Adversarial Music:

Real world audio adversary against wake-word detection systems

Juncheng B. Li, Shuhui Qu, Xinjian Li, Joseph Szurley, Zico Kolter, Florian Metze

Carnegie Mellon University
Bosch Center for Artificial Intelligence
Stanford University
Motivation

Adversarial Attack
not just a problem in vision

Existing audio attacks against Automatic Speech Recognition systems
not robust over-the-air

Li et al. [2019]

Sample adversarial noise
Schönherr et al. [2019]

Environment noise at home
Fish tank + clock
Two Big Challenges

The actual Alexa model is a black box

Unstructured adversarial noises are not robust in practice
Contributions

Gray-box over-the-air Denial of Service (Dos) Attack against commercial voice assistant

- A “gray-box” attack that leverages the domain transferability of our perturbation. We demonstrated its effect in the real world under separate audio source settings.

- A novel threat model that allows us to disguise our adversarial attack as a piece of music with tunable parameters playable over the air in the physical space.

- Jointly optimizing the attack nature while fitting the threat model to the perturbation achievable by the microphone hearing response of Amazon Alexa. Our attack budget is very limited compared with previous works, which makes this challenging.
“Grey Box” Attack

Figure 1: Emulated Model Architecture based on Panchapagesan et al. [2016], Kumatani et al. [2017], Guo et al. [2018]

Figure 2: Detection Error Tradeoff Curve. The curve of Alexa model is shown in a flat line as its false alarm rate is not published
Adversarial Music Generation using Physical Modeling Synthesizer

Physical Modeling Synthesizer

Key

Duration

Volume

Karplus Strong Generator

Initialize White Noise

Iteratively decay average

\(\theta_{\text{Duration}}\), \(\theta_{\text{Key}}\), \(\theta_{\text{Volume}}\)

\(\delta_\theta\)


Unstructured noise

\(\delta\)
Combat Distortion with Limited Attack Budget

Psychoacoustic Effect

Final Loss: \[ \max_{l(x, \delta, y)} = \mathbf{E}_{t \in \mathcal{T}, x, y \sim D}[L_{\text{wake}}(f(t(x + \delta), y)) - \alpha \cdot L_{\eta}(x, \delta)] \]
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Digital/Physical</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th># of Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/o Attack</td>
<td>Attack</td>
<td>w/o Attack</td>
<td>Attack</td>
<td>w/o Attack</td>
</tr>
<tr>
<td>Emulated Model</td>
<td>Digital</td>
<td>0.97</td>
<td>0.14</td>
<td>0.94</td>
<td>0.11</td>
</tr>
<tr>
<td>Emulated Model</td>
<td>Physical</td>
<td>0.96</td>
<td>0.12</td>
<td>0.91</td>
<td>0.09</td>
</tr>
<tr>
<td>Alexa</td>
<td>Physical</td>
<td>0.93</td>
<td>0.11</td>
<td>0.92</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Table 1. Performance of the models with and without attacks in digital and physical testing environments given the number of testing samples.
Over-the-air Experiment Setup

Over-the-air testing illustration

Spectrogram of the generated adversarial music

Adversarial Music Source

Tester
Over-the-air Evaluation

<table>
<thead>
<tr>
<th>Test Against Alexa</th>
<th>$\phi = 0^\circ$</th>
<th>$d_i =$</th>
<th>$\phi = 90^\circ$</th>
<th>$d_i =$</th>
<th>$\phi = 180^\circ$</th>
<th>$d_i =$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_a =$</td>
<td>Volume</td>
<td>4.2 ft</td>
<td>7.2 ft</td>
<td>10.2 ft</td>
<td>4.2 ft</td>
<td>7.2 ft</td>
</tr>
<tr>
<td>4.7 ft</td>
<td>70 dbA</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>6.2 ft</td>
<td>70 dbA</td>
<td>1/10</td>
<td>0/10</td>
<td>0/10</td>
<td>1/10</td>
<td>0/10</td>
</tr>
<tr>
<td>7.7 ft</td>
<td>70 dbA</td>
<td>2/10</td>
<td>0/10</td>
<td>0/10</td>
<td>3/10</td>
<td>1/10</td>
</tr>
<tr>
<td>4.7 ft</td>
<td>60 dbA</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>6.2 ft</td>
<td>60 dbA</td>
<td>1/10</td>
<td>1/10</td>
<td>0/10</td>
<td>3/10</td>
<td>1/10</td>
</tr>
<tr>
<td>7.7 ft</td>
<td>60 dbA</td>
<td>2/10</td>
<td>1/10</td>
<td>0/10</td>
<td>3/10</td>
<td>2/10</td>
</tr>
<tr>
<td>4.7 ft</td>
<td>50 dbA</td>
<td>1/10</td>
<td>2/10</td>
<td>1/10</td>
<td>2/10</td>
<td>2/10</td>
</tr>
<tr>
<td>6.2 ft</td>
<td>50 dbA</td>
<td>2/10</td>
<td>3/10</td>
<td>2/10</td>
<td>3/10</td>
<td>3/10</td>
</tr>
<tr>
<td>7.7 ft</td>
<td>50 dbA</td>
<td>2/10</td>
<td>3/10</td>
<td>2/10</td>
<td>3/10</td>
<td>2/10</td>
</tr>
</tbody>
</table>
We tested our Adversarial Music against Amazon Echo’s wake word detection: “Alexa” in a normal household environment. In this case, the tester is standing 7.2 ft away from the Amazon Echo.
Thank you!

See you
on Thursday, Dec 12th 10:45-12:45
at East Exhibition Hall B + C #10

Adversarial Music:
Real world audio adversary against wake-word detection systems

Juncheng B. Li, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, Florian Metze