Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems

Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, Feng Qian
Voice Assistant Devices

- Alexa, play Today’s Hits on Pandora
- Alexa, turn on Living Room lights
- Alexa, ask PayPal to send 10 dollars to Sam
- Alexa, ask Medical Assistant to give me my diagnosis
Smart Enough to be Secure?
Not Yet
How it works?

Voice assistants work like a relay, proxying and translating conversation between users and skills.

- Alexa, play Today’s Hits on Pandora
- Alexa, turn on Living Room lights
- Alexa, ask PayPal to send 10 dollars to Sam

User -> Smart Speaker -> Voice Assistant Cloud -> Third-party Skill Clouds
Security requirements and gaps

Route the source payload to the **CORRECT** destination

Voice Assistant Platforms
## Security requirements and gaps

<table>
<thead>
<tr>
<th>Requirements for Reliable Payload Routing</th>
<th>Network Routing System</th>
<th>Voice Assistant Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destinations should be assigned with addresses</td>
<td>IP addresses</td>
<td>Skill Invocation Names in text forms</td>
</tr>
<tr>
<td>Different destinations should have unique addresses</td>
<td>Different network hosts are with different IP addresses</td>
<td>Alexa allows skills to have same invocation names</td>
</tr>
<tr>
<td>The traffic should embed the destination address</td>
<td>Each IP packet has dest IP address as the header field</td>
<td>Users are not machines &amp; natural language is diverse</td>
</tr>
<tr>
<td>The routing system should correctly retrieve destination address</td>
<td>Well-defined IP packet format</td>
<td>Complicated AI systems</td>
</tr>
<tr>
<td>Conflicting Paths</td>
<td>Longest prefix matching</td>
<td>Longest prefix matching</td>
</tr>
</tbody>
</table>
Voice Squatting

Voice assistants may fail to understand user’s intention, and mistakenly invoke wrong skills.

Example:
Alexa, ask PayPal to send 10 dollars to Sam
Voice Masquerading

Skill switching is not well supported, allowing a skill to masquerade itself as other skills or even the system.

User

Smart Speaker

Voice Assistant Cloud

Third-party Skill Clouds

Alexa, open PayPal please

Yes, I am PayPal, give me your credentials
Potential Consequences of Voice Squatting

- Compromise of user's sensitive data or devices
- Propagate fake or controversial information
- Traditional Phishing
- Compromise reputation of the victim skill

- Money, historical transactions, bank accounts
- Access to home devices
Potential Consequences of Voice Squatting

- Compromise of user's sensitive data or devices
- Traditional Phishing
- Propagate fake or controversial information
- Compromise reputation of the victim skill

President Trump didn’t twitter last week

We regret to tell you our diagnosis shows that XX
Potential Consequences of Voice Squatting

- Compromise of user's sensitive data or devices
- Propagate fake or controversial information
- Traditional Phishing
- Compromise reputation of the victim's skill

Account Closed

Capital One

You account is locked due to suspicious activity. Please contact fraud department immediately at (800) XXX-XXXX to activate your account.
Potential Consequences of Voice Squatting

- Compromise of user's sensitive data or devices
- Traditional Phishing
- Propagate fake or controversial information
- Compromise reputation of the victim skill
Potential Consequences of Voice Masquerading

- Fake Skill Switching
- Fake Skill Termination

Same consequences as the voice squatting
Potential Consequences of Voice Masquerading

- Fake Skill Switching
- Fake Skill Termination

- Record user’s conversations
- Skill recommendation
How realistic are those attacks?

- Study how users invoke skills
- Study how well the platforms can understand voice commands
- Experiment proof-of-concept attack skills
- Identify real-world attacks
How realistic are those attacks?

Study how users invoke skills

Study how well the platforms can understand voice commands

Identify real-world attacks

Experiment proof-of-concept attack skills
How realistic are those attacks?

- “Sleep Sounds”, “Cat Facts”
- Multi-choice questions combined with open questions

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Amazon</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, “open Sleep Sounds please”</td>
<td>64%</td>
<td>55%</td>
</tr>
<tr>
<td>Yes, “open Sleep Sounds for me”</td>
<td>30%</td>
<td>25%</td>
</tr>
<tr>
<td>Yes, “open Sleep Sounds app”</td>
<td>26%</td>
<td>20%</td>
</tr>
<tr>
<td>Yes, “open my Sleep Sounds”</td>
<td>29%</td>
<td>20%</td>
</tr>
<tr>
<td>Yes, “open the Sleep Sounds”</td>
<td>20%</td>
<td>14%</td>
</tr>
<tr>
<td>Yes, “play some Sleep Sounds”</td>
<td>42%</td>
<td>35%</td>
</tr>
<tr>
<td>Yes, “tell me a Cat Facts”</td>
<td>36%</td>
<td>24%</td>
</tr>
</tbody>
</table>

When invoking skills, Users tend to use diverse and natural-language utterances

Longest prefix matching creates attack space for voice squatting

Users’ preference when invoking skills
How realistic are those attacks?

- Study how users invoke skills
- Study how well the platforms can understand voice commands
- Experiment proof-of-concept attack skills
- Identify real-world attacks
How realistic are those attacks?

- Invocation Names → Record → Voice Recordings → Play → Voice Assistant Platforms → Recognition → Helper Skill

- 100 invocation names for each platform
- Human subjects & TTS services

Those voice assistant platforms are error-prone when recognizing voice commands

### Recognition Mistake Rates

<table>
<thead>
<tr>
<th></th>
<th>TTS services</th>
<th>Human subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa</td>
<td>30%</td>
<td>57%</td>
</tr>
<tr>
<td>Google</td>
<td>9%</td>
<td>10%</td>
</tr>
</tbody>
</table>

- Florid state quiz: ✔️
- Florid snake quiz: ✗
- Rent Europe: ✔️
- Read your app: ✗
How realistic are those attacks?

- Study how users invoke skills
- Study how well the platforms can understand voice commands
- Identify real-world attacks
- Experiment proof-of-concept attack skills
How realistic are those attacks?

- Compose attacks skills
- Register attacks skills
- Generate and record voice commands
- Play voice commands and decide whether attack skills get invoked

Voice Squatting through invocation name extending:
- Capital One
- Capital One Please
- My Capital One
- Capital One App

Voice Squatting through similar pronunciation:
- Capital One
- Capital Won
- Captain One
- Capitol One

Attack skills were not published to the skill market.
How realistic are those attacks?

**Voice Squatting through invocation name extending**

<table>
<thead>
<tr>
<th>Alexa</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>invocation name + “please”</td>
<td>10/10</td>
</tr>
<tr>
<td>“my” + invocation name</td>
<td>7/10</td>
</tr>
<tr>
<td>“the” + invocation name</td>
<td>10/10</td>
</tr>
<tr>
<td>invocation name + “app”</td>
<td>10/10</td>
</tr>
<tr>
<td>“mai” + invocation name</td>
<td>-</td>
</tr>
<tr>
<td>invocation name + “plese”</td>
<td>-</td>
</tr>
</tbody>
</table>

**Voice Squatting through similar pronunciation**

<table>
<thead>
<tr>
<th>Alexa</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon TTS</td>
<td>Google TTS</td>
</tr>
<tr>
<td>10/17</td>
<td>12/17</td>
</tr>
</tbody>
</table>
How realistic are those attacks?

Study how users invoke skills

Study how well the platforms can understand voice commands

Identify real-world attacks

Experiment proof-of-concept attack skills
How realistic are those attacks?

Identify Skills with Competing Invocation Names (CIN)

1. Collect Available Skills
2. Generate CINs for each invocation name
3. Identify Competing Skills

Alexa: 19,670
Google: 1001

Invocation names on the market

Invocation name → Text Paraphrasing → Pronunciation comparison → CINs on the market
Real-World Attack Measurement

Invocation name → Text Paraphrasing → Pronunciation comparison → CINs on the market

Capital One ➔ Capital One
Capital One ➔ Capital One please
Capital One app ➔ The Capital One
Capital One ➔ ...
Capital One ➔ ...
Capital One ➔ Captain One
Captain One ➔ ...
Captain One ➔ ...
Capital One ➔ Captain One
Real-World Attack Measurement

19% (3718) skills: same pronunciation

66 skills were named as “cat facts”, and provided similar functions.

2.7% (531) skills: same pronunciation, but different spelling

1.8% (345) skills: longest prefix matching

Interesting cases

dog fact → search me a dog fact

“SCUBA Diving Trivia” Skill and “Soccer Geek” skill, registered “space geek” as invocation names
Defense

UIC: User Intention Classifier
- Classify user's intention as context switching or not

SRC: Skill Response Checker
- Identify suspicious skill response, such as fake skill recommendation

Guess Game Web Service
Defense

User Request → Sentence Embedding → Classifier

- System commands
- Invocation name of other skills

For current skill

For context switch

Request and response of same session
Skill Description

User Intention Classifier (UIC)
Defense

Skill Response Checker (SRC)

- Skill Response
- Sentence Embedding
- Black List
  - System Response
  - Empty Response
  - ......
Summary

- Two attack scenarios: Voice Squatting & Voice Masquerading
- Both attacks were found to be practical, and dangerous
- We explored a set of mitigation solution: CIN generator, User Intention Classifier, and Skill Response Checker.
- Both platform vendors acknowledged our attacks, and discussed the mitigation solutions.
Attack Demos: https://sites.google.com/site/voicevpasec/
How realistic are those attacks?

What would you say when invoking a skill?

Have you ever invoked a wrong skill?

Did you try context switch when talking to a skill?

Have you experienced any problem closing a skill?

How do you know whether a skill has terminated?

Recruit participants on Amazon Mechanical Turk

Filter out invalid response

105 valid responses from Amazon Echo users and 51 valid responses from Google Home users
How realistic are those attacks?

- “Sleep Sounds”, “Cat Facts”
- Multi-choice questions combined with open questions

Users tend to use diverse and natural-language utterances

<table>
<thead>
<tr>
<th>Question</th>
<th>Amazon</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, “open Sleep Sounds please”</td>
<td>64%</td>
<td>55%</td>
</tr>
<tr>
<td>Yes, “open Sleep Sounds for me”</td>
<td>30%</td>
<td>25%</td>
</tr>
<tr>
<td>Yes, “open Sleep Sounds app”</td>
<td>26%</td>
<td>20%</td>
</tr>
<tr>
<td>Yes, “open my Sleep Sounds”</td>
<td>29%</td>
<td>20%</td>
</tr>
<tr>
<td>Yes, “open the Sleep Sounds”</td>
<td>20%</td>
<td>14%</td>
</tr>
<tr>
<td>Yes, “play some Sleep Sounds”</td>
<td>42%</td>
<td>35%</td>
</tr>
<tr>
<td>Yes, “tell me a Cat Facts”</td>
<td>36%</td>
<td>24%</td>
</tr>
</tbody>
</table>

• Multi-choice questions combined with open questions

- “Sleep Sounds”, “Cat Facts”
- Users tend to use diverse and natural-language utterances
How realistic are those attacks?

<table>
<thead>
<tr>
<th>Question</th>
<th>Amazon</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invoked a wrong skill</td>
<td>29%</td>
<td>27%</td>
</tr>
<tr>
<td>Tried to switch to another skill</td>
<td>26%</td>
<td>24%</td>
</tr>
<tr>
<td>Failed to quit a skill</td>
<td>30%</td>
<td>29%</td>
</tr>
</tbody>
</table>

**Interaction context switching is not well supported**

**Longest prefix matching creates attack space for voice squatting**
How realistic are those attacks?

- Select skills
- Generate and record voice commands
- Play voice commands and get recognition results

100 skills per platform

Invocation name, open + invocation name

TextToSpeech Services & Human subjects

<table>
<thead>
<tr>
<th></th>
<th>Invocation Name</th>
<th>Open + Invocation Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon TTS</td>
<td>5 x 100</td>
<td>5 x 100</td>
</tr>
<tr>
<td>Google TTS</td>
<td>5 x 100</td>
<td>5 x 100</td>
</tr>
<tr>
<td>Human Subject</td>
<td>-</td>
<td>2 x 100</td>
</tr>
</tbody>
</table>
How realistic are those attacks?

Those Voice assistant platforms are error-prone when recognizing voice commands.

- **Amazon TTS**: 232/500 (Pronounce invocation name only), 62/100 (Pronounce “Open” + Invocation Name)
- **Google TTS**: 164/500 (Invocation name only), 41/100 (Pronounce “Open” + Invocation Name)
- **Human (Avg)**: 115/500 (Invocation name only), 69/100 (Pronounce “Open” + Invocation Name)

<table>
<thead>
<tr>
<th>VPA</th>
<th>Source</th>
<th># of misrecognized utterances</th>
<th># of misrecognized skills</th>
<th># of misrecognized utterances</th>
<th># of misrecognized skills</th>
<th># of skills misrecognized every time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa</td>
<td>Amazon TTS</td>
<td>232/500</td>
<td>62/100</td>
<td>125/500</td>
<td>33/100</td>
<td>17/100</td>
</tr>
<tr>
<td></td>
<td>Google TTS</td>
<td>164/500</td>
<td>41/100</td>
<td>104/500</td>
<td>26/100</td>
<td>17/100</td>
</tr>
<tr>
<td></td>
<td>Human (Avg)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>45/100</td>
</tr>
<tr>
<td>Google</td>
<td>Amazon TTS</td>
<td>96/500</td>
<td>24/100</td>
<td>42/500</td>
<td>12/100</td>
<td>7/100</td>
</tr>
<tr>
<td></td>
<td>Google TTS</td>
<td>62/500</td>
<td>19/100</td>
<td>26/500</td>
<td>6/100</td>
<td>4/100</td>
</tr>
<tr>
<td></td>
<td>Human (Avg)</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>6/100</td>
</tr>
</tbody>
</table>
How realistic are those attacks?

<table>
<thead>
<tr>
<th>Attack Skill</th>
<th>Skill Name</th>
<th>Invocation Name</th>
<th>Victim Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>Smart Gap</td>
<td>smart gap</td>
<td>smart cap</td>
</tr>
<tr>
<td></td>
<td>Soothing Sleep Sounds</td>
<td>sleep sounds please</td>
<td>sleep sounds</td>
</tr>
<tr>
<td></td>
<td>Soothing Sleep Sounds</td>
<td>soothing sleep sounds</td>
<td>sleep sounds</td>
</tr>
<tr>
<td></td>
<td>My Sleep Sounds</td>
<td>the sleep sounds</td>
<td>sleep sounds</td>
</tr>
<tr>
<td></td>
<td>Super Sleep Sounds</td>
<td>sleep sounds</td>
<td>sleep sounds</td>
</tr>
<tr>
<td></td>
<td>Incredible Fast Sleep</td>
<td>incredible fast sleep</td>
<td>N/A</td>
</tr>
<tr>
<td>Google</td>
<td>Walk Log</td>
<td>walk log</td>
<td>work log</td>
</tr>
</tbody>
</table>

All Passed vetting processes, and got published
How realistic are those attacks?

Users might notice the system invoked the wrong skills, therefore, quickly exited.

Those higher numbers of attack skills suggest we have actually stolen users from the victim skill.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep sounds please</td>
<td>325</td>
<td>3,179</td>
<td>9.58</td>
<td>1.11</td>
<td>0.61</td>
<td>0.73</td>
</tr>
<tr>
<td>soothing sleep sounds</td>
<td>294</td>
<td>3,141</td>
<td>10.44</td>
<td>1.28</td>
<td>0.73</td>
<td>0.87</td>
</tr>
<tr>
<td>the sleep sounds</td>
<td>144</td>
<td>1,248</td>
<td>8.49</td>
<td>1.11</td>
<td>0.33</td>
<td>0.45</td>
</tr>
<tr>
<td>sleep sounds</td>
<td>109</td>
<td>1,171</td>
<td>10.18</td>
<td>1.59</td>
<td>0.51</td>
<td>0.82</td>
</tr>
<tr>
<td>incredible fast sleep</td>
<td>200</td>
<td>1,254</td>
<td>6.12</td>
<td>0.56</td>
<td>0.06</td>
<td>0.11</td>
</tr>
</tbody>
</table>
### Real-World Attack Measurement

<table>
<thead>
<tr>
<th># of Skills</th>
<th># of unique invocation names</th>
<th>Transformation cost</th>
<th>Skills has CIN$^*$ in market</th>
<th>Skills has CIN in market excluding same spelling</th>
<th>Skills has CIN in market through utterance paraphrasing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td># of skills</td>
<td>Avg. CINs per skill</td>
<td>Max CINs</td>
</tr>
<tr>
<td>19,670</td>
<td>17,268</td>
<td>0</td>
<td>3,718 (19%)</td>
<td>5.36</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt; 1</td>
<td>345 (1.8%)</td>
<td>1.04</td>
<td>38</td>
</tr>
</tbody>
</table>

66 skills were named as “cat facts” and provided similar functions.

**Interesting cases**

- **dog fact**
- **me a dog fact**

“SCUBA Diving Trivia” Skill and “Soccer Geek” skill, registered “space geek” as invocation names.