CommanderGabble: A Universal Attack Against ASR Systems Leveraging Fast Speech

Zhaohe Zhang, Edwin Yang, Song Fang
University of Oklahoma
ACSAC 2021
Background

- Automatic Speech Recognition (ASR) systems are widely available; their accuracy has been greatly improved over time.

- However, ASR misinterpretations still happen frequently in practice.
Existing Attacks on ASRs

• According to the knowledge available for an attacker:

- If specialized hardware is available:

  - Input audio → Audio modification → Adversarial audio → ASR

  - Low: No specifics about the victim system
  - High: Internal structure of the target system

  - The attacker’s knowledge

• If specialized hardware is available:

  - Transmitter hardware → Inaudible signal → ASR

  - Ultrasonic speaker

  - Microphone circuits
  - Processor

  - Nonlinearity: high frequency → low frequency
**Existing Attacks on ASRs (contd.)**

- According to how adversary audio is delivered to ASR:

  - **Over-the-wire**
    - Audio file (WAV, FLAC) → Model → ASR
    - Audio is directly passed to the target ASR.
    - Environmental factors (e.g., noise) have **no impact**.

  - **Over-the-air**
    - Adversarial audio → Microphone → ASR
    - Audio is played via a speaker towards the target ASR.
    - Environmental factors **matter**.
Phoneme VS. Syllable

• What are phonemes?
  ✓ The smallest units of sound which can distinguish two words, e.g., /k/ and /b/ → ‘cat’ vs. ‘bat’ => two different words
  ✓ Classification
    ➢ Vowel vs. consonant

• What is a syllable?
  ✓ A single, unbroken sound within a spoken or written word, e.g., ‘cat’ vs. ‘water’ => 1 syllable vs. 2 syllables

<table>
<thead>
<tr>
<th>Syllable Structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>I</td>
</tr>
<tr>
<td>CV</td>
<td>me, see</td>
</tr>
<tr>
<td>VC</td>
<td>up, in</td>
</tr>
<tr>
<td>CVC</td>
<td>cat, map</td>
</tr>
<tr>
<td>CCV</td>
<td>try, sly</td>
</tr>
<tr>
<td>CCVC</td>
<td>slip</td>
</tr>
</tbody>
</table>
Motivation

- Impact of fast speech

What if we carefully manipulating the phonetic structure of a target voice command?
Attack Scenario

- ASR
- Playing news
What is happening?
Who's there?!
Types of Misinterpretation

• An example command: “Open the door”

Original command

“Open the door”

Original phonemes

[OW P AH N] [DH AH] [D AO R]

Recognized command

“Oh panda our”

Recognized phonemes

[OW] [P AE N D AH] [AW ER]

✓ Reduction: some phonemes are omitted;
Types of Misinterpretation (contd.)

- **Reduction**: some phonemes are omitted;
- **Replacement**: some phonemes are replaced with similar phonemes;
Types of Misinterpretation (contd.)

- Reduction: some phonemes are omitted;
- Replacement: some phonemes are replaced with similar phonemes;
- Coalescence: some neighboring phonemes are merged together.

Original command: “Open the door”
Recognized command: “Oh panda our”

Original phonemes: [OW P AH N] [DH AH] [D AO R]
Recognized phonemes: [OW] [P AE N D AH] [AW ER]
Phonetic reconstruction
- Extract syllables from target command’s phonetic representation.
- Map each word to a new word to generate an adversarial command.

Speech synthesis
- Generate fast speech of the adversarial command.

Winnowing and refining
- Verify incomprehensibility and effectiveness.
- Update syllabification rules.
Phonetic Reconstruction

✓ Phoneme syllabification: [C C V C C V C C]
✓ Phoneme morpher: [R AO D K AE S]
✓ Phonemes-to-word translation: ‘Rode Cass’
Speech Synthesis

- Generate adversarial audio of a candidate command.
  ✓ Utilize Google Cloud TTS

- Achieve fast speech by controlling playback speed (2.0x - 3.0x).
  ✓ Normal speed (≈ 1.0x): Easy to understood by human
  ✓ Too fast (> 3.0x): ASR fails to recognize due to excessive distortion

- Generated audio is transmitted to target ASR according to attack scenario
Winnowing and Refining

• Winnow out ineffective candidate adversarial audio.
  ✓ Intelligibility check

  ![Diagram](image)

  Adversarial Audio → Human

  ![Diagram](image)

  Adversarial Audio → ASR → Success?

✓ Execution check

• Syllabification modifier
  ❖ If either check fails, the adversary modifies syllabification rules correspondingly.

  ![Diagram](image)

  Target Command → Phonetic Reconstruction → Speech Synthesis → Winnowing and Refining → Adversarial Audio
Evaluation Setup

• Over-the-wire attack
  ✓ Select 100 ASR commands

[Images of Amazon Transcribe, IBM Watson, Google Cloud, and Microsoft Azure]

• Over-the-air attack
  ✓ 6 commands for each environment

<table>
<thead>
<tr>
<th>Environment</th>
<th>ID</th>
<th>Command</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>C1</td>
<td>Stop</td>
</tr>
<tr>
<td></td>
<td>C2</td>
<td>Continue</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td>Unlock the door</td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>Call my phone</td>
</tr>
<tr>
<td></td>
<td>C5</td>
<td>Show me the back door camera</td>
</tr>
<tr>
<td></td>
<td>C6</td>
<td>Turn off the light in living room</td>
</tr>
<tr>
<td></td>
<td>C7</td>
<td>Bluetooth</td>
</tr>
<tr>
<td></td>
<td>C8</td>
<td>Location</td>
</tr>
<tr>
<td></td>
<td>C9</td>
<td>Call my phone</td>
</tr>
<tr>
<td></td>
<td>C10</td>
<td>Recent messages</td>
</tr>
<tr>
<td></td>
<td>C11</td>
<td>Turn on the light</td>
</tr>
<tr>
<td></td>
<td>C12</td>
<td>Set the alarm at 3am</td>
</tr>
<tr>
<td></td>
<td>C13</td>
<td>News</td>
</tr>
<tr>
<td></td>
<td>C14</td>
<td>Home</td>
</tr>
<tr>
<td></td>
<td>C15</td>
<td>Enable Tollway</td>
</tr>
<tr>
<td></td>
<td>C16</td>
<td>Cancel Route</td>
</tr>
<tr>
<td></td>
<td>C17</td>
<td>How long will it take to drive to library</td>
</tr>
<tr>
<td></td>
<td>C18</td>
<td>What is my current location</td>
</tr>
</tbody>
</table>
Over-the-wire Translation Accuracy

- Most of adversarial audios are correctly recognized.
- Highest accuracy (95%) for medium length commands.
- Low accuracy (28%) for normal commands.

<table>
<thead>
<tr>
<th>Command Length</th>
<th>Number of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>2-3</td>
</tr>
<tr>
<td>Long</td>
<td>&gt;3</td>
</tr>
</tbody>
</table>

OTW Translation accuracy for fast speech audio files
Over-the-air Attack Success Rate

- Target ASRs

- Adversarial wake-up word test

<table>
<thead>
<tr>
<th>Wake-up Word</th>
<th>Adversarial Command</th>
<th>Playback Speed</th>
<th>Successful?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ok Google</td>
<td>kaye go oh</td>
<td>2.0x-2.1x</td>
<td>✓</td>
</tr>
<tr>
<td>Alexa</td>
<td>a leh sa</td>
<td>2.0x-2.1x</td>
<td>✓</td>
</tr>
<tr>
<td>Hey Cortana</td>
<td>hye core ta</td>
<td>2.0x-2.1x</td>
<td>✓</td>
</tr>
</tbody>
</table>

❖ All wake-up words are correctly recognized by target ASRs.
Over-the-air Attack Success Rate (contd.)

Attack performance on different ASRs

<table>
<thead>
<tr>
<th>Command ID</th>
<th>Amazon Alexa</th>
<th>Google Assistant</th>
<th>Microsoft Cortana</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C2</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C3</td>
<td>7/10</td>
<td>8/10</td>
<td>8/10</td>
</tr>
<tr>
<td>C4</td>
<td>10/10</td>
<td>10/10</td>
<td>9/10</td>
</tr>
<tr>
<td>C5</td>
<td>10/10</td>
<td>10/10</td>
<td>9/10</td>
</tr>
<tr>
<td>C6</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C7</td>
<td>8/10</td>
<td>9/10</td>
<td>7/10</td>
</tr>
<tr>
<td>C8</td>
<td>9/10</td>
<td>8/10</td>
<td>8/10</td>
</tr>
<tr>
<td>C9</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C10</td>
<td>8/10</td>
<td>9/10</td>
<td>9/10</td>
</tr>
<tr>
<td>C11</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C12</td>
<td>10/10</td>
<td>10/10</td>
<td>10/10</td>
</tr>
<tr>
<td>C13</td>
<td>5/10</td>
<td>6/10</td>
<td>5/10</td>
</tr>
<tr>
<td>C14</td>
<td>6/10</td>
<td>6/10</td>
<td>5/10</td>
</tr>
<tr>
<td>C15</td>
<td>6/10</td>
<td>8/10</td>
<td>4/10</td>
</tr>
<tr>
<td>C16</td>
<td>8/10</td>
<td>8/10</td>
<td>-</td>
</tr>
<tr>
<td>C17</td>
<td>8/10</td>
<td>8/10</td>
<td>6/10</td>
</tr>
<tr>
<td>C18</td>
<td>9/10</td>
<td>9/10</td>
<td>7/10</td>
</tr>
</tbody>
</table>

- Human comprehensibility test
  - Recruited 28 volunteers
  - None could comprehend any adversarial audio

✓ Average success rates for three ASRs:
  - Home: 95%, 97%, 93%
  - Teleconference: 92%, 93%, 90%
  - Noisy environment results decreased success rates.
Conclusion

✓ We systematically explore misinterpretations introduced by fast speech and analyze the consequent phonetic structure variations.

✓ By combining phoneme manipulation with fast speech, we develop CommanderGabble for a model-agnostic and easily-constructed adversarial attack against ASR systems.

✓ We perform extensive experiments to evaluate feasibility robustness, and suspiciousness of CommanderGabble.
Thank you!
Any questions?

commandergabble.info