Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems

Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R.B. Butler, and Joseph Wilson
Voice as an Interface
Injecting Commands
You might be thinking...

- **Benign?**
- **Easy to Defend?**
Is there a generic, transferrable way to produce audio that:

- sounds like noise to humans,
- sounds like a valid command to the system?
- works against both speech and speaker recognition systems
- with Black-Box access to target system
Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference
Modern Speech Recognition Systems

Audio Sample → Preprocessing → Feature Extraction → Inference

Most Attacks* → Our Attack → MODEL DOES NOT MATTER!!

Our Attack
Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway
• Designed to **approximate** the human ear
• **Retains** the most important **features**
• Magnitude Fast Fourier Transform (**mFFT**)
- Converts time domain to frequency domain
- Multiple Inputs can have same output
- mFFT is lossy
Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway
How the Attack Works (Types)

- Grouped into 4 types

**Original Audio**

**Perturbed Audio**
How the Attack Works (Psychoacoustics)

- Intelligibility hard to measure
- Fundamentals of psychoacoustics
- Spread energy across spectrum
## How the Attack Works (Evaluation)

<table>
<thead>
<tr>
<th>Task</th>
<th>Speech</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise -&gt; text</td>
<td>22 speakers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>Speech</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 20,000 successful attack samples</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Queries</th>
<th>Speech</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;10 queries to model (a few seconds!)</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>Speech</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>KALDI</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>HOUNDIFY</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>Intel</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>CMU Sphinx</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>Microsoft Azure</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>Bing</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>Firefox</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
<tr>
<td>wit.ai</td>
<td>Noise -&gt; user</td>
<td></td>
</tr>
</tbody>
</table>
Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway
More at: https://sites.google.com/view/practicalhiddenvoice/home
More at:
https://sites.google.com/view/practicalhiddenvoice/home
More at:

https://sites.google.com/view/practicalhiddenvoice/home
Modern Speech Recognition Systems

Feature Extraction

How the Attack Works

Demo

Takeaway
Simple, efficient audio transformations yield “noise” that is understood as commands by speech systems.

The model is irrelevant.

All systems we tested are vulnerable.

Achieve the same goals as traditional Adversarial ML.

---

Project webpage: sites.google.com/view/practicalhiddenvoice/home

hadi10102@ufl.edu
hadiabdullah.github.io
hadiabdullah1
- Easy to get around
- Artificially generate a target’s speech
  - LyreBird
- Capture and stitch together target’s speech
Defenses

- Must be implemented at or before feature extraction
- Adversarial Training?
- Voice Activity Detection?
- Environmental Noise
- Liveness Detection
  - Blue et al.*