Life After Speech Recognition: Fuzzing Semantic Misinterpretation for Voice Assistant Applications

Yangyong Zhang, Lei Xu, Abner Mendoza, Guangliang Yang, Phakpoom Chinprutthiwong, Guofei Gu

Texas A&M University

NDSS’19 San Diego Feb 27th
Amazon’s Alexa Is Totally Baffled by My Bilingual Family

Alexa skills top 80,000 after a big Alexa-powered holiday season
Do you really know/trust who (Voice Assistants or VAs) you are talking to?

A couple says that Amazon’s Alexa recorded a private conversation and randomly sent it to a friend [read.bi/2xbra0R via @businessinsider](https://read.bi/2xbra0R).

Today in Google Home accidents I wanted to watch Nifty videos on YouTube and instead it played some really messed up “Patriotism for Kids” album on YouTube music. Oh no.

The spy inside your car
Voice Assistant

noun

• a computer program that can hold a conversation with somebody and complete particular tasks by responding to instructions or to information that it gathers from that person's digital device
a computer program that can hold a conversation with somebody and complete particular tasks by responding **to instructions** or **to information that it gathers** from that person's digital device.

Existing Work

- unrecognizable or inaudible malicious voice commands
  - Close Range
  - Physical
  - Speech Recognition
• a computer program that can hold a conversation with somebody and **complete particular tasks**
  by responding **to instructions** or **to information that it gathers** from that person's digital device

Existing Work

• unrecognizable or inaudible malicious voice commands
  • Close Range
  • Physical
  • Speech Recognition

• speech misinterpretation
  • Remote
  • Application
  • Blackbox
• **a computer program** that can hold a conversation with somebody and **complete particular tasks** by responding **to instructions** or **to information that it gathers** from that person's digital device

Our Work
• unboxing the black box

• **speech misinterpretation**
  • Remote
  • Application
  • Blackbox
• a computer program that can hold a conversation with somebody and complete particular tasks by responding to instructions or to information that it gathers from that person's digital device

Our Work
• unboxing the black box

• speech misinterpretation
  • Remote
  • Application
  • Blackbox
• **a computer program** that can hold a conversation with somebody and **complete particular tasks** by responding **to instructions** or **to information that it gathers** from that person's digital device.

Our Work
- unboxing the black box
- a systematically study of **semantic misinterpretation** (what’s after speech-to-text)
1. What’s in the Voice Assistant “blackbox”?

2. What’s the root cause of semantic misinterpretation?

3. How to explore the problem systematically?
Voice Assistant Application (vApp) Architecture
what’s inside the box

- Speech Recognition: from audio to text
- Semantic interpretation: from text to intents
- Audio Response: from text to audio
Our Work..

- Semantic interpretation: from text to intents
Semantic Misinterpretation

- Even when the speech recognition works fine, NLU yields incorrect intents.

Today in Google Home accidents I wanted to watch Nifty videos on YouTube and instead it played some really messed up “Patriotism for Kids” album on YouTube music. Oh no.
Semantic Misinterpretation
Semantic Misinterpretation

- Not just Jennifer’s problem, but more
Problematic Intent Classifier
root cause of semantic misinterpretation

Developer Defined
“Alexa, install the true bank skill.”

Template: “Alexa, tell me my balance.”

SYSTEM.InstallIntent
CUSTOM.BalanceQueryIntent
vApp Squatting

Victim vApp
The *True* Bank Skill

Malicious vApp:
The *Truth* Bank Skill
Fuzzing Semantic Misinterpretation
explore the problem systematically

- Fuzz potentially dangerous voice command variants
  - Problem: too many options! ➔ large search space
Learn from Linguistic Knowledge

- Reduce the search space
- Find voice command utters that people likely to speak
- **Lapsus** is a concept beyond speech errors, it is any thing that makes machine misunderstand your intents

- Regional language
- New Vocabulary
- Logical Fallacies
  
  ....
LipFuzzer

- Learn from existing linguistic knowledge
- Fuzz potentially dangerous voice command utters

- Pinpoint the problematic voice commands
- Help start to eliminate the threats
LipFuzzer

- Learn from existing linguistic knowledge
- Fuzz potentially dangerous voice command utters

- Pinpoint the problematic voice commands
- Help start to eliminate the threats
LipFuzzer

• Learn from existing linguistic knowledge
• Fuzz potentially dangerous voice command utters

- Pinpoint the problematic voice commands
- Help start to eliminate the threats
Evaluation
for Lapsus and Lipfuzzer

• A misinterpreted voice commands can either due to ASR or the Intent Classifier
  • How much does the Intent Classifier (using developer defined templates) contribute?

• We collected 109 real Lapsus from Mturk (IRB approved)
• 77% of them are caused by improper intent classification
Evaluation for Lapsus and Lipfuzzer

- A misinterpreted voice command can either be due to ASR or the Intent Classifier
  - How much does the Intent Classifier (using developer-defined templates) contribute?

- We collected 109 real Lapsus from Mturk (IRB approved)
- 77% of them are caused by improper intent classification

Collected Lapsus Audio Record → Automatic Speech Recognition (ASR) → Textual Data

Is text result transcribed correctly? 23% from ASR
Evaluation for Lapsus and Lipfuzzer

- A misinterpreted voice commands can either due to ASR or the Intent Classifier
  - How much does the Intent Classifier (using developer defined templates) contribute?

- We collected 109 real Lapsus from Mturk (IRB approved)
- 77% of them are caused by improper intent classification

23% from ASR

77% (84/109)
Store-wide Scanning

• We use our trained Lapsus Model to find potentially vulnerable voice commands and vApps

• We let our chatbot speak with an Amazon Echo and a Google Home device
  • A verified vulnerable vApp is the one returns unintended result

<table>
<thead>
<tr>
<th>Store Name</th>
<th>Crawled vApp #</th>
<th>LipFuzzer-generated LAPSUS #</th>
<th>Potentially vulnerable vApp #</th>
<th>Verified vulnerable vApp % (Sampled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Alexa</td>
<td>32,892</td>
<td>497,059</td>
<td>26,790</td>
<td>71.5%</td>
</tr>
<tr>
<td>Google Assistant</td>
<td>2,328</td>
<td>11,390</td>
<td>2,295</td>
<td>29.5%</td>
</tr>
</tbody>
</table>
Attacking The True Bank Skill

• Use Amazon Mechanical Turk (MTurk) to let users “use” The True Bank Skill
  • We show the web user interface to MTurk workers
  • We collect their voice command audio records
  • Play these records to Alexa
• We check if users would eventually be using those generated “Malicious” vApp
• 4 Malicious ones are found to be “triggered”
Summary

• **Unboxes** the black box of VA platforms

• Pinpoint the **Intent Classifier** problem and **3-party vApp** development

• Find potentially dangerous voice commands in large scale (**LipFuzzer**) 

• An early signal that the lack of security considerations for vApp backend processing

• Limitations?
  • Need more data, more knowledge
  • Better Modelling and Machine Learning techniques
Take Away
go beyond

• A similar trend: vApp and Mobile App

• More discovery even after the NLU component?

• How to deal with the human factors in Voice User Interface (VUI)?
Contact Us:
yangyong@tamu.edu

Project Release:
success.cse.tamu.edu/lab/lipfuzzer