Home Digital Voice Assistants: use cases and vulnerabilities

Oleksandra Baga
Master Computer Science, Freie Universität Berlin
Sommersemester 2021, Seminar Technische Informatik
oleksandra.baga@gmail.com

***Abstract*— If you want to own a new modern digital voice assistant like Alexa or Google Echobot you must read this paper to get know about potential risks and lacks of your personal data**

I. INTRODUCTION

II. DIGITAL VOICE ASSISTANCES

A. Usage by people of every age

III. USE CASES

IV. VULNERABILITIES

A. Voice-based remote attacks

1) Voice squatting attack (VSA):

2) Voice Masquerading Attack (VMA):

V. COPY AND PASTE CORE IDEAS

A. About

VPA on IoT devices. Amazon and Google are two major players in the market of smart speakers with voice-controlled personal assistant capabilities. Since the debut of the first Amazon Echo in 2015, Amazon has now taken 76% of the U.S. market with an estimate of 15-million devices sold in the U.S. alone in 2017. A unique property of these four devices is that they all forgo conventional I/O interfaces, such as the touchscreen, and also have fewer buttons (to adjust volume or mute), which serves to offer the user a hands-free experience. In another word, one is supposed to command the device mostly by speaking to it. For this purpose, the device is equipped with a microphone circular array designed for 360-degree audio pickup and other technologies like beamforming that enable far-field voice recognition.

Behind these smart devices is a virtual personal assistant, called Alexa for Amazon and Google Assistant for Google, engages users through a two-way conversation. Unlike those serving a smartphone (Siri, for example) that can be activated by a button push, the VPAs for these IoT devices are started with a wake-word like “Alexa” or “Hey Google”. These assistants have a range of capabilities, from weather report, timer setting, to-do list maintenance to voice shopping, hands-free messaging and calling. The user can manage these capabilities through a companion app running on her smartphone.

B. Skills and actions

Both Amazon and Google enrich the VPAs’ capabilities by introducing voice assistant function called skill by Amazon or action by Google. Skills are essentially third-party apps, like those running on smartphones, offering a variety of services the VPA itself does not provide. Examples include Amex, Hands- Free Calling, Nest Thermostat and Walmart. These skills can be conveniently developed with the supports from Amazon and Google, using Alexa Skills Kit [32] and Actions on Google. Indeed, we found that up to November 2017, Alexa already has 23,758 skills and Google Assistant has 1,001. HOW MANY DO WE HAVE NOW?!

Both Amazon Alexa and Google Assistant run a skill market that can be accessed from their companion app on smartphones or web browser for users to discover new skills.

Skills can be started either explicitly or implicitly. Explicit invocation takes place when a user requires a skill by its name from a VPA: for example, saying ?Alexa, talk to Amex? to Alexa triggers the Amex skill for making a payment or checking bank account balances. Such a type of skills is also called custom skills on Alexa.

Implicit invocation occurs when a user tells the voice assistant to perform some tasks without directly calling to a skill name. For example, ?Hey Google, will it rain tomorrow?? will invoke the Weather skill to respond with a weather forecast. Google Assistant identifies and activates a skill implicitly whenever the conversation with the user is under the context deemed appropriate for the skill. This invocation mode is also supported by the Alexa for specific types of skills.

Specifically, to invoke a skill explicitly, the user is expected to use a wake-word, a trigger phrase, and the skill?s invocation name. For example, for the spoken sentence ?Hey Google, talk to personal chef?, ?Hey Google? is the wake-word, ?talk to? is the trigger phrase, and ?personal chef? is the skill invocation name. Note that skill invocation name could be different from skill name, which is intended to make it simpler and easier for users to pronounce. For example, ?The Dog Feeder? has invocation name as the dog; ?Scryb? has invocation name as scribe. When a user invokes a VPA device with its wake-word, the device captures her voice command and sends it to the VPA service provider?s cloud for processing. The cloud performs speech recognition to translate the voice record into text, finds out the skill to be invoked, and then delivers the text, together with the

timestamp, device status, and other meta-data, as a request to the skill?s web service. Note that the skill will only receive requests in text format rather than the users? voice recordings.To publish a skill, the developer needs to submit the informa- tion about her skill like name, invocation name, description and the endpoint where the skill is hosted for a certification process. This process aims at ensuring that the skill is functional and meets the VPA provider?s security requirements and policy guidelines.

C. Adversary Model

today anyone can publish her skill through Amazon and Apple markets, given that these markets have only minimum protection in place to regulate the functions submitted: almost nothing on Amazon before our attacks were reported2, and only the basic check is performed on Google to find duplicated invocation names. once a malicious skill is published, it can be transparently launched by the victim through her voice commands, without being downloaded and installed on her device. Therefore, they can easily affect a large number of VPA IoT devices.

we found that for Amazon, such names are not unique skill identifiers: multiple skills with same invocation names are on the Amazon market. Also, skills may have similar or related names.66 different Alexa skills are called cat facts, 5 called cat fact and 11 whose invocation names contain the string ?cat fact?, e.g. fun cat facts, funny cat facts. When such a common name is spoken, Alexa chooses one of the skills based on some undisclosed policies (possibly random as observed in our research). When a different but similar name is called, however, longest string match is used to find the skill. For example, ?Tell me funny cat facts? will trigger funny cat facts rather than cat facts. This problem is less serious for Google, which does not allow duplicated invocation names. However, it also cannot handle similar names

D. Voice-based remote attacks.

In our research, we analyzed the most popular VPA IoT systems - Alexa and Google Assistant, focusing on the third-party skills deployed to these devices. It is completely feasible for an adversary to remotely attack the users of these popular systems, collecting their private information through their conversations with the systems [1].

Voice squatting attack (VSA): the adversary exploits how a skill is invoked (by a voice command), and the variations in the ways the command is spoken (e.g., phonetic differences caused by accent, courteous expression, etc.) to cause a VPA system to trigger a malicious skill instead of the one the user intends [1]. For example, one may say “Alexa, open Capital One please”, which normally opens the skill Capital One, but can trigger a malicious skill Capital One Please once it is uploaded to the skill market. In response to the commands, a malicious skill can pretend to yield control to another skill (switch) or the service (terminate), yet continue to operate stealthily to impersonate these targets and get sensitive information from the user.

More specifically, we first surveyed 156 Amazon Echo and Google Home users and found that most of them tend to use natural languages with diverse expressions to interact with the devices: e.g., “play some sleep sounds” . These expressions allow the adversary to mislead the service and launch a wrong skill in response to the user’s voice command, such as *some sleep sounds* instead of *sleep sounds* [1].

Our further analysis of both Alexa and Google Assistant demonstrates that indeed these systems identify the skill to invoke by looking for the longest string matched from a voice command. From these responses, we found that 50% of the Amazon Echo users used ?please? at least once in their invocation examples, so did 41% of the Google Home users. Also, 28% users reported that they did open unintended skills when talking to their devices. [1].

found that an adversary can intentionally induce confusion by using the name or similar one of a target skill, to trick the user into invoking an attack skill when trying to open the target. For example, the adversary who aims at Capital One could register a skill Capital Won, Capitol One, or Captain One. All such names when spoken by the user could become less distinguishable, particularly in the presence of noise, due to the limitations of today?s speech recognition techniques.

To study voice squatting, we randomly sampled 100 skills each from Alexa and Google assistant markets. For this purpose, we studied two types of the attacks: voice squatting in which an attack skill carries a phonetically similar invocation name to that of its target skill, and word squatting where the attack invocation name includes the target?s name and some strategically selected additional words (e.g., ?cat facts please?). During this study, we found that a mispronounced invocation name would also trigger the right skill if their pronunciation is close and there is no other registered skills using the mispronounced invocation name. ...we then register ?captain one? as the attack skill?s invocation name, play the original invocation utterance... Such skills were invoked five times each in the test modes of Alexa and Google Assistant.

To study word squatting, we randomly sampled ten skills from each skill markets as the attack targets. For each skill, we built four new skills whose invocation names include the target?s name together with the terms identified from our survey study (Section III-A): for example, ?cat facts please? and ?my cat facts?. On Alexa, an attack skill with the extended name (that is, the target skill?s invocation name together with terms ?please?, ?app?, ?my? and ?the?) was almost always launched by the voice commands involving these terms and the target names. On Google Assistant, however, only the utterance with word ?app? succeeded in triggering the corresponding attack skill, which demonstrates that Google Assistant is more robust against such an attack. However, when we replaced ?my? with ?mai? and ?please? with ?plese?, all such attack skills were successfully invoked by the commands for their target skills (see Table IV). This indicates that the protection Google puts in place (filtering out those with suspicious terms) can be easily circumvented.

Voice Masquerading Attack (VMA):

Google Assistant seems to have protection in place against the impersonation. Specifically, it signals the launch of a skill by speaking ?Sure, here is?, together with the skill name and a special earcon, and skill termination with another earcon. Both Alexa and Google Assistant support voluntary skill termination, allowing a skill to terminate itself right after making a voice response to the user. according to our survey study, 78% of the participants rely on the response of the skill (e.g. ?Goodbye? or silence) to determine whether a skill has been terminated. This allows an attack skill to fake its termination by providing ?Goodbye? or silent audio in its response while keeping the session alive.

When sending back a response, both Alexa and Google Assistant let a skill include a reprompt (text content or an audio file), which is played when the VPA does not receive any voice command from the user within a period of time. If the user continues to keep quiet, after another 6 seconds for Alexa and one additional reprompt from Google and follow-up 8-second waiting, the running skill will be forcefully terminated by the VPA. On the other hand, we found in our research that as long as the user says something (even not meant for the skill) during that period, the skill is allowed to send another response together with a reprompt. Adding a silent audio file (up to 90 seconds for Alexa and 120 seconds for Google Assistant) will make it to be able to continue to run at least 102 seconds on Alexa and 264 seconds on Google. This running time can be further extended considering the attack skill attaching the silent audio right after its last voice response to the user (e.g., ?Goodbye?), which gives it 192 seconds on Alexa and 384 on Google Assistant), and indefinitely whenever Alexa or Google Assistant picks up some sound made by the user. In this case, the skill can reply with the silent audio and in the meantime, record whatever it hears.

Only ?exit? is processed by the VPA service and used to forcefully terminate the skill. Through survey study, we found that 91% of Alexa users used ?stop? to terminate a skill, 36% chose ?cancel?, and only 14% opted for ?exit?, which suggests that the user perception is not aligned with the way Alexa works and therefore leaves the door open for the VMA. Consequently, all the information stealing and Phishing attacks caused by the VSA can also happen here.

THIS WAS SOURCE 1 [1]

SOURCE 2 [2] In the last years, the heavy use of smartphones led to the appearance of voice assistants such as Apple?s Siri, Google?s Assistant, Microsoft?s Cortana and Amazon?s Alexa. Voice as-sistants use technologies like voice recognition, speech synthesis, and Natural Language Processing (NLP) to provide services to the users. Cloud platforms are now enabling voice assistants in millions of homes. Voice as-sistants rely on a cloud-based architecture, since data has to be sent back and forth to centralized data centers. A smart speaker is relatively simple by design, which means most of the computing and artificial intelligence processing happens in the cloud and not in the device itself.

The basic idea is that the user makes a request through the voice-activated device, and then, the voice request gets streamed through the cloud, and here voice gets converted into text. Then, the text request goes to the backend and after processing, the backend replies with a text response. Finally, the text response goes through the cloud and gets transformed into voice, which will be streamed back to the user. Most smart speakers come without a screen although there are smart speak-ers with screens such as the Amazon Echo Show and Echo Spot, the Facebook Portal, and the Google Home Hub.

Interesting findings came from Sciuto et al. (2018), where authors explored how households incorporate conversational agents into their lives. Specifically, authors ana-lyzed the logs of 75 Alexa users, for a total of 278,654 voice commands. Participants who have owned an Alexa device for at least six months, answered survey questions related to their household use of Alexa. Of the 75 participants, 26 reported having chil-dren although data from the log files did not provide any insights into which household member gave each command. Parents that were interviewed, positively recalled their children successfully interacting with Alexa even before interacting with smartphones and other technology devices.

VI. PROCEDURE FOR PAPER SUBMISSION

A. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns.

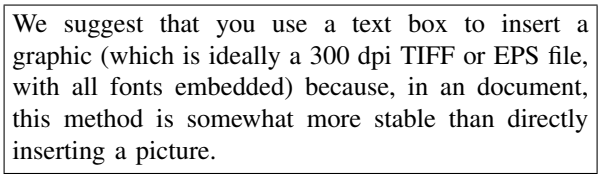

Fig. 1. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

VII. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an e

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] Nan Zhang, Xianghang Mi, Xuan Feng. Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems. https://wiki.aalto.fi/download/attachments/116657996/IoT-attestation.pdf. date accessed: 29.04.2021

[2] George Terzopoulos, Maya Satratzemi. Voice Assistants and Smart Speakers in Everyday Life and in Education. https://files.eric.ed.gov/fulltext/EJ1267812.pdf. date accessed: 01.05.2021