Defending Against Adversarial Attacks in Speaker Verification Systems

Li-Chi Chang, Zesheng Chen, Chao Chen, Guoping Wang, and Zhuming Bi

Purdue University Fort Wayne
Outlines

Motivation
Our Proposed Defense System
Experiments
Conclusions and Future Works
Motivation - Speaker Verification Systems

Speaker verification systems are important to apply human voice as biometrics:
- Accurately identify a legitimate user
- Avoid illegal access

Speaker Verification Systems
- GMM
- I-Vector
- D-Vector
- X-Vector
Motivation
- Attack Against SV Systems

There are many attacks targeted on the speaker verification systems.

Attacks against Speaker Verification Systems
Replay attack
Cloning attack
Adversarial attack
Motivation - Adversarial Attack Against SV Systems

There are many attacks targeted on the speaker verification systems.

Attacks against Speaker Verification Systems
- Replay attack
- Cloning attack
- Adversarial attack

Adversarial attack
- Machine learning or deep learning models
- Most dangerous
- Very difficult to detect and defend
Motivation
- Adversarial Attacks

Attack the weakness of machine learning and deep learning models (Goodfellow, Shlens, and Szegedy ICLR 2015)

High Attack Success Rate (ASR)
Motivation
- FakeBob Attack


One of adversarial attacks on SV systems

~99% ASR

---

Algorithm 3.3 FakeBob Attacks

| Input: an audio signal array, threshold of target SV system |
| Output: an adversarial audio |

Require: Threshold of target SV system $\theta$, Audio signal array $A$, Maximum iteration $m$, Score function $S$, Gradient decent function $f_G$, Clip function $f_c$, Learning rate $lr$, Sign function $f_{sign}$

1: begin
2:    $adver \leftarrow A$
3: for $i = 0; i < m; i + +:$
4:    $score \leftarrow S(adver)$
5: if $score \geq \theta$:
6:    return $adver$
7: end if
8:    $adver \leftarrow f_c(adver - lr \times f_{sign}(f_G(adver)))$
9: end for
10: end
Motivation
- Perturbations
Motivation
- Intuition

Noise-like

- - - - - - - -

Adversarial Sample

Denoise it!

Perturbations

Unique

- - - - - - - -

Attacker’s voice

Adversarial Sample

Noise-add (Distort) it!

Perturbations
Our Proposed Defense System - Goal of Our Approaches

Simple
- Easy to implement
- Compatible with any existing SV system
- Modalized

Light weight
- Low computation load
- Real-time task

Effective
- Greatly increase the adversarial processing time
- Reduce the attack success rate
Our Proposed Defense System
- Defense Systems

Plugin functions
Denoising
Noise-Adding
Our Proposed Defense System - Denoising Plugin Effect

Our Proposed Defense System - Noise-Adding Plugin Effect

(a) $\sigma = 0.001$

Noise-added Audio

Original Audio

Difference Between Noise-added Audio and Original Audio

(a) $\sigma = 0.001$
Experiments - Setup

• Environment
  Google Cloud Platform
  Local GPU server

• SV systems
  GMM
  i-Vector

• Tools
  Kaldi speech recognition toolkit
  Pre-trained models from VoxCeleb 1

• Adversarial Attack
  FakeBob

• Audio dataset
  LibriSpeech
Experiments
- Efficiency Evaluation (Equal Error Rate)

EER = CER = FAR_i = FRR_j, where Threshold(FAR_i) = Threshold(FRR_j)

Crossover Error Rate
False Acceptance Rate
False Rejection Rate

Good Performance, low EER
Bad Performance, high EER
# Experiments

## Normal Operations in GMM

<table>
<thead>
<tr>
<th>Plugin</th>
<th>$\sigma$</th>
<th>EER (%)</th>
<th>Processing Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0</td>
<td>1.05</td>
<td>18.44</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.001</td>
<td>1.61</td>
<td>30.67</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.002</td>
<td>2.95</td>
<td>30.41</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.005</td>
<td>3.36</td>
<td>30.79</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.001</td>
<td>1.21</td>
<td>19.34</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.002</td>
<td>1.92</td>
<td>19.78</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.005</td>
<td>3.94</td>
<td>20.31</td>
</tr>
</tbody>
</table>
## Experiments
- Normal Operations in I-Vector

<table>
<thead>
<tr>
<th>Plugin</th>
<th>$\sigma$</th>
<th>EER (%)</th>
<th>Processing Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0</td>
<td>0</td>
<td>433.45</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.001</td>
<td>0.15</td>
<td>447.37</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.002</td>
<td>0.05</td>
<td>447.82</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.005</td>
<td>0.49</td>
<td>446.20</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.001</td>
<td>0.44</td>
<td>435.35</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.002</td>
<td>0.39</td>
<td>435.89</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.005</td>
<td>1.14</td>
<td>435.51</td>
</tr>
</tbody>
</table>
# Experiments

- Against FakeBob Attacks in GMM

<table>
<thead>
<tr>
<th>Plugin</th>
<th>$\sigma$</th>
<th>Avg Iterations</th>
<th>Avg Time (sec)</th>
<th>Avg ASR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0</td>
<td>23.00</td>
<td>158.68</td>
<td>100.00</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.001</td>
<td>18.90</td>
<td>192.02</td>
<td>77.20</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.002</td>
<td>22.85</td>
<td>235.96</td>
<td>56.05</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.005</td>
<td>22.30</td>
<td>235.78</td>
<td>51.00</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.001</td>
<td>92.6</td>
<td>614.92</td>
<td>24.35</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.002</td>
<td>604.95</td>
<td>3992.88</td>
<td>5.20</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.005</td>
<td>694.95</td>
<td>4350.35</td>
<td>4.10</td>
</tr>
</tbody>
</table>

Max iterations = 1000
Experiments
- Against FakeBob Attacks in I-Vector

<table>
<thead>
<tr>
<th>Plugin</th>
<th>$\sigma$</th>
<th>Avg Iterations</th>
<th>Avg Time (sec)</th>
<th>Avg ASR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0</td>
<td>168.88</td>
<td>6080.47</td>
<td>95.00</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.001</td>
<td>97.40</td>
<td>3702.36</td>
<td>55.68</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.002</td>
<td>100.58</td>
<td>3825.02</td>
<td>38.63</td>
</tr>
<tr>
<td>Denoising</td>
<td>0.005</td>
<td>344.53</td>
<td>13130.24</td>
<td>17.73</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.001</td>
<td>556.33</td>
<td>20041.00</td>
<td>8.98</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.002</td>
<td>918.23</td>
<td>33017.30</td>
<td>0.50</td>
</tr>
<tr>
<td>Noise-Adding</td>
<td>0.005</td>
<td>921.48</td>
<td>33103.39</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Max iterations = 1000
Conclusions

**Simple**
Modalized as a small plugin
Does not need to change the internal structure of an existing SV system

**Light weight**
Low computation load
Minor effect on EER

**Effective**
Reduce the targeted ASR from 100% to 5.2% in GMM and 0.5% in i-vector
Slow down the adversarial attack processing speed 25 times in GMM and 5.43 times in i-vector
Future Works

Future works
X-vector
D-vector
Other type of noise like rustle noise
Thank You