CommanderSong: A Systematic Approach For Practical Adversarial Voice Recognition

Xuejing Yuan¹,², Yuxuan Chen³, Yue Zhao¹,², Yunhui Long⁴, Xiaokang Liu¹,², Kai Chen¹,², Shengzhi Zhang³,⁵, Heqing Huang, XiaoFeng Wang⁶, and Carl A. Gunter⁴

¹SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China  
²School of Cyber Security, University of Chinese Academy of Sciences, China  
³Department of Computer Sciences, Florida Institute of Technology, USA  
⁴Department of Computer Science, University of Illinois at Urbana-Champaign, USA  
⁵Department of Computer Science, Metropolitan College, Boston University, USA  
⁶School of Informatics and Computing, Indiana University Bloomington, USA
Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion
Background

Automatic speech recognition (ASR)
● Traditional attack

● Adversarial sample

STOP + [image] = Speed limit 50

Stop
Outline

- Background
- **Motivation**
- Approach
- Evaluation
- Conclusion
Motivation

Hidden voice command attack: noise-like voice command is abnormal

Dolphin attack: need a proper transmitter

Recent adversarial audio sample: is not effective in the physical world
So can we design an approach that is:
using normal sound to make a physical world attack?

✓ Automatical
✓ Practical
✓ Surreptitious
✓ Spread
✓ Transferable
CommanderSong Attack
CommanderSong Attack

I think the command in the sound wants me to set the temperature to 24 degree.

----IVC device
Challenges Of The Attack

- Human realization
- Influence of the speakers and environment
- Transfer
Outline

- Background
- Motivation
- **Approach**
- Evaluation
- Conclusion
Approach

• step 1: WTA (WAV-To-API) attack
• step 2: WAA (WAV-Air-API) attack

ASR system: Kaldi (open source platform)
Decoding Principle Of Kaldi

Deep neural network (DNN): represents the probability between features and phonemes. Phoneme: the smallest unit composing a word.

Weighted Finite State Transducers (WFST): probability distribution over sequence of words.
Decoding Principle Of Kaldi

(audio)

(observe state)

O₁O₂O₃O₄…

(transference between HMM states)

Feature extraction → Acoustic model → Language model → Open the door.
Decoding Principle Of Kaldi

- **pdf-id**: indicates the probability of every phoneme (column number of the DNN output matrix)
- **transition-id**: uniquely identifies the HMM state transition (a sequence of transition-ids can identify a phoneme)
Example Of Kaldi Decoding Results

Example of the relationship among the phoneme, pdf-id and transition-id.

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>HMM state</th>
<th>Pdf-id</th>
<th>Transition-id</th>
<th>Transition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$eh_B$</td>
<td>0</td>
<td>6383</td>
<td>15985</td>
<td>0→1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15986</td>
<td>0→2</td>
</tr>
<tr>
<td>$eh_B$</td>
<td>1</td>
<td>5760</td>
<td>16189</td>
<td>self-loop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16190</td>
<td>1→2</td>
</tr>
</tbody>
</table>

Transition-ids sequence of the decoding “Echo”.

df-ids sequence: 6383, 5760, 5760, 5760, 5760, 5760, 5760, 5760, 5760, 5760 ……
WTA Attack Approach

Objective function:
\[ m_i = \arg \max a_{i,j}, \quad g(x(t)) = m. \]
WTA Attack Approach

Song audio $x(t)$ → Feature extraction → DNN → HCLG → Lyrics

Crafted song $x'(t)$ → Feature extraction → DNN → Gradient descent → Pdf-id sequence $b$ → DNN → HCLG → Command text

Objective function:

$$m_i = \arg \max a_{i,j}, \quad g(x(t)) = m.$$  
$$\text{argmin} \| g(x(t) + \delta(t)) - b \|_1. \quad \delta(t)$$
WTA Attack Approach

Objective function:

\[ m_i = \arg \max a_{i,j}, \quad g(x(t)) = m. \]

\[ \arg \min ||g(x(t)+\delta(t))-b||_1. \]

\[ \delta(t) \]
WTA Attack samples for the real world attack?
WAA Attack Approach

- Noise model (background noise & electronic noise of speakers) (needs to access to the speaker and receiver)
- Random noise model (easily generate and universally applicable)
Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion
# Evaluation

## WTA attack results

<table>
<thead>
<tr>
<th>Command</th>
<th>Success rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okay google restart phone now.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google flashlight on.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google read mail.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google clear notification.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google airplane mode on.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google turn on wireless hot spot.</td>
<td>100</td>
</tr>
<tr>
<td>Okay google read last sms from boss.</td>
<td>100</td>
</tr>
<tr>
<td>Echo open the front door.</td>
<td>100</td>
</tr>
<tr>
<td>Echo turn off the light.</td>
<td>100</td>
</tr>
</tbody>
</table>
## Evaluation

### WAA attack results

<table>
<thead>
<tr>
<th>Command</th>
<th>Speaker</th>
<th>Success rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Echo ask capital one to make a credit card payment.</td>
<td>JBL speaker</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>ASUS Laptop</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>SENMATE Broadcast</td>
<td>72</td>
</tr>
<tr>
<td>Okay google call one one zero one one nine one two zero.</td>
<td>JBL speaker</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>ASUS Laptop</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>SENMATE Broadcast</td>
<td>70</td>
</tr>
</tbody>
</table>
Evaluation

Human comprehension (a survey on Amazon Mechanical Turk)

- Have you ever heard this original song before?
- Do you think the song is abnormal?
- Where do you think the noise in the abnormal song comes from?
- How many times have you listened before you can recognize the words.
## Evaluation

Human comprehension of the WTA attack samples

<table>
<thead>
<tr>
<th>Music classification</th>
<th>Listened (%)</th>
<th>Abnormal (%)</th>
<th>Recognize Command (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft music</td>
<td>13</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Rock</td>
<td>33</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Popular</td>
<td>32</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>Rap</td>
<td>41</td>
<td>23</td>
<td>0</td>
</tr>
</tbody>
</table>
## Evaluation

### Human comprehension of the WAA attack samples

<table>
<thead>
<tr>
<th>Song name</th>
<th>Listened (%)</th>
<th>Abnormal (%)</th>
<th>Noise-speaker (%)</th>
<th>Noise-song (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did You Need It</td>
<td>15</td>
<td>67</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>Outlaw of Love</td>
<td>11</td>
<td>63</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>The Saltwater Room</td>
<td>27</td>
<td>67</td>
<td>39</td>
<td>3</td>
</tr>
<tr>
<td>Sleepwalker</td>
<td>13</td>
<td>67</td>
<td>41</td>
<td>0</td>
</tr>
<tr>
<td>Under neath</td>
<td>13</td>
<td>68</td>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>Feeling Good</td>
<td>38</td>
<td>59</td>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>19.5</strong></td>
<td><strong>65.2</strong></td>
<td><strong>40</strong></td>
<td><strong>2.2</strong></td>
</tr>
</tbody>
</table>
## Evaluation

### Transferability from Kaldi to iFLYTEK

<table>
<thead>
<tr>
<th>Command</th>
<th>iFLYREC (%)</th>
<th>iFLYTEK Input (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airplane mode on.</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>Open the door.</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Good night.</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Evaluation

Spread and attack iFlytek

In this demo, we suppose the attacker will send the victim one scam email with YouTube link, which will open the video with our sample.
Understanding Of The Attacks

- Kaldi recognize as command
- Human recognize as command
- Human recognize as song

Original song

CommanderSong1 (no noise)

CommanderSong2 (noise++)

CommanderSong3 (noise++++)

Explanation of Kaldi and human recognize of the audios.
Defense

- Audio turbulence defense
- Audio squeezing defense
Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion
Conclusion

- **Practical** adversarial attack automatic speech recognition
- Can be **transferred** to iFlytek
- Can be **spread** through the Internet and radio
- **Surreptitious** to human