Learning to Fool the Speaker Recognition
ICASSP 2020

Jiguo Li,
jiguo.li@vipl.ict.ac.cn

Joint work with Xinfeng Zhang, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, Wen Gao

1 Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS)
2 Institute of Digital Media (IDM), Peking University (PKU)
3 Bytedance Inc
4 University of Chinese Academy of Science (UCAS)
Biometric Systems

◆ Image-based Systems
  ✔ Face, fingerprint

◆ Speech-based Systems
  ✔ Speaker recognition
Security Risks for Biometric Systems

◆ Image-based Systems
   ✓ Face attack

◆ Speech-based Systems
   ✓ Speech attack

Motivation

◆ How to attack the speech-based biometric systems?
◆ Is the speech-based biometric systems vulnerable to the adversarial attack?
◆ Is it possible to design a biometric systems robust to the adversarial attack?
## Our attack results

<table>
<thead>
<tr>
<th>Non-targeted attack</th>
<th>Targeted attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>dr1/fcjf0/si1027.ogg</td>
<td>dr1/fcjf0/si1027.ogg</td>
</tr>
<tr>
<td>dr1/fcjf0/sx37.ogg</td>
<td>dr1/fdaw0/si1046.ogg</td>
</tr>
<tr>
<td>dr2/faem0/si762.ogg</td>
<td>dr2/faem0/si762.ogg</td>
</tr>
<tr>
<td>dr8/fbcg1/sx82.ogg</td>
<td>dr8/fbcg1/si982.ogg</td>
</tr>
<tr>
<td>real</td>
<td>real</td>
</tr>
<tr>
<td>fake</td>
<td>target0</td>
</tr>
<tr>
<td></td>
<td>target100</td>
</tr>
<tr>
<td></td>
<td>target200</td>
</tr>
</tbody>
</table>

Proposed Attack Framework

- **Our Framework**
  - An attacker for all samples
  - Optimize the speech via phoneme recognition module
Proposed Attack Framework

Speaker/Phoneme Recognition Model: Sincnet[1]

- Frequency filters in the first layer
- Process on the raw waveform
- More interpretable

Proposed Attack Framework

◆ Attacker: a Residual Block

✓ Referring to Adversarial Transformer Networks (ATNs)[1]
✓ Additive perturbations
✓ The scale of the perturbation is controllable
✓ Training once for all testing samples

Proposed Attack Framework

**Adversarial training/Optimization**

✓ **Non-targeted attack**

\[
L_{total} = L_{spk} + \lambda_{phn}L_{phn} + \lambda_{norm}L_{norm}
\]
\[
L_{spk} = \begin{cases} 
    x'_{spk}[I_{1st}] - x'_{spk}[I_{2nd}], & I_{1st} = y_{spk} \\
    0, & \text{else}
\end{cases}
\]
\[
L_{phn} = KL(p_{phn}||p'_{phn})
\]
\[
L_{norm} = [\max(s - s' - m, 0)]^2
\]

✓ **Targeted attack**

\[
L_{spk} = \begin{cases} 
    x'_{spk}[I_{1st}] - x'_{spk}[y_{target}], & I_{1st} \neq y_{target} \\
    0, & \text{else}
\end{cases}
\]
Experimental Results

Datasets and Metrics

Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Label</th>
<th>Speaker number</th>
<th>Samples (train+test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>Speaker+phoneme</td>
<td>462</td>
<td>3694(2309+1385)</td>
</tr>
</tbody>
</table>

Metric

- Sentence Error Rate (SER): used for non-targeted attack
- Prediction Target Rate (PTR): used for targeted attack
- Signal-noise Ratio (SNR)
- Perceptual Evaluation of Speech Quality (PESQ): 0.5~4.5
Experimental Results

Can our proposed model attack the pretrained speaker recognition model?

- Non-targeted attack
- SER 90.5% with SNR 59.01 dB
- SER 90.5% with PESQ 4.28

<table>
<thead>
<tr>
<th>$\lambda_{phon}$</th>
<th>$\lambda_{norm}$</th>
<th>SER(%)↑</th>
<th>SNR(dB)↑</th>
<th>PESQ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>1.52*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>99.7</td>
<td>18.56</td>
<td>1.09</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>96.5</td>
<td>56.39</td>
<td>3.72</td>
</tr>
<tr>
<td>0</td>
<td>2000</td>
<td>86.7</td>
<td>57.79</td>
<td>3.61</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>99.2</td>
<td>57.20</td>
<td>4.20</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>93.9</td>
<td>58.00</td>
<td>4.25</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>90.5</td>
<td>59.01</td>
<td>4.28</td>
</tr>
</tbody>
</table>
Experimental Results

Can our proposed model attack the pretrained speaker recognition model?

- Targeted attack
- Average success rate 72.1%
- Average SNR 57.64dB
- Average PESQ 3.48

<table>
<thead>
<tr>
<th>Target ID</th>
<th>PTR(%)↑</th>
<th>SNR(dB)↑</th>
<th>PESQ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>91.4</td>
<td>57.55</td>
<td>3.36</td>
</tr>
<tr>
<td>100</td>
<td>89.3</td>
<td>56.83</td>
<td>3.16</td>
</tr>
<tr>
<td>200</td>
<td>63.3</td>
<td>58.42</td>
<td>3.69</td>
</tr>
<tr>
<td>300</td>
<td>58.7</td>
<td>56.92</td>
<td>3.52</td>
</tr>
<tr>
<td>400</td>
<td>57.6</td>
<td>58.36</td>
<td>3.68</td>
</tr>
<tr>
<td>avg</td>
<td>72.1</td>
<td>57.64</td>
<td>3.48</td>
</tr>
</tbody>
</table>
Experimental Results

Does our design work? (the phoneme recognition model)

✓ With fixed $\lambda_{norm}$, larger $\lambda_{phn}$ results a higher SNR and PESQ

✓ The phoneme branch works for obtaining a trade-off between SER and SNR/PESQ

<table>
<thead>
<tr>
<th>$\lambda_{phn}$</th>
<th>$\lambda_{norm}$</th>
<th>SER(%)↑</th>
<th>SNR(dB)↑</th>
<th>PESQ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>1.52*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>99.7</td>
<td>18.56</td>
<td>1.09</td>
</tr>
<tr>
<td>0</td>
<td>1000</td>
<td>96.5</td>
<td>56.39</td>
<td>3.72</td>
</tr>
<tr>
<td>0</td>
<td>2000</td>
<td>86.7</td>
<td>57.79</td>
<td>3.61</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>99.2</td>
<td>57.20</td>
<td>4.20</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>93.9</td>
<td>58.00</td>
<td>4.25</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>90.5</td>
<td>59.01</td>
<td>4.28</td>
</tr>
</tbody>
</table>
Experimental Results

◆ Other findings

- The perturbations concentrate on high frequency
- Can we design robust speaker recognition models focusing on the low frequency? (future works)
The questions

◆ How to attack the speech-based biometric systems?
  ✔ Our proposed framework successfully attacked the SOTA speaker recognition model

◆ Is the speech-based biometric systems vulnerable to the attacker?
  ✔ Yes

◆ Is it possible to design a biometric systems robust to the adversarial attack?
  ✔ The future works
Thanks

Q & A

Codes, data and more results: https://smallflyingpig.github.io/speaker-recognition-attacker/main