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*— Smart speakers with voice assistants achieved last years impressive results in speech recognition enabling more seamless interactions between user and a machine but also raise privacy concerns due to their continuously listening microphones. A better understanding of these aspects can help future smart speaker users to make a right decision about the digitalisation of their homes. For these purposes this paper contains as well a result of research about the functionality of digital voice assistance and the real use cases how people are tending to use a device as the research of actual security and privacy concerns including attack surfaces and vulnerabilities.**

I. INTRODUCTION

The development of the Deep Learning algorithms and Internet of Things last years is opening up a new era in the use of the digital tools that surround us. A combination of various algorithms in Machine Learning, Deep Learning, speech synthesis, and Natural Language Processing (NLP) to providing services to the users makes it possible to achieve impressive results in speech recognition enabling more seamless interactions between user and a machine. It is realistic now to say that in the coming years digital voice assistances probably will come into the use of every household. They could become embedded in users' day-to-day routines, particularly people in a dependent situation, whether elderly or disabled.

However, these undeniable advances should not obscure the questions that voice assistants raise from a data protection perspective, in particular from the point of view of transparency in the way their system functions [9]. In the survey by Lau et al. [5] smart speaker users and non-users were interviewed to find out their arguments for and against adopting this new technology and their privacy perceptions and concerns. Many non-users believe that these devices are not useful at all and companies are not to be trusted. On the other hand, smart speaker users have fewer privacy concerns and rely on companies to safeguard their personal data which think are not interesting to others [2]. The goal of this research is reality check of privacy concerns and myths circulating about voice assistants and the abilities that they are assumed to have. This paper presents the closer look at digital voice assistance functions for a clearer understanding of the logic behind these systems and the security questions they raise for their users.

II. DIGITAL VOICE ASSISTANCES

A. What is Voice Assistance?

Over the last years a very significant progress in the development of digital voice assistance is made with various factors contributing to this: improved methods, a significant increase in computing capacity and greater volumes of data available. This is enabling voice assistants in millions of homes today. In this connection a new report from Juniper Research has found that consumers will interact with voice assistants on over 8.4 billion devices by 2024; overtaking the world's population and growing 113% compared to the 4.2 billion devices expected to be in use by year end 2020 [1].

Even if someone as an author of this paper is a non-user whatever the reason is, like privacy concerns or confidence that the device is useless, the understanding of what is digital voice assistance is and how it works is important regarding the rising amount of devices. It could be easy imagined that a best friend who oft invites to have a dinner at his place has bought a digital voice assistance or got it as a present on Christmas and even forgot to tell that there is a constantly listening device in their home.

What is actually a digital voice assistance? A voice assistant is a complex system consisting of several modules to perform different tasks. From a hardware side there are embodied speaker with microphones and some computing capabilities (more or less developed depending on the case). As smart speaker is relatively simple by design and small be size most of the computing and artificial intelligence processing happens in the cloud and not in the device itself [2]. Because data has to be sent back and forth to centralized data centers a user has to make a request first through the voice-activated device, and then, the voice request gets streamed through the cloud, and here voice gets converted into text [2]. From a software side there are many modern algorithms used on the backend for processing the user request and implementing human-machine interaction as such and which includes built-in modules for automatic speech recognition, natural language comprehension and generation, dialogue and speech synthesis. As was mentioned above in many cases it is done remotely using cloud-based architecture. After the request processing on the backend a text response will be generated. Finally, the text response goes through the cloud and gets transformed into voice using speaker of the physical

[1] https://www.juniperresearch.com/press/number-of-voice-assistant-devices-in-use

digital device and streamed back to the user. After that the voice assistant returns to standby and is constantly listening again to hear a specific wake word uttered by a user ("Alexa" is the default in the case of the Echo Dot, and "OK Google" in the case of Google Home) with no need for activation by pressing buttons or doing anything else. The word or phrase is detected locally on the device, and only once it is matched is a recording made and sent back to the Amazon or Google servers, although a tiny fraction of sound from just before when the matched keyword is said is also sent back [13]. The text of the response that was sent to a user is stored by the voice assistant system so that users of personal devices can review past answers using their application.

B. Audio processing in the cloud

The main power of digital voice assistance is that after being activated through the trigger they can access all the intelligence and computational power in the backend. Device uses on-device technology to detect when the wake word is spoken and then turn on the audio stream to the cloud with a backend. While the microphone is active and the VA's system is processing the request, the user is notified that streaming is occurring by a visual signal (such as a light), an audible signal, or both. When the interaction is complete, no audio is processed by the device and sent to the VA's cloud.

An appropriate response to the user's request is identified and, if necessary, remote resources are used. They can be publicly accessible knowledge database (online encyclopaedia, etc.) or resources accessed by authentication (bank account, music application, customer account for online purchase, etc.) [9].

The fact that a device is constantly listening with no need for activation by pressing buttons or doing anything else raises privacy concerns. All the popular devices have a hardware button that allows you to mute the microphone. However, this does mean that when you want to use the voice assistant again you will have to physically unmute the device, which somewhat defeats the purpose of voice activation [13]. Manufacturers state that before the trigger was said no audio processing and sending to the cloud should be happened. But some usage patterns open the back door for accidental recordings as will be discussed in more detail in the chapter Security and Privacy Concerns. For example, if user says, "Alexa, set the timer!" Alexa will respond with "Timer for how long?" and will open the audio stream to wait for user's response. If a user assumes that the timer will be set to some default value or last settings will be used, they could not proceed with a response that device can understand and a device will continue to listen and record. After 6 seconds for Alexa and one additional reprompt from Google and follow-up 8-second waiting, the running request will be forcefully terminated by the device [1].

After request phrases are processed they will be stored in the cloud to respond to the user's requests on subsequent repeated calls to improve the user's experience so a system can better understand user's requests. All user's request made to their device will be used by a company to train a speech recognition and natural language understanding using machine learning algorithms. Thus the backend system of digital voice assistance is a non-disclosure bunch a self-learning algorithms and a training process with real world requests from a diverse range of customers is necessary for a system to respond properly to the variation in users' speech patterns, dialects, accents, and vocabulary and the acoustic environments where customers use their devices. Moreover it is possible for a user to review their records and even correct them thus such training relies in part on supervised machine learning where humans review an extremely small sample of requests to help to understand the correct interpretation of a request and provide the appropriate response in the future.

Amazon and Google give users multiple ways to manage their data. As stated above, audio recordings are used to improve system services. For personal devices, customers can review voice recordings associated with their account and delete those voice recordings one by one or all at once by visiting their settings page. Users were enabled to review and delete prior voice interactions with the device if they feel uncomfortable or not want companies to keep particular voice recordings on their servers [5]. However in a survey by Malkin et al. (2019) [4] it was found that 56% of active users did not know that their recordings were being permanently stored and that they could review them.

C. Third party features

Amazon and Google are two major players in the market of smart speakers with voice-controlled personal assistant capabilities. However the functionality of devices is not limited with those tasks that was developed and written on the device when it was sold to a user. A modern smartphones have a huge market of applications named Google Play Store for an Android phone and App Store for an Apple Iphone. The main goal is that using such markets user must not looking for a new applications somewhere and download a software that potentially has a virus or intended to damage their hardware or brings privacy risks. Amazon and Google has invented Skills and Actions that can be installed from a safe place that meets hight company's standards for privacy, security, and content. Skills are essentially third-party apps offering a variety of services the voice assistant itself does not provide. Examples include Amex, Hands- Free Calling, Nest Thermostat and Walmart [1]. These skills can be conveniently developed with the supports from Amazon and Google, using Alexa Skills Kit and Actions on Google. Indeed, in survey by Nan Z., Xianghang M et al. [1] it was found that up to November 2017, Alexa already has 23,758 skills and Google Assistant has 1,001. The total number of Amazon Alexa skills continues to grow at a steady pace in selected countries. As of January 2021, the skill count for Amazon Alexa has grown to 80,111 in the United States [2]. The growing amount of third party features easy to see on the figure 1 below. Google proposes lower amount of

[2]https://www.statista.com/statistics/917900/selected-countries-amazon-alexa-skill-count

third party features, thus on January 2019, there were 4,253 official Google Actions in the U.S. [3]. Both companies state that customer's personal information (e.g. name, address) are not released to the 3rd-party unless specifically requested to be shared by the customer.

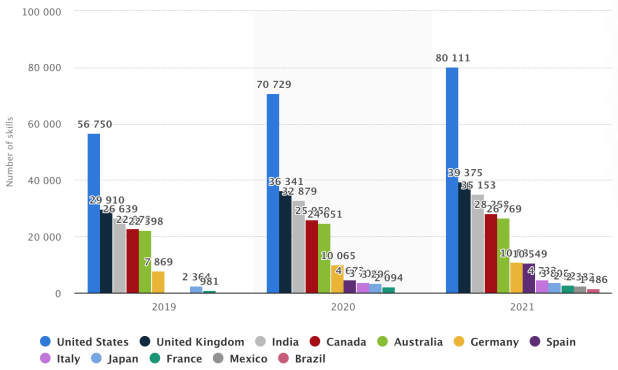

Fig. 1. Total number of Amazon Alexa skills in selected countries as of January 2021. Source www.statista.com/statistics/917900/selected-countries-amazon-alexa-skill-count/

Third party features can be started either explicitly or implicitly. Explicit invocation takes place when a user requires a feature by its name from a digital vice assistance: 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 [1].

Implicit invocation occurs when a user tells the voice assistant to perform some tasks without directly calling to a skill name. When listening device receives a request from a user without a skill name, such as "Alexa, play piano music" Alexa recognizes that no skill name has been specified, selects top candidate skills to fulfill the request, and then queries these skills to determine if skill can fulfill the intent that the customer wanted. If requested skill is not enabled by the user but supports the specific interface for skill understanding, then requested skill may be chosen and auto-enabled to fulfill a query for users who have not enabled requested skill yet [4].

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. When a user invokes a device with its wake-word the almost same procedure as was described above will be started: the device captures user's voice command and sends it to the company'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 cloud server [1]. Using skill's server the requested service will be provided for a user, for example the requested music track will be streamed and played to a user using a music streaming service that is enabled on the digital voice assistance.

III. USE CASES

Voice assistants are not new and researches has been carried out in this field for many years and the knowledge gained in this way is used, for example, for automated telephone calls. Since the launch of Siri in 2011 people have been surprised by the application but many people were not necessarily enthusiastic about the new possibilities. It took a while until Amazon entered the market with Alexa in 2014. Amazon then brought Alexa devices to customers relatively aggressively at very low prices. In 2016, Google finally followed suit with its Voice Assistant. They launched this simultaneously on the Android operating system (mobile phone) and on their first Google Home Smart Speaker. Voice Assistant became a part of everyday life and can be found in many households. In this chapter the typical usage of devices will be examined. As part of the WIK's[5] online survey, respondents were also asked about the usage functions that are typical for assistants. The search function is used most frequently - asking about weather, sporting events or other less complex information, which is mostly retrieved from the Internet. Around 72% of users used this function in 2018. Less often a control of external devices such as lights, heating or stoves are established (20%). Around 40% of users use the call function, control the device on which the voice assistant is installed, set up reminders, appointments or an alarm clock and play music [14]. The survey presented in the CSA-Hadopi report introduces the idea of basic use (see Figure 2) when the majority of items purchased are small and can be used quick and are often some things that someone could buy without necessarily having to see it physically [9]. In chapters below the most important usage patterns will be covered and discussed.

A. Entertainment, music and media

Interviews with 19 participants were conducted to to understand how people use voice assistants [7]. Amazon Alexa and Google Home histories, automatically generated 82 logs and 193,665 commands for Amazon Alexa and 65,499 commands for Google Home were analysed in the research by Lau et al [7]. It was found during the log analysis that playing music was the most common use of Amazon Alexa (at 28.5%) and the second most used command category for Google Home (at 26.1%). As it might be expected, Alexa devices tend to assume that user wants to hear tunes from Amazon Music. If user prefers Spotify, Apple Music, or another music service, it's not all that hard to set it up and use it with their Amazon devices. Same is applied to Google devices that support many music services nowadays: YouTube Music, Apple Music, Spotify, iHeartRadio, TuneIn,

[3]https://voicebot.ai/2019/02/15/google-assistant-actions-total-4253-in-january-2019-up-2-5x-in-past-year-but-7-5-the-total-number-alexa-skills-in-u-s/

[4]https://developer.amazon.com/en-US/docs/alexa/custom-skills/understand-name-free-interaction-for-custom-skills.html

[5]WIK - Wissenschaftliches Institut fr Infrastruktur und Kommunikationsdienste GmbH. https://www.wik.org/

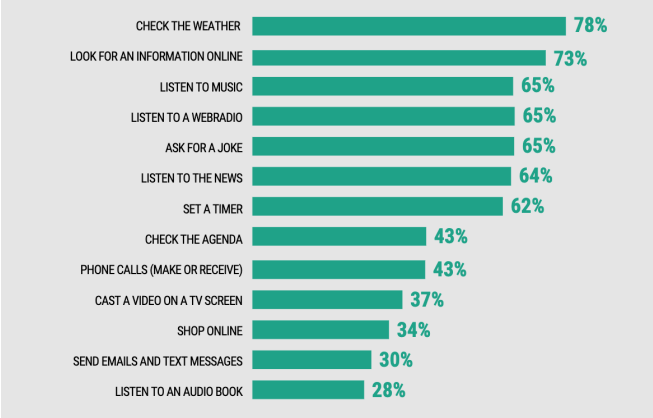

Fig. 2. How are voice assistant-enabled speakers used. Source CSA-Hadopi report, users of smart speakers over the 30 days leading up to the survey, 287 individuals. L'impact de la voix sur l?offre et les usages culturels et mdias, may 2019. https://www.csa.fr

Pandora, Deezer. It mostly depends on user location, which music services they will find in their Google Home app. All these advantages make voice assistances extremely useful for playing music during daily routines such cooking, cleaning or mental work. Users played music based on genre (e.g., classical music), album (e.g., "Load" by Metallica), or artist (e.g., Moby). Most of bluetooth speakers can be connected with a voice assistant and are really popular and easy to use with no wires and little setup.

In research by Lau et al [7] the music search on both voice assistants over the 24-hour time line was aggregated and thus the research findings present specific music command category as a portion of all other commands throughout that period of time. For Amazon Alexa, the music command was used most heavily between 6 and 10 pm, while peaking between 6 and 8 pm [7]. Similar to the Amazon Alexa was found that music for Google Home was used most heavily between 6 pm and 9 pm [7]. This might arise because users are listening to music while preparing meals at the end of the workday. Delicious process of cooking food can be very relaxing if accompanied with a favourite music and the the ability to give a command with a voice, select a favourite song, change the volume or change the album without using hands, which may get dirty during cooking or just be wet, is very convenient. It is interesting that the ability to listen a music may be the deciding factor in choosing the location of the device. One responder in the research by Lau et al [7] told that his wife placed their voice assistant in the kitchen, because she is a music teacher who loves to listen music at home and who loves to cook.

Around 4.9% of Amazon Alexa interactions and 5.9% of Google Home commands were volume related.[7]. Thus these volume related commands could be easily extracted as a separated class because it was found that the ratio of "volume up" to "volume down" commands for Alexa was 37% and 30% for Google Home commands, the authors of research by Lau et al [7] made a suggestion that both Alexa?s and Google Home?s default volume may be set too high.

Interviewees did not limit their voice assistant use to music. Some interviewees indicated that they used their voice assistant to access other media. For example, one used Google Home, along with Google Chromecast o operate their Netflix account connected to their smart TV models [7].

B. Timers and alarms

In research by Lau et al [7] can also be seen that the use of timer command category in both Google Home and Amazon Alexa logs is very popular between 5 and 7 pm. This corresponds to the time users might be cooking dinner at the end of the workday. Users could use timers mostly for cooking purposes. It is assumed to be much easier to set a timer for a spaghetti boiling just with a voice instead of using hands on order to set a mechanical timer on a cooking plate. If several timers, for example for cooking rice and vegetables, were set and when timer's alarm will start to sound it can confuse the user how to distinguish what timer is run out. There is an easy way to keep them all straight. User can name their timers. For example, they can say, "Alexa, set a pizza timer for 10 minutes," and then, "Set a veggies timer for 15 minutes. When the timers are done, Alexa's alarm tone will sound, and she'll say "Your pizza timer is done", followed (in five minutes) by "Your veggies timer is done."

Timers could also be used to set reminders for users. User can also set sleep timers, ask for reminders, check how much time you have left on a timer, and more. It can slowly dim an device-enabled light, or can turn off tunes after a set period of time. User can set a timer to remind them to make a smoothie or to take vitamins. Author of this paper still doesn't have a voice assistant and use for such reminders the mobile application ToDoist[6] and after each execution of a task from the list of tasks the manually checking the box next to the task is required. This often leads to the situation that only the next day, when the list of tasks for the next day is examined it could be noticed that yesterday's tasks were completed, but not marked as completed. Or even something worse for someone who tries to manage their day routines could happen. It could be easily forgotten to complete a task since no active reminders were used and user did not have time to look into the list of tasks. During the research of use cases of voice assistant the author find out that it is possible to manage a Todoist tasks hands-free with a little help from the voice-controlled assistant that turns words into actions. With simple voice commands, assistant can be asked to add new tasks to Todoist, update shopping list, or read out the tasks on lists. When it's time for a reminder, voice assistant will produce the beeping alarm and say (for example), "I'm reminding you, to take your vitamins". The reminder is repeated one more time in a few seconds, and also push notifications to mobile devices are sent. The finding this ability to user reminders and alarms the author of this paper regards as a real use case of the device in her home in case she would finally decide to buy one. While most interviewees noted that they used voice assistants as Alarms,

[6]https://todoist.com/app

the logs show the Alarm category includes words like “set” as in set the Alarm and “snooze” when snoozing the alarm when triggered [7].

C. Search and source of information

Users of smart speakers usually ask for music 70% of the time, about the weather 64% of the time, and fun questions 53%. It is expected that by 2021, 50% of all searches will be voice-activated. Google?s voice searches demographic statistics show that 27% of the global population with access to the internet use voice searches. In the US, voice assistants are used by over 111 million people[7].

When user purchased a Google Home device, it could be expected that they are going to be using Google’s search engine to answer all their questions. But, with Alexa being a large competitor of Google in certain spaces, it’s to assume that they aren’t going to play well together. Alexa utilizes Bing’s search engine for all of her search queries. Google isn’t directly available with any Amazon approved Alexa skills. There is, however, a way to search Google with Alexa if user uses a workaround skill by a third party. Once set up, user can use the Google Assistant Skill to search Google on their Alexa device. After a needful set up of a favourite search engine user can use it and ask a voice assistant to perform a search and read aloud the results. Search or informational queries was the most prevalent use of Google Home (at 26%) and second most prevalent use for Amazon Alexa (at 19.4%) [7]. The frequency of search command use was highest for both Amazon Alexa and Google Home was between 5 and 7 pm followed by the time between 8 am and noon [7]. One of the most popular terms was “song” for both Amazon Alexa and Google Home. Users used the search command to ask questions about music they listened to, specifically the name of a song they are listening to, or the name of the artist singing a particular song, etc. [7]

The significant change that voice search brings is to the results page. If user uses their smartphone and asks with their voice ?find a pizza near me? the smartphone displays just three listings. A search engine result page on a computer/laptop lists at least ten options. For Google Home and Amazon Echo, Google and Alexa respond with just one result. To reach a first-page placement now is not enough for a success for many companies because of limited results offered by voice search. Also it must be mentioned that Google Home, Amazon Echo, and even Siri deliver answers based on the personal data they collect. This is another move to ensure the answers are relevant and helpful.

Some respondents emphasized the use of the search feature when interacting with family and friends. – random questions, like trivia questions, or like some facts, sports scores or check stock market value. Voice assistant can be used to quick access to a knowledge database when having a dispute with friends. However it changes a way we communicate with our friends: instead of an hour-long discussion trying to find the truth in a dispute, now you can easily get an answer and refute your opponent or confirm your point of view without taking out your mobile phone.

Additionally the family can enjoy plenty of tales and kid-friendly news by asking their device to play a podcast. In the research by Ammari et al. was found that users also prefer to ask their device for a help during cooking process like for converting measurements (how many teaspoons are in a cup) ot for an additional help with some difficult cooking terms [7]. Users also asked about the temperature on that particular time as well as future forecasts, at times asking for a specific day, for example, “Alexa, is it gonna snow two days from now?” [7]. Weather-related requests (39 from a total of 136 times during the four days of study) were the most frequently reported by all age-group partici-pants followed by requests to play music 29 times.

D. Smart Home

E. Usage by children

The studies that explore the usage of smart speakers in homes are mostly focused on young and middle age people who are tending to be the most active users of the modern technologies. Comparing to children and elderly people adults usually make a decision to adopt some new technology on their own. Children normally are allowed (or not) to use an available at home device that was bought by their parents and elderly people are rather tending to get something innovative new as a present from their younger relatives than to buy it on their own.

In the research by Sciuto et al [15] involving children authors explored how households incorporate conversational agents into their lives and the interesting findings came from this research. It was reported by users in the research that they have children although data from the log files did not provide any insights into which household member gave each command, Specifically, authors analyzed the logs of 75 Alexa users, who have owned an Alexa device for at least six months for a total of 278,654 voice commands. Of the 75 participants, 26 reported having children [2]. Parents that were interviewed, positively recalled their children successfully interacting with Alexa even before interacting with smartphones and other technology devices [2]. Such findings rises a security questions since digital voice assistants nowadays don’t recognise voices and can not distinguish a children and an adult giving commands. It as great specificities of the voice such as physically the voice is only a trace left by air movements caused by the phenomenon of phonation, i.e. the production of sounds specific to the spoken language. As a voice assistant doesn’t recognise children and adult voices it literally just being in permanent standby mode can activated and inadvertently record a conversation as soon as device assumes to have detected a wake word. Once recorded, the interactions might be listened to by persons, employees or service providers of the company providing the voice assistant, in order to improve the various algorithms implemented (wake word detection, automatic speech transcription, language comprehension, etc.). Being a mother of the small child an author of this paper can easily

[7]https://review42.com/resources/voice-search-stats/

imagine that a children can say a wake word many times a day just as a part of their imaginary game. Choosing to place such a device at the heart of one's home therefore implies responsibilities towards the various persons whose personal data may be processed.

Interesting are the findings from different research about the nature of interaction of children with a devices. Beirl et al. conducted a research about the home usage of Alexa, in a period of three weeks. Six families with children in the age group up to 13 years old were interviewed [16]. Results showed that children interacted with Alexa with much enthusiasm and natural interest and the shorts conversations became easy part of their family rituals [2]. When a more competent family member helped a younger member interact with Alexa the interaction continued with more encouragement and interest. Children behaviour is investigated also by Druga et al. where 26 participants (3-10 years old) interacted with 4 voice assistants, Amazon Alexa, Google Home, Cozmo, and Julie Chatbot. Children enjoyed interaction with voice assistants, while older children perceived their intelligence and thought they could learn from them. The main issue of the interaction with children was getting the assistants to understand their questions although with the help of facilitators and parents, children altered their strategy and became fluent in voice interaction [3]. None of the children expressed suspicion or inquired about how the system worked. After reviewing the logs and audio recordings of all the participants, authors came to the conclusion, that children preferred personified interfaces rather than non-personified and that age played an important role in children's performance [2]. Older children could get the answer that they needed using less help from provided hints [3]. Since the interaction required children to reformulate questions, most of them needed hints to complete the task. Analysis of the results of conversations of children aged 5 to 6 showed that 89% of children's questions were transcribed correctly, although only 50% of children's questions received a full answer [2]. Children and their parents reported that the provided answers were long or required interpretation. Most children's questions were about the world around them and they believed that the device is a source of information.

Voice assistants became quickly a digital interface particularly appreciated by children for its (relative) ease of use. While there is no doubt that a computer or smartphone should not be left in the hands of a young child without parental supervision, it is essential to note that the same is true for voice interfaces [9].

IV. VULNERABILITIES

A. Security and Privacy Concerns

B. Accidental Recordings

V. VOICE-BASED REMOTE ATTACKS

1) Voice squatting attack (VSA):

2) Voice Masquerading Attack (VMA):

VI. COPY AND PASTE CORE IDEAS

A. 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

B. 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]

Security and Privacy Concerns [2]:

Privacy is also considered by Hoy (2018) since anyone with access to a voice-activated device can ask it questions, gather information about the accounts and services associated with the device, and ask it to perform tasks. As stated by Horn (2018), since Voice Assistants and Smart Speakers can distinct children?s voices, their specific learning needs are certain to raise questions with the Children?s Online Privacy Protection Act (COPPA) and the Family Educational Rights and Privacy Act (FERPA).

In a survey by Malkin et al. (2019) [4] sample group was approximately gender-balanced,with 44.0% self-identifying as female, and the median reported age was 34. Households of 2 or more accounted for 83.6% of all participants, with a median household size of 3. Regarding 116 smart speaker owners? beliefs, attitudes, and concerns about the recordings that are made and shared by their devices (Amazon and Google), it was found that 56% of them did not know that their recordings were being permanently stored and that they

could review them. It was observed, that while participants did not consider their own recordings important or sensitive, they were more protective of others? recordings, such as children and guests, and were strongly opposed to the use of their data by third parties. Only 3% of the participants review their recordings and deleted them. Additionally, only 5% of the participants used the mute button on their device while only 4% unplugged their device in order to stop listening. The survey concludes that privacy controls are underutilized[2]

Regarding RQ1, early findings from a small number of studies, show that adults (Purington et al., 2017; Sciuto et al., 2018; McLean and Osei-Frimpong, 2019; Rzepka, 2019; Song, 2019; Bunyard, 2019) use voice assistants for entertainment purposes, seek-) use voice assistants for entertainment purposes, seek-ing information, making purchases and listening to music. Adults also enjoy interaction and find voice assistants easy to use. Studies involving children (Sciuto et al., 2018; Beirl et al., 2019; Druga et al., 2017; Yuan et al., 2019; Lovato et al., 2019) conclude that children can interact with smart speakers and voice assistants, making basic requests.[2]

To detect when a user makes a request, a smart speaker?s multiple microphones continuously listen for the device?sactivation keyword (e.g., ?Alexa? or ?Hey Google?). The smart speaker responds to a request through virtual or physical actions and audio feedback [5]. Current smart speakers are equipped with some privacy features. While the device?s microphonesare always listening, speech recognition is performed locally by the device until the activation-keyword has been detected, at which point the subsequent voice command is forwarded to themaker?s servers for processing. In addition, most smart speakers are equipped with a physicalbutton to mute the microphones. Companion mobile apps and websites enable users to reviewand delete prior voice interactions with the device should they feel uncomfortable or not wantcompanies to keep particular voice recordings on their servers.[2]

user participants included mostly primary users [2]? people who have set up their smart speakerthemselves and connected it to their own accounts. Users mostly placed their smart speakers incentral locations in their homes to maximize utility. Since some rooms in people?s homes are more private than others, we investigated if users considered this when placing their speakers. Smart speakers were commonly placed in central locations in participants? homes, sometimes on dedicated tables situated at intersection points of multiple rooms. In study made by Lau et al. (2018a) [5] participants reported various uses for their speakers, which were also reflected in their diary responses (frequencies based on diary responses): music playback (14),checking the weather (13), probing smart speaker capabilities (11), controlling other devices (8),asking knowledge questions (6), setting timers (6), checking the time (4), checking the news (3),setting alarms (3), setting reminders (3), checking their calendar (2), purchasing items online (2),playing games (1), and listening to the radio (1). Some utilized multiple different features of their smart speakers, but many stated they used their smart speakers almost exclusively to play music. Seven users controlled other IoT devices in their home with the smart speaker, such as lights,thermostats, and home security systems. This required users to purchase other, potentially expensive,IoT devices that can be linked with smart speakers [2]

This is a source [9] When the user utters the wake word, the assistant "wakes up". A listening channel opens and the audio content is streamed. Processing (NLP) technology, speech is interpreted. Automatic speech recognition used to involve three distinct steps to: 1) determine which phonemes had been pronounced using an acoustic model; 2) determine which words were pronounced using a phonetic dictionary; 3) transcribe the sequence of words (sentence) most likely to have been spoken using a language model. Today, with the progress made possible by deep learning (a machine learning technique), a large number of systems offer automatic transcription of speech from end to end.

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. bla

REFERENCES

[1] Nan Z., Xianghang M., Xuan F. 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] Terzopoulos G., Satratzemi M.. Voice Assistants and Smart Speakers in Everyday Life and in Education. https://files.eric.ed.gov/fulltext/EJ1267812.pdf. date accessed: 01.05.2021

[3] Druga, S., Williams, R., Breazeal, C., Resnick, M. Hey Google is it OK if I eat you?: Initial explorations in child-agent interaction. https://dl.acm.org/doi/pdf/10.1145/3078072.3084330. date accessed: 02.05.2021

[4] Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., Wagner, D. Privacy Attitudes of Smart Speaker Users. https://www.researchgate.net/publication/336184996_Privacy_Attitudes_of_Smart_Speaker_Users. date accessed: 02.05.2021

[5] Lau, J., Zimmerman, B., Schaub, F. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. https://www.key4biz.it/wp-content/uploads/2018/11/cscw102-lau-1.pdf. date accessed: 02.05.2021

[6] Lau, J., Zimmerman, B., Schaub, F. Alexa, Stop Recording! : Mismatches between Smart Speaker Privacy Controls and User Needs. https://www.usenix.org/sites/default/files/soups2018posters-lau.pdf. date accessed: 03.05.2021

[7] Ammari T., Kaye J., Tsai J.T., Bentley F. Music, search and IoT: How people (really) use voice assistants. https://www.researchgate.net/publication/332745214_Music_Search_and_IoT_How_People_Really_Use_Voice_Assistants. date accessed: 03.05.2021

[8] Amazon.com. Alexa features. https://www.amazon.com/b?ie=UTF8&node=21576558011. date accessed: 03.05.2021

[9] Commission Nationale de l'Informatique et des Liberts. White Paper Collection. Exploring the ethical, technical and legal issues of voice assistants. https://www.cnil.fr/sites/default/files/atoms/files/cnil_white-paper-on_the_record.pdf. date accessed: 27.04.2021

[10] Lei X., Tu L.,Liu A.X., Chi-Yu Li. The Insecurity of Home Digital Voice Assistants: Vulnerabilities, Attacks and Countermeasures. https://ieeexplore.ieee.org/abstract/document/8433167. date accessed: 03.05.2021

[11] Ren J., Dubois D. J., Choffnes D., Mandalar A. M. Information Exposure From Consumer IoT Devices. https://www.researchgate.net/publication/336657694_Information_Exposure_From_Consumer_IoT_Devices_A_Multidimensional_Network-Informed_Measurement_Approach. date accessed: 03.05.2021

[12] Knote R., Janson A., Eigenbrod L., Sllner M.. The What and How of Smart Personal Assistants: Principles and Application Domainsfor IS Research. https://mkwi2018.leuphana.de/wp-content/uploads/MKWI_285.pdf. date accessed: 03.05.2021

[13] Wueest C.. An ISTR Special Report. A guide to the security of voice-activated smart speakers. https://docs.broadcom.com/doc/istr-security-voice-activated-smart-speakers-en. date accessed: 08.05.2021

[14] Tas S., Hildebrandt C., Arnold R. Sprachassistenten in Deutschland. https://www.researchgate.net/publication/334597267_Sprachassistenten_in_Deutschland. date accessed: 24.05.2021

[15] Sciuto A., Saini A., Forlizzi J., Hong J. Hey Alexa, What's Up?: A Mixed-Methods Studies of In-Home Conversational Agent Usage. https://www.researchgate.net/publication/325704495_Hey_Alexa_What's_Up_A_Mixed-Methods_Studies_of_In-Home_Conversational_Agent_Usage date accessed: 25.05.2021

[16] Beirl D.,,Rogers Y. Using Voice Assistant Skills in Family Life https://discovery.ucl.ac.uk/id/eprint/10084820/1/Using%20Voice%20Assistant%20skills%20in%20Fam%20life_Beirl,%20Rogers,%20Yuill.pdf date accessed: 25.05.2021