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

All existing physical attacks against ASR are not robust

Li et al. [2019],

Sample adversarial noise
Schönherr et al. [2019]

Environment noise at home
Fish tank + clock
2 Big Challenges

The model behind Alexa is a black box to us.

Unstructured noise is hard to control and is not robust.
This Paper: a physical perturbation works against a commercial grade voice assistant

- We create a **parametric threat model** in audio domain that allows us to disguise our adversarial attack as a **piece of music** playable over the air in the physical space.

- Our adversarial attack is 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, which is the first real-time “gray-box” adversarial.

- Our adversarial attack is 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

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 Synthesizer

Figure 3: String instruments with the one-zero low-pass filter approximation. The synthesis process first generates a short excitation $D$-length waveform. It is then fed into the filter iteratively to generate the sound.

Algorithm 1: Karplus-Strong algorithm

Simulate the plucking phase of each note $i$ by initialize $y_i[0 : D - 1] \sim \mathcal{N}(0, \beta)$;

for $i = 1, \ldots, L$ do

for $n = D, \ldots, d_i$ do

\[ y_i[n] = \gamma v_{\text{output}} (\omega \times y_i[n - D] + (1 - \omega) \times y_i[n - D - 1]) / 2; \]

Return $y$;
Psychoacoustic Effect and Room Impulse Response

\[ \tilde{p}_\delta(k) = 92 - \max_k p_x(k) + p_\delta(k) \]

where \( p_\delta(k) = 10 \log_{10} \left| \frac{1}{N} s_\delta(k) \right|^2 \), \( p_x(k) = 10 \log_{10} \left| \frac{1}{N} s_x(k) \right|^2 \)

are power spectral density estimation of the perturbation and the original audio input. \( s_x(k) \) is the \( k \)th bin of the spectrum of frame \( x \).

\[
L_\eta(x, \delta) = \frac{1}{|N/2| + 1} \sum_{k=0}^{|N/2|} \max \{ \tilde{p}_\delta(k) - \eta_x(k), 0 \}
\]

where \( N \) is the predefined window size and \([x]\) outputs the greatest integer no larger than \( x \).

Final Loss:

\[
\max L(x, \delta_\theta, y) = \mathbb{E}_{t \in \mathcal{T}} \left[ L_{WW} f(t(x + \delta_\theta)) - \alpha \cdot L_\eta(x, \delta) \right]
\]
Experiment

Figure 4: Physical Testing Illustration

Figure 5: A sample adversarial music
## Results

<table>
<thead>
<tr>
<th>Models</th>
<th>Attack</th>
<th>Digital / Physical</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emulated Model</td>
<td>No</td>
<td>Digital</td>
<td>0.97</td>
<td>0.94</td>
<td>0.955</td>
<td>4000</td>
</tr>
<tr>
<td>Emulated Model</td>
<td>No</td>
<td>Physical</td>
<td>0.96</td>
<td>0.91</td>
<td>0.934</td>
<td>100</td>
</tr>
<tr>
<td>Alexa</td>
<td>No</td>
<td>Physical</td>
<td>0.93</td>
<td>0.92</td>
<td>0.925</td>
<td>100</td>
</tr>
<tr>
<td>Emulated Model</td>
<td>Yes</td>
<td>Digital</td>
<td>0.14</td>
<td>0.11</td>
<td>0.117</td>
<td>4000</td>
</tr>
<tr>
<td>Emulated Model</td>
<td>Yes</td>
<td>Physical</td>
<td>0.12</td>
<td>0.09</td>
<td>0.110</td>
<td>100</td>
</tr>
<tr>
<td>Alexa</td>
<td>Yes</td>
<td>Physical</td>
<td>0.11</td>
<td>0.10</td>
<td>0.110</td>
<td>100</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

<table>
<thead>
<tr>
<th>Test Against Alexa</th>
<th>$\phi = 0^\circ$</th>
<th>$\phi = 90^\circ$</th>
<th>$\phi = 180^\circ$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_t =$</td>
<td>4.2 ft</td>
<td>7.2 ft</td>
<td>10.2 ft</td>
</tr>
<tr>
<td>$d_a = 4.7 ft, 70 dB$</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>$d_a = 6.2 ft, 70 dB$</td>
<td>1/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>$d_a = 7.7 ft, 70 dB$</td>
<td>2/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>$d_a = 4.7 ft, 60 dB$</td>
<td>0/10</td>
<td>0/10</td>
<td>0/10</td>
</tr>
<tr>
<td>$d_a = 6.2 ft, 60 dB$</td>
<td>1/10</td>
<td>1/10</td>
<td>0/10</td>
</tr>
<tr>
<td>$d_a = 7.7 ft, 60 dB$</td>
<td>2/10</td>
<td>1/10</td>
<td>0/10</td>
</tr>
</tbody>
</table>

Table 2: Times of the real Amazon Alexa being able to respond to the wake-word under the influence of our adversarial music with different settings. (The female and male tester each tests 5 utterances.) The testing setup is illustrated in Figure 4, and it is also shown in the demo video. In this paper, dB indicates dB SPL.
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 7.2ft 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