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. 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. Voice squatting attack

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.

II. INTRODUCTION

This template provides authors with most of the formatting specifications needed for preparing electronic versions of their papers.

III. PROCEDURE FOR PAPER SUBMISSION

A. Figures and Tables

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

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

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

IV. 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