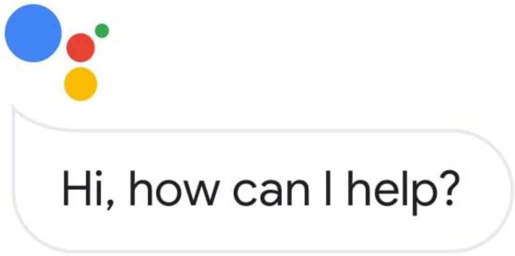
Life after Speech Recognition: Fuzzing Semantic Misinterpretation for Voice Assistant Applications

Presented by: Jeffrey Young, Margi Engineer, Bhavik Suthar

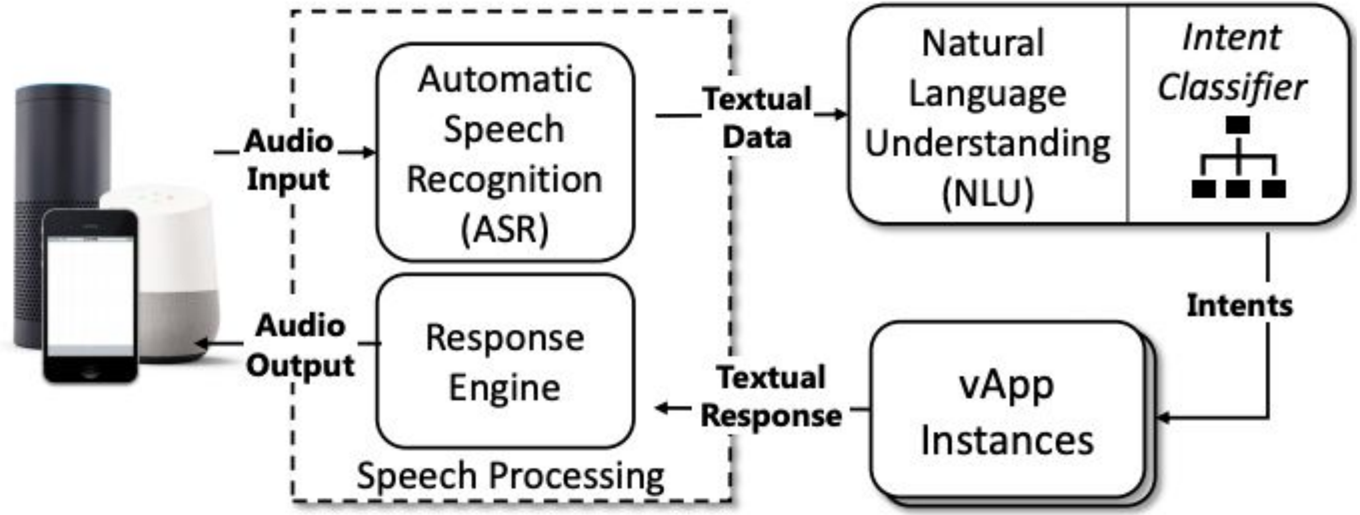
Introduction

- What are Virtual Assistant (VA)?
- Fastest growing Voice User Interface (VUI)-based Technology
  - Amazon Alexa Skills
  - Google Assistant Actions
- 30,000 voice assistant applications (vApps) are available for Amazon Alexa
- Attacks on Automatic Speech Recognition(ASR)
  - Acoustic-based attack
  - Squatting attack



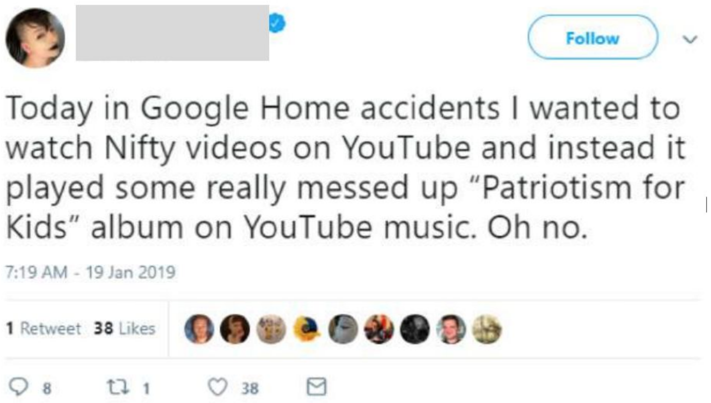
VUI-Based VA Architecture

- Speech Recognition
- Semantic interpretation
- Audio Response
- Speech recognition
  - ASR - Automatic Speech recognition
  - NLU - Natural Language Understanding
    - Intent Classifier



Challenges

- NLU yielding incorrect intends.
- Closed Development (Amazon Alexa and Google Assistant)
    - Difficult to conduct white box analysis
- Strong Privacy Enforcement
    - Impossible to get real users' speech input
    - vApp response output of the user
- Fuzzing Scheme
    - Limitation of I/O in vApp
    - Squatting attack on vApps



[redacted]

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.

7:19 AM - 19 Jan 2019

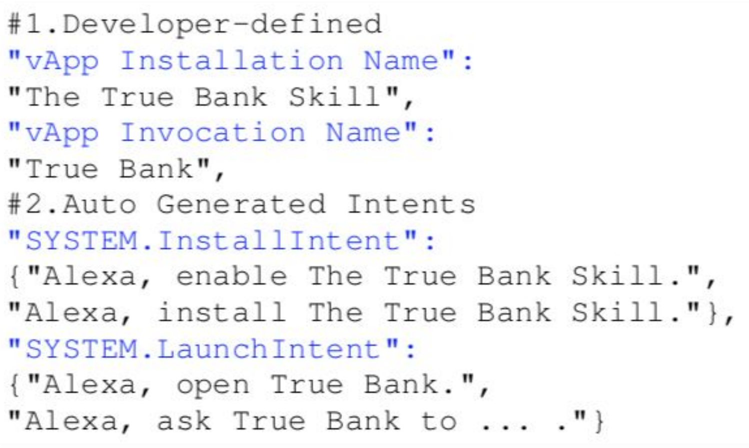
1 Retweet 38 Likes

8 1 38

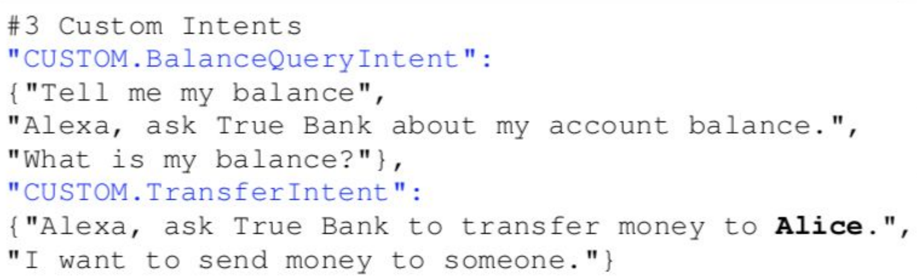
Jennifer

Background

- Natural Language Understanding of vApp
  - Intent Classifier
  - Slot Values
  - Fuzzing Matching
- Linguistic-related LAPSUS (Latin word for Slip)
- Motivation Example
- Threat
  - Denial of Service
  - Privacy Leakage
  - Phishing

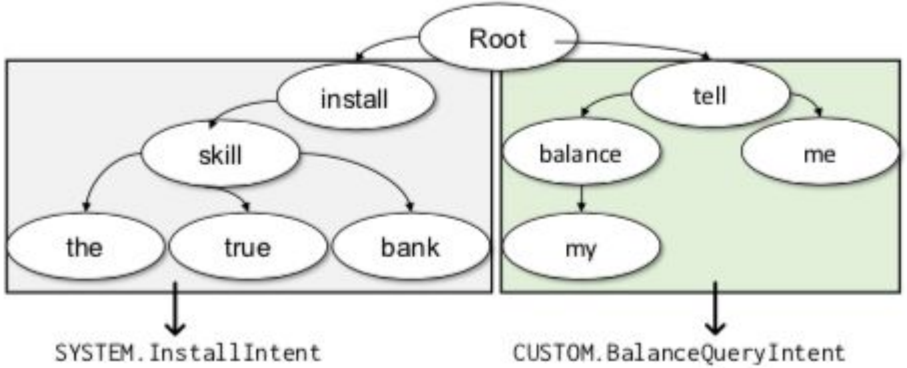


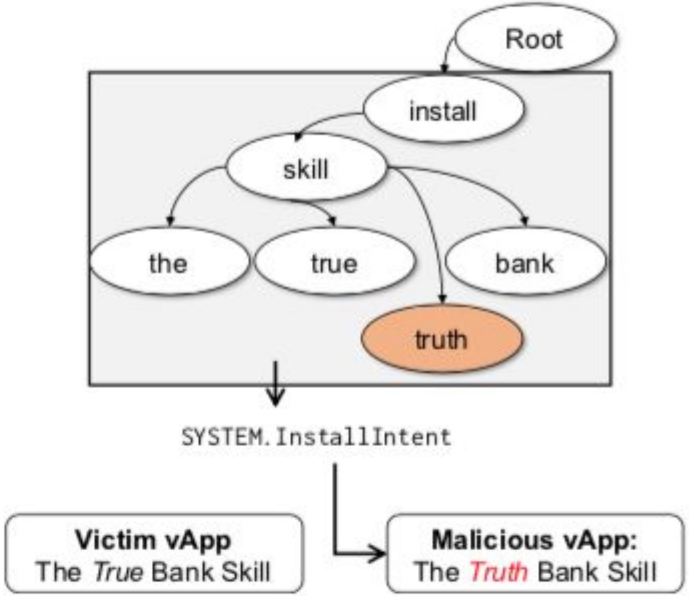
```
#1.Developer-defined
"vApp Installation Name":
"The True Bank Skill",
"vApp Invocation Name":
"True Bank",
#2.Auto Generated Intents
"SYSTEM.InstallIntent":
{"Alexa, enable The True Bank Skill.",
"Alexa, install The True Bank Skill."},
"SYSTEM.LaunchIntent":
{"Alexa, open True Bank.",
"Alexa, ask True Bank to ... ."}
```



```
#3 Custom Intents
"CUSTOM.BalanceQueryIntent":
{"Tell me my balance",
"Alexa, ask True Bank about my account balance.",
"What is my balance?"},
"CUSTOM.TransferIntent":
{"Alexa, ask True Bank to transfer money to Alice.",
"I want to send money to someone."}
```

Intent Classification and Squatting

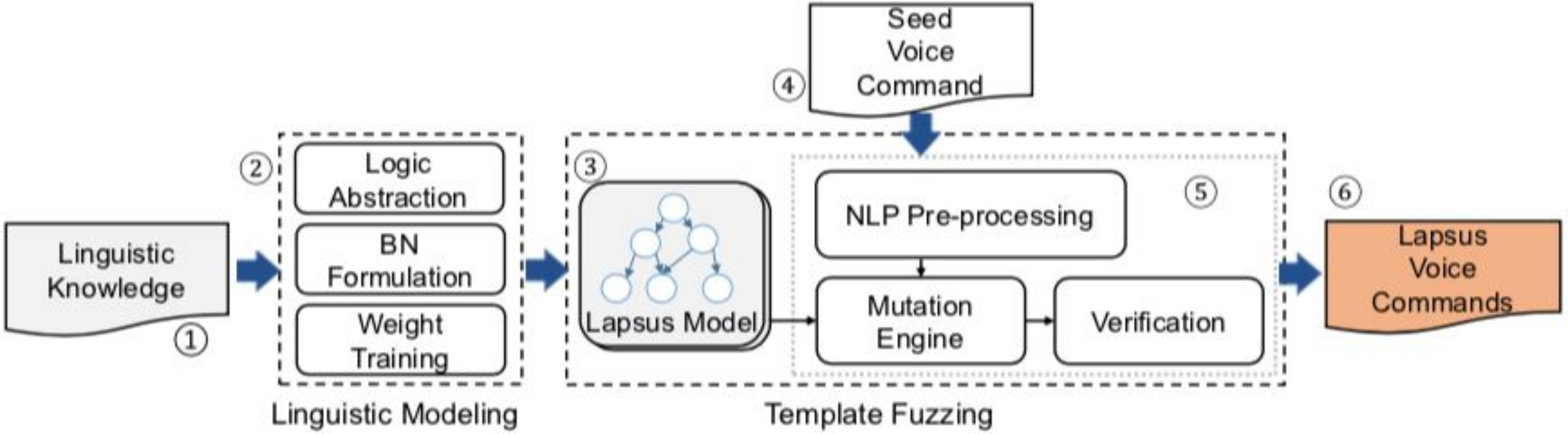




Fuzzing Speech Misinterpretation

- Challenges
  - Mutation Field
    - Low-level phonetic
    - High-level phrases
  - Space Explosion
    - Random fuzzing to generate LAPSUS
    - Voice command can be mutated with phonemes or words
- What is LipFuzzer?
  - Learn for existing Linguistic Knowledge
  - Potential dangerous voice command utters are fuzzed

LipFuzzer Architecture



- Helps in starting to eliminate threats
- Points out voice commands that are problematic

Linguistic Model- Guided Fuzzing

LipFuzzer's Design

A. Fuzzing Input & Output
   a. Linguistic Knowledge
   b. Lapsus Model
   c. Seed Input
   d. Fuzzing Output
B. Linguistic Model
   a. Logic Abstraction
   b. Bayesian Network Formulation
   c. Weight Training
C. Template Fuzzing
   a. NLP Pre-processing
   b. Mutation Engine
   c. Verification

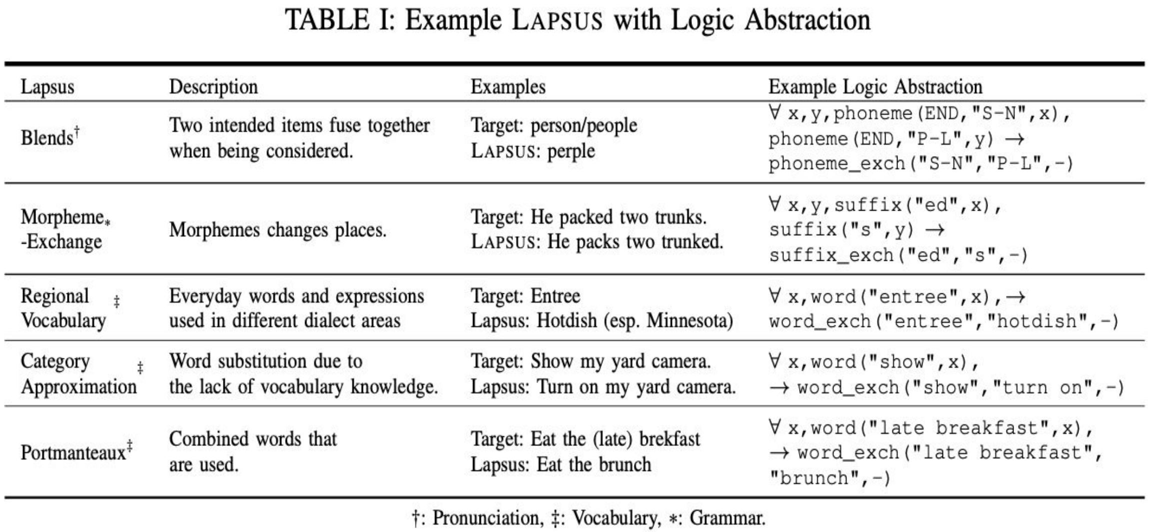


TABLE I: Example LAPSUS with Logic Abstraction

Lapsus	Description	Examples	Example Logic Abstraction
Blends†	Two intended items fuse together when being considered.	Target: person/people LAPSUS: perple	`∀ x,y,phoneme(END,"S-N",x), phoneme(END,"P-L",y) → phoneme_exch("S-N","P-L",-)`
Morpheme∗ -Exchange	Morphemes changes places.	Target: He packed two trunks. LAPSUS: He packs two trunked.	`∀ x,y,suffix("ed",x), suffix("s",y) → suffix_exch("ed","s",-)`
Regional ‡ Vocabulary	Everyday words and expressions used in different dialect areas	Target: Entree Lapsus: Hotdish (esp. Minnesota)	`∀ x,word("entree",x),→ word_exch("entree","hotdish",-)`
Category ‡ Approximation	Word substitution due to the lack of vocabulary knowledge.	Target: Show my yard camera. Lapsus: Turn on my yard camera.	`∀ x,word("show",x), → word_exch("show","turn on",-)`
Portmanteaux‡	Combined words that are used.	Target: Eat the (late) brekfast Lapsus: Eat the brunch	`∀ x,word("late breakfast",x), → word_exch("late breakfast", "brunch",-)`

†: Pronunciation, ‡: Vocabulary, ∗: Grammar.

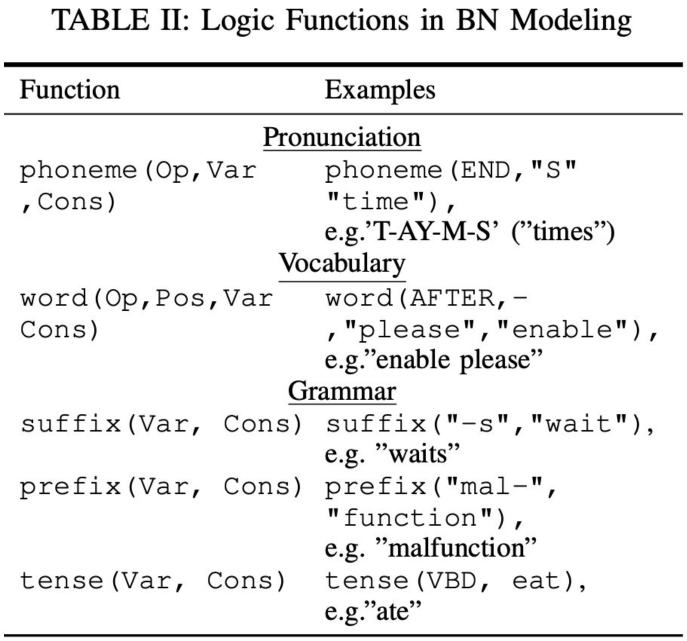
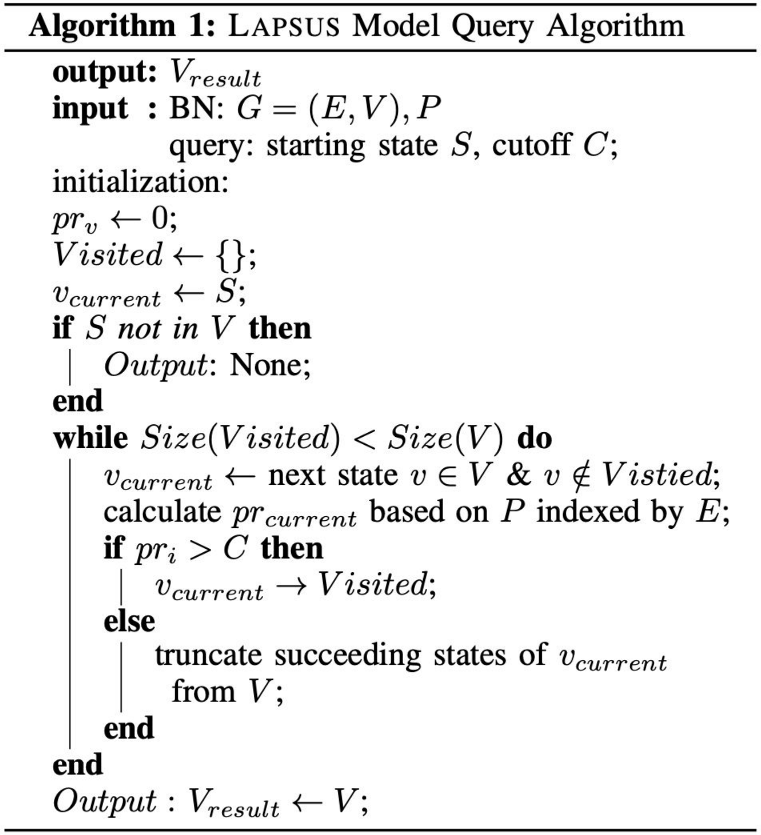


TABLE II: Logic Functions in BN Modeling

Function	Examples
Pronunciation	
`phoneme(Op,Var ,Cons)`	`phoneme(END,"S" "time"),` e.g.’T-AY-M-S’ (”times”)
Vocabulary	
`word(Op,Pos,Var Cons)`	`word(AFTER,- ,"please","enable"),` e.g.”enable please”
Grammar	
`suffix(Var, Cons)`	`suffix("-s","wait"),` e.g. ”waits”
`prefix(Var, Cons)`	`prefix("mal-", "function"),` e.g. ”malfunction”
`tense(Var, Cons)`	`tense(VBD, eat),` e.g.”ate”



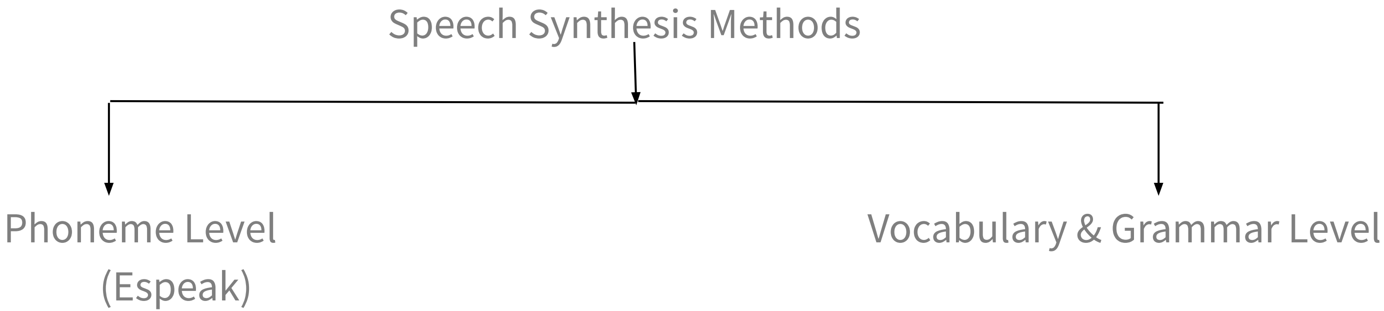
Algorithm 1: LAPSUS Model Query Algorithm

```
output: $V_{result}$
input : BN: $G = (E, V), P$
        query: starting state $S$, cutoff $C$;
initialization:
$pr_v \leftarrow 0$;
$Visited \leftarrow \{\}$;
$v_{current} \leftarrow S$;
if $S$ not in $V$ then
    $Output$: None;
end
while $Size(Visited) < Size(V)$ do
    $v_{current} \leftarrow$ next state $v \in V$ & $v \notin Vistied$;
    calculate $pr_{current}$ based on $P$ indexed by $E$;
    if $pr_i > C$ then
        $v_{current} \rightarrow Visited$;
    else
        truncate succeeding states of $v_{current}$
         from $V$;
    end
end
$Output$ : $V_{result} \leftarrow V$;
```

Weight Based

Implementation

- LipFuzzer



Speech Synthesis Methods

Phoneme Level
(Espeak)

Vocabulary & Grammar Level

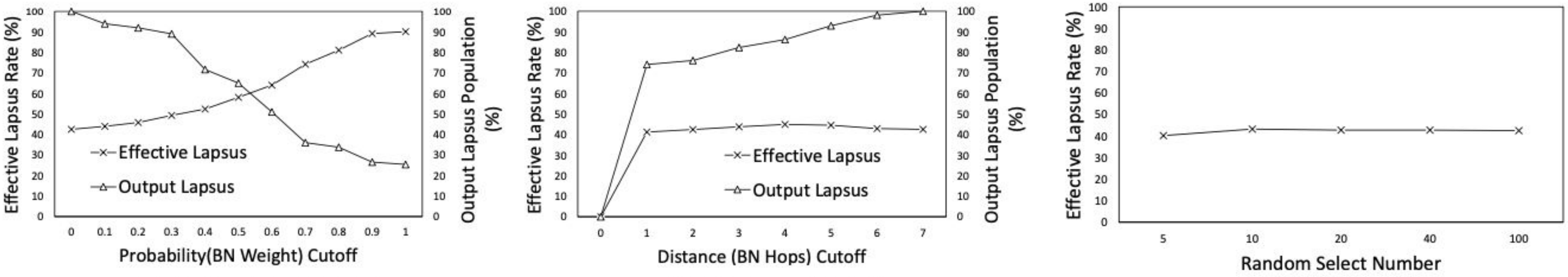
- User Study: Amazon MTurk (Mechanical Turk) crowdsourcing for collecting data (users and developers)

Evaluation

Three Goals:

- Empirically verify that the problematic Intent Classification can lead to speech misinterpretation related to LAPSUS
- Show LipFuzzer's performance in terms of LAPSUS model's accuracy and effectiveness
- LipFuzzer to reveal that problematic templates widely exist in both Amazon Alexa and Google Assistant platform

- Intent Classifier
  - Leverage user-group data to locate Lapsus
  - Amazon echo to check if semantic inconsistency still exists
  - Recorded audio—>text—>speech synthesis tool: 77% (89/109) semantic Inconsistency
  - **Concluded: Empirically verify that the problematic Intent Classification can lead to speech misinterpretation related to LAPSUS**
- LipFuzzer Evaluation
  - Different cutoff to generate LAPSUS
  - Fuzzing accuracy through generated LAPSUS
  - Fuzzing Effectiveness with average LAPSUS produced from seed template
- vApp StoreEvaluation
  - 71.5% Amazon Alexa vaApp and 29.5% of Google Assistant are verified to be vulnerable
- Case Study
  - vApp squatting attack

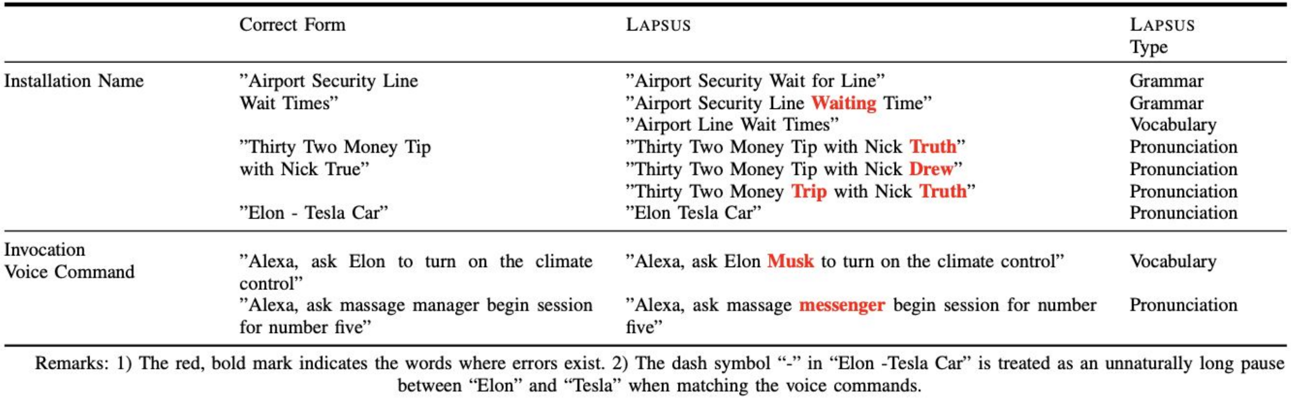


(a) Probability Cutoff

(b) Distance Cutoff

(c) Random Select

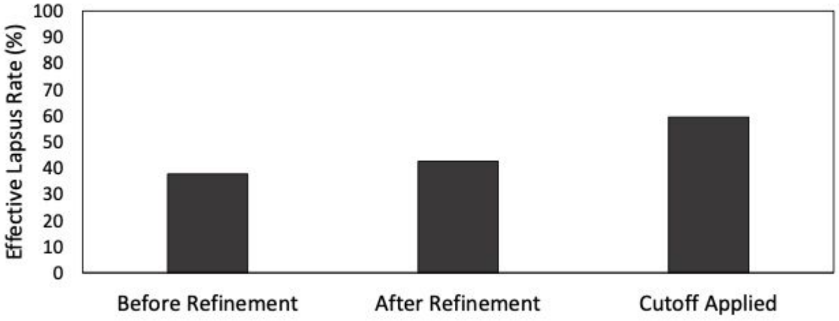
Different cutoff to generate LAPSUS



	Correct Form	LAPSUS	LAPSUS Type
Installation Name	”Airport Security Line Wait Times”	”Airport Security Wait for Line”	Grammar
		”Airport Security Line **Waiting** Time”	Grammar
		”Airport Line Wait Times”	Vocabulary
	”Thirty Two Money Tip with Nick True”	”Thirty Two Money Tip with Nick **Truth**”	Pronunciation
		”Thirty Two Money Tip with Nick **Drew**”	Pronunciation
		”Thirty Two Money **Trip** with Nick **Truth**”	Pronunciation
	”Elon - Tesla Car”	”Elon Tesla Car”	Pronunciation
Invocation Voice Command	”Alexa, ask Elon to turn on the climate control”	”Alexa, ask Elon **Musk** to turn on the climate control”	Vocabulary
	”Alexa, ask massage manager begin session for number five”	”Alexa, ask massage **messenger** begin session for number five”	Pronunciation

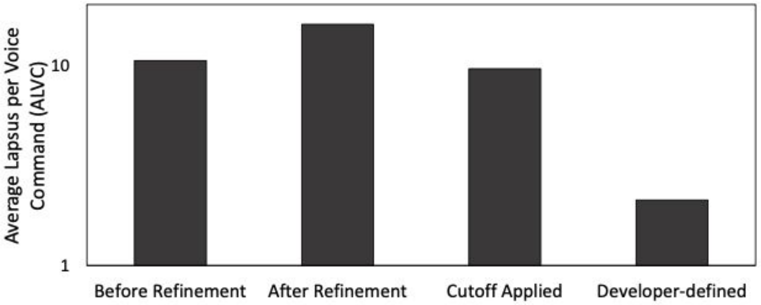
Remarks: 1) The red, bold mark indicates the words where errors exist. 2) The dash symbol “-” in “Elon -Tesla Car” is treated as an unnaturally long pause between “Elon” and “Tesla” when matching the voice commands.

LAPSUS Example collected from users



Fuzzing Accuracy

- After refinement effective LAPSUS rate increases from 37.7% to 42.5%
- With cutoff applied it increases to 59.52%



Fuzzing Effectiveness

- Uses ALVC (Average Lapsus per voice command)
- Applying cutoff reduces the ALVC to 9.57 as fewer states in models are present

- Evaluated 300 voice commands with ALVC of 2.12 (So empirically proved that the finding LAPSUS is better than manual effort)
- Implementation takes 11.4 seconds/ seed template fuzzing

TABLE IV: vApp Store-wide Fuzzing Results

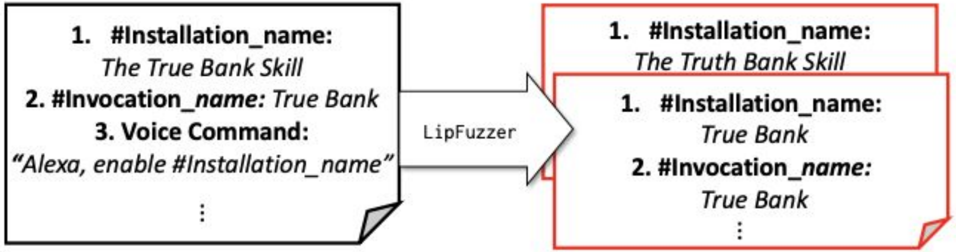
Store Name	Crawled vApp #	LipFuzzer-generated LAPSUS #	Potentially vulnerable vApp #	Verified vulnerable vApp % (Sampled)	Potentially vulnerable vApps # Zhang et al. [37]	Vulnerable vApps # Kumar et al. [28]
Amazon Alexa	32,892	497,059	26,790	71.5%	531	25
Google Assistant	2,328	11,390	2,295	29.5%	4	N/A

Remark: N/A means not applicable because Google Assistant was not evaluated in the work.

TABLE V: LAPSUS for Example vApps

Intended Voice Command	LAPSUS	Effective LAPSUS?
"Paypal" (installation)	"Pay-ple"	✓
	"Pay-ples"	✓
"ask PayPal to check my balance"	"ask PayPal to check my balances"	✗
	"ask PayPal to check my balancing"	✗
	"ask PayPal to check my balancing"	✗
	"ask PayPal to checks my balance"	✗
	"ask PayPal to checking my balance"	✗
"Skyrim Very Special Edition" (installation)	"Skyrim Very Special Edit"	✓
	"Skyrim Special Edition"	✓
	"Skyrim Very Specially Edition"	✗
	"Sky-ram Special Edition"	❍
	"Sky-im Special Edition"	❍

✓: Effective, ✗: Ineffective, ❍: Maybe Effective



Related Work - Discussion

Attacking ASR through Acoustic Channels:

1. **Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone**
2. **Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition**
3. **Hidden Voice Commands**
4. **CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition**

More powerful attacks using inaudible voice commands:

5. **BackDoor: Making Microphones Hear Inaudible Sounds**
6. **Dolphinattack: Inaudible voice commands**

Related Work - Discussion

Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone

Basic Attack: Malicious Number Dialing

Extended Attack: Sensitive Permissions Bypassing

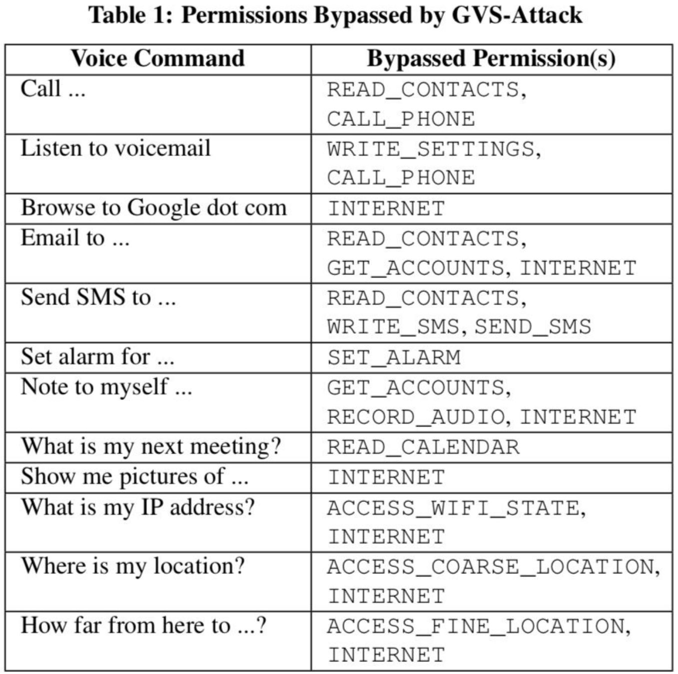


Table 1: Permissions Bypassed by GVS-Attack

Voice Command	Bypassed Permission(s)
Call ...	`READ_CONTACTS`, `CALL_PHONE`
Listen to voicemail	`WRITE_SETTINGS`, `CALL_PHONE`
Browse to Google dot com	`INTERNET`
Email to ...	`READ_CONTACTS`, `GET_ACCOUNTS`, `INTERNET`
Send SMS to ...	`READ_CONTACTS`, `WRITE_SMS`, `SEND_SMS`
Set alarm for ...	`SET_ALARM`
Note to myself ...	`GET_ACCOUNTS`, `RECORD_AUDIO`, `INTERNET`
What is my next meeting?	`READ_CALENDAR`
Show me pictures of ...	`INTERNET`
What is my IP address?	`ACCESS_WIFI_STATE`, `INTERNET`
Where is my location?	`ACCESS_COARSE_LOCATION`, `INTERNET`
How far from here to ...?	`ACCESS_FINE_LOCATION`, `INTERNET`

Related Work - Discussion

Attacking ASR through Acoustic Channels:

Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition

The gap between human and machine speech recognition leads to an unmonitored channel by which an adversary can inject commands.

- ***Initiate a drive-by-download****:* An attacker can issue commands to open a webpage maintained by the adversary that contains a drive-by-download. This effectively serves as a stepping stone, enabling other attacks that exploit vulnerabilities in the device's browser;
- ***Earn money via pay-based SMS services****:* An attacker can construct audio that causes phones to send text messages to pay-based SMS short code services that it operates;
- ***Enumerate devices in a physical area****:* Similarly, an attacker may use a loudspeaker to cause nearby phones to send SMS messages to a number that the adversary controls, allowing it to enumerate the devices that are physically located within "earshot" of the broadcast (e.g., those belonging to dissidents attending a rally);
- ***Earn money via premium rate services:*** An attacker can operate a premium rate number (i.e., a "900 number") and monetize his attack by causing nearby phones to call it (for some mobile devices and calling plans);
- ***Perform a denial-of-service attack****:* Using a public announcement system, an attacker can issue commands to turn on airplane mode on all devices, preventing them from receiving calls and other communications.

Related Work - Discussion

More powerful attacks using inaudible voice commands:

BackDoor: Making Microphones Hear Inaudible Sounds

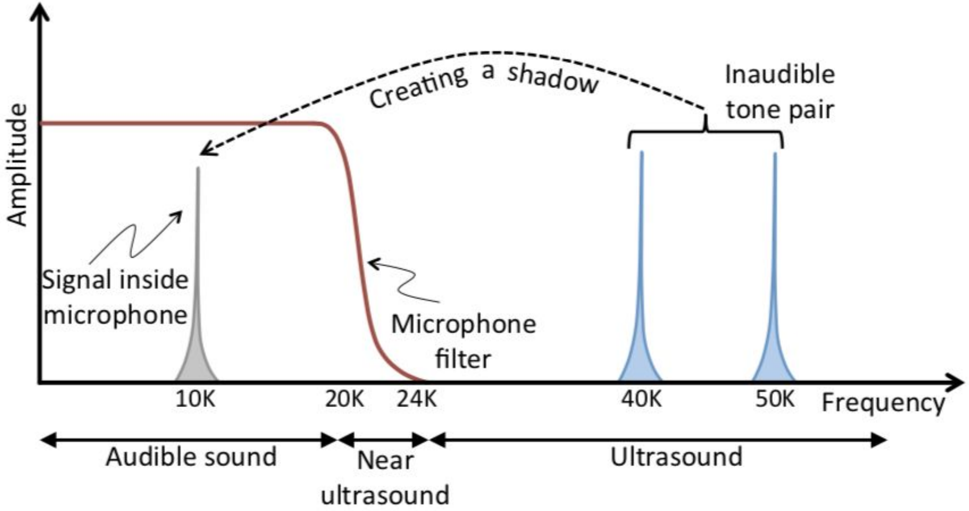


Figure 1: The main idea underlying *BackDoor*.

Related Work - Discussion

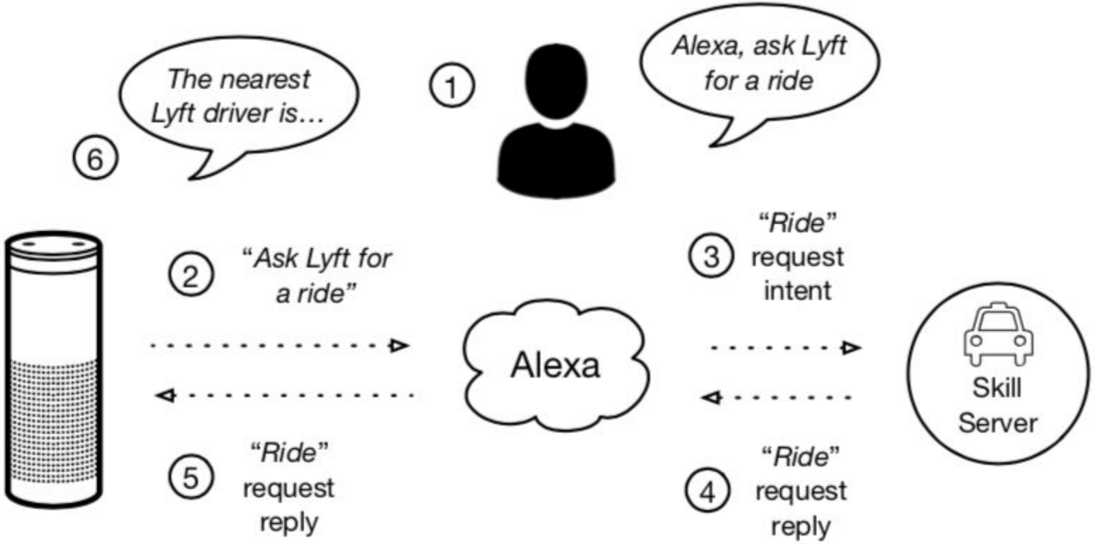
Attacking ASR with Misinterpretation:

1. **Skill Squatting Attacks on Amazon Alexa**
2. **Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems**

Related Work - Discussion

Attacking ASR with Misinterpretation:

1. **Skill Squatting Attacks on Amazon Alexa**



Related Work - Discussion

Discussion

Bayesian Networks are sufficient to demonstrate the effectiveness.

With a more complete dataset, LAPSUS can be produced more effectively.

The model can be improved through more complicated logical representation. (i.e. use predicate logic to represent hybrid LAPSUS).

Another direction for LipFuzzer is to improve the training and mutation process.

Conclusion

1. Systematically studied how Intent Classifier affects the security of popular vApps.

2. Currently used vApp templates can incur dangerous semantic inconsistencies.

3. Design the first linguistic-guided fuzzing tool to systematically discover the the speech misinterpretations that lead to such inconsistencies.

Thank You!

Questions?