automated Speaker Recognition*

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

Overview of Speaker Verification

Detecting and mitigating bias in voice activated technologies
Speaker Verification Evaluation

Detection Error Trade-off Curve

- Score = -1.024
- FPR = 0.27%
- FNR = 10.36%

Detecting and mitigating bias in voice activated technologies
Fairness, Bias and Discrimination in ML

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

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

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

---

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:

\[
\text{subgroup bias} = \frac{C_{Det}(\theta_{\text{overall min}})^{SG}}{C_{Det}(\theta_{\text{overall min}})^{overall}}
\]
7 Sources of Harm in the ML Life Cycle

**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

---

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.*

VoxCeleb 1 Utterance Demographics

![Graph showing utterance count by nationality and sex](image-url)
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.

DET Curves for ResNetSE34V2 evaluated on VoxCeleb1-H

- sex = f
- sex = m

- nationalsities: Australia, Canada, Germany, India, Ireland, Italy, Mexico, New Zealand, Norway, UK, USA
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.*

DET Curves for ResNetSE34V2 evaluated on VoxCeleb1-H (1190 speakers)

- **nationality = India**
- **nationality = UK**
- **nationality = USA**

*false negative rate vs. false positive rate*
Design Guidelines for Inclusive Speaker Verification Evaluation Datasets*

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

## Schema for Grading Utterance Pairs

<table>
<thead>
<tr>
<th>Utterance Pairs</th>
<th>Difficulty</th>
<th>Same Gender</th>
<th>Same Nationality</th>
<th>Same Channel</th>
<th>Same Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Speaker</td>
<td>cat 1 (trivial)</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>cat 3 (medium)</td>
<td>-</td>
<td>-</td>
<td>No</td>
<td>n.k.</td>
</tr>
<tr>
<td>Different Speakers</td>
<td>cat 1 (trivial)</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>n.k.</td>
</tr>
<tr>
<td></td>
<td>cat 2 (easy)</td>
<td>No</td>
<td>Yes</td>
<td>-</td>
<td>n.k.</td>
</tr>
<tr>
<td></td>
<td>cat 3 (medium)</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>n.k.</td>
</tr>
<tr>
<td></td>
<td>cat 4 (hard)</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>n.k.</td>
</tr>
</tbody>
</table>

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

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

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.
Detecting and mitigating bias in voice activated technologies

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.

Proud recipient of a Mozilla Tech Fund Award
Fair EVA Projects

- **bt4vt**: 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 technology

Find out more: [https://www.faireva.org/](https://www.faireva.org/)
Discussion


Contact w.toussaint@tudelft.nl
@wiebketous