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.

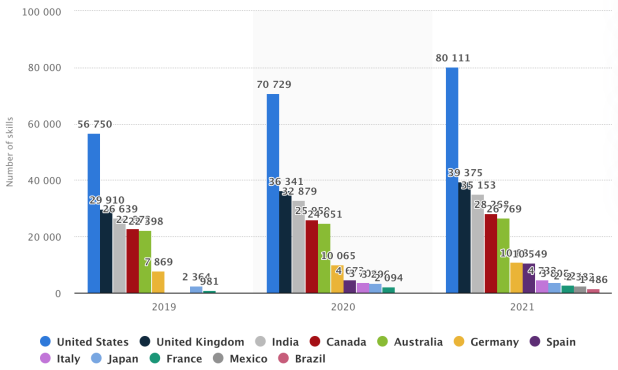

Fig. 1. Total number of Amazon Alexa skills in selected countries as of January 2021. Source www.statista.com/statistics/917900/selected-countries-amazon-alexa-skill-count/

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/

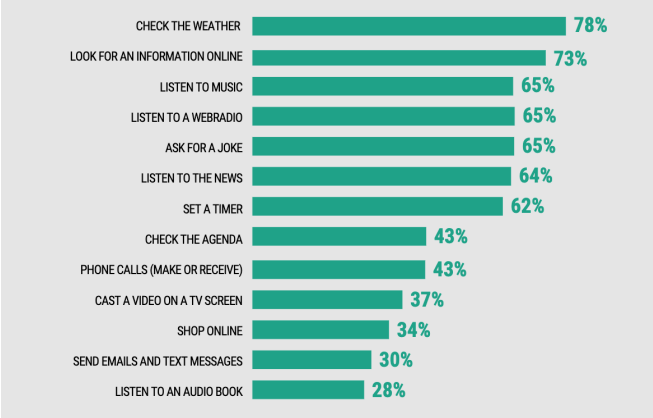

Fig. 2. How are voice assistant-enabled speakers used. Source CSA-Hadopi report, users of smart speakers over the 30 days leading up to the survey, 287 individuals. L'impact de la voix sur l?offre et les usages culturels et mdias, may 2019. https://www.csa.fr

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

It was found in the research by Ammari et al. that voice assistants have been used to control IoT devices in different parts of the house like kitchen, bedroom, living room etc.. Mostly they have been used to turn lights on and off. It can happen because there is still no well developed market for a connected devices with a lot of competition and choice of not overpriced devices. In the last few years it has started to change, for example kettles are getting connected. If user would like to try to living in a smart home, and control their lights and other fittings with just their phone or voice, they can also get a brew boiling without having to stand up and do it themselves. Ammari et al mentioned that the value of voice assistants and IoT devices around the home of users depends on the form of home ownership and daily routines of a specific user that could be automated. Respondents in the research noted that they would be more willing to install more IoT devices if they owned the house since they thought making their domicile smarter added to its value. For example, in the research it was mentioned that an user plans to buy a Nestthermostat to use it along with Google Home when they ?become a homeowner.?. Despite the fact that a smart thermostat will save money and energy even in a rent house, users are more likely ready to invest money in purchasing the smart home devices only if it adds a value to their property [7].

Integrating problems were mentioned by interviewed people. It was found that an average user probably underestimated how complicate can be a connection process. The integration between voice assistant and other devices was mentioned as not sufficient when for connecting IoT devices from different manufacturers rises the need to use a smart hub to connect the different devices to a chosen voice assistant. The luck of the context understand when giving a command to a voice assistant was also mentioned. What is normal for a human iteration seems to be difficult for an artificial intelligence, for example device could better interpret user’s commands with relation to their location in the house at the time of issuing the command to a device. So that when user is in the living room and they ask their

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

assistant to shut down the lamp, they want to shut down the lights in the living room [7].

A fact the someone have purchased a voice assistant leads to the new investments in smart home appliance. Responders in the research by Ammari et al. mentioned that after buying a voice assistant they suddenly realised that the device itself is probably just a very expensive clock with an ability to play music via built-in speaker and tell a weather forecast. Users wanted rather minimise their costs to home automation by installing the cheaper smart switches then just purchasing additionally expensive smart lights like the Hue. As users installed more IoT devices, the need for more voice assistants in different parts of the house arose. This fact should be considered more careful by users when deciding to purchase a voice assistant since it is brings the new additional costs and need to purchase a new IoT devices. Even the size of a house is decisive for the purchase of additional equipment including additional voice assistants.

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 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].

F. PRIVACY CONCERNS

As was mentioned above in the survey by Lau et al. [5] smart speaker users and non-users were interviewed and it was found that smart speaker users have fewer privacy concerns and either think the company that developed a smart assistant can be trusted and they will protect their privacy

and data; or think they have nothing to hide and third party companies have no real interest to their personal data.

However the security and privacy concerns regarding using the voice assistants should be seriously considered by a new potential and existing users and probably requires new studies as new features and versions will be developed and released to a market. Anyone with access to a voice-activated device can utter a wake-up word, ask it questions, gather information about the accounts and services associated with the device, and ask it to perform tasks. It is claimed that smart speakers can distinct children's voices but in the research by Sciuto et al [15] was found that the log files did not provide any insights into which household member gave each command. Interviewed parents clearly remember as their children successfully interacting with Alexa and ask her questions but these responds can not be extracted from the log data. Since the voice assistant can not distinguish the children voices it arises questions with the children's online privacy protection. Children's specific learning needs must be considered, access to the skills installed by parents (such a bank account) must be limited if a child gives a command; age rating and parents guide must be used or playing music as songs can contain offensive words, aggression, sexual texts or suicidal romantic.

A smart speaker's multiple microphones continuously listen for the device's activation keyword (e.g., "Alexa" or "Hey Google") in order to detect when a user makes a request. The smart speaker responds to a request through actions and audio feedback [5]. From security reasons speech recognition is performed locally by the device until the activation keyword has been detected. Most smart speakers are equipped with a physical button to mute the microphones but only 5% of the participants used the mute button on their device while only 4% unplugged their device in order to stop listening [2]. When the user utters the wake word, the assistant "wakes up". A listening channel opens and the audio content is streamed to the cloud or to servers of third party. Remote cloud using deep learning and artificial intelligence will try to correct determine which words were pronounced using a phonetic dictionary and then to create the sequence of words as c complete sentence most likely to have been spoken using a language model.

Responders concerned that to not knowing whether their device is listening when they did not want it to listen. Some of them reported physically unplugged their device when discussing financial issues because they did not trust that device would not be listening if it were muted. In the research by Malkin et al. only 18.6% of respondents described taking any steps to limit their devices and most commonly (43% of respondents who took privacy actions, 7.8% of all participants), users described turning off the microphone [4]. It was found that when not muted, Alexa sometimes does interact with the Amazon service, even when a wake word was not used. In 2018 a lot of cases about Alexa's random laugh were reported that was freaking people out. It turns out that in rare circumstances, Alexa can mistakenly hear the phrase "Alexa, laugh" even when that's not what was said. Alexa then interprets the phrase as a command and laughs. After that the phrase necessary to make Alexa laugh was changed by Amazon to "Alexa, can you laugh?" which less likely can generate false positives. A trigger "Alexa, laugh" was disabled. In addition, Alexa does not longer respond to that question with simple laughter but instead will say, "Sure, I can laugh" followed by laughter.

It is not the only one example of accidental recordings. Because the natural language processing even today is not perfect the mistakes occur regular, sometimes with drastic consequences. In an emailed statement to The Washington Post[8] Amazon said in May 2018 that the Echo sometimes can misunderstand pronounced word and think it sounded like "Alexa." In come weird circumstances it can not only wake up, record the following conversation, it can even then silently sent recordings to the person from contact list without the owners?s permission. The occurred accidental recordings is one of the major privacy concerns with smart speakers for today. The entire conversations already being emailed to random contacts. Since it is possible to review all records, done by voice assistant, the participants of Malkin et al. research reported that 1.72% and 2.93% of all 53.5% recordings were television/radio/music or just noise, respectively [4]. Respondents said that the speaker was not addressing the device 6.33% of the time. Thus, over 10% of the recordings in our study were unintentional.

It is worth to be mentioned, that it is impossible to change the Alexa's name. User can choose from Alexa, Amazon, Computer, or Echo. One participant in the research by Malkin et al. provided an example of how this leads to accidental recordings done by Alexa: "I have a friend also named Alexa who comes over, and Amazon Echo thinks we are giving it commands" [4]. The author of this paper is named Alexandra (Oleksandra in Ukrainian) and usually introduces herself as "Alex" despite that fact that this name more often given to boys. The one of the reasons is similarity with the name of the device that it is impossible to change thus most of the users of device (and potential interlocutors of the author) are using this name every day to give it commands.

Additionally, findings suggest that voice interactions elevated the feelings of having a social conversation between the user and the voice assistant and this led to positive evaluation toward the voice assistant. A study[9] published earlier this year in the Journal of Social and Personal Relationships suggests that it takes 50 hours with a person to consider that person a casual friend. The study further reports that it takes 200 hours to make a close friend. It is assumed that Alexa will soon have the capability to hold a 20-minute conversation. These all can lead to the fact the some people especially social isolated older adults who do not see their friends or family for most of days can consider their voice assistant is kinda a digital friend and can talk much more

[8]https://www.washingtonpost.com/news/the-switch/wp/2018/05/24/an-amazon-echo-recorded-a-familys-conversation-then-sent-it-to-a-random-person-in-their-contacts-report-says/

[9]http://journals.sagepub.com/doi/abs/1.1177/0265407518761225?journalCode=spra

privately about their lives and provide too much additional information and personal data. All these conversations thus will be recorded and stored as user used a device's activation keyword in order to get a feedback from a device.

IV. SECURITY CONCERNS

In a survey by Malkin et al. [4] sample group was gender-balanced with the median reported age of 34. 83.6% of all participants households have 2 or more people using voice assistant so that a median household size was reported to be 3 people. It included mostly primary users who have set up their smart speaker themselves and connected it to their own accounts. 56% of smart speaker owners did not know that their recordings were being permanently stored and that they could review them. Only 3% of the participants review their recordings and deleted them [2]. This finding echos findings by by Lau et al. [5] indicating that while users might know of the logs, they might find accessing and editing them too complicate. Devices was placed mostly in central locations in their homes to maximize utility sometimes on dedicated tables situated at intersection points of multiple rooms. So that device can hear the "wake up" word and further command easily as well as everything the household participants and their guests will discuss every day. Since some rooms in people's homes are more private than others (eg. bedroom), it was found be researches by Lau et al. [5], [6] that this fact was not considered when placing their speakers. They were placed not only on the middle of the apartment so that all rooms can be observed, but also in bedrooms so that user might control a light in their bedroom, sleep sounds and others tasks. Voice assistants share the same physical space inside the house and can be used by every household member but they are not designed for multiple users environments with different privacy needs, they don't have a profiles for different rooms of human houses and they don't recognise person who a talking to them right now. For example, in the bedroom device could be muted automatically after 9 pm until the morning alarm. The placing of device must be also considered not only from the point of view of personal security (what can be heart when user talks) but also from the point of view of security of the device itself. Keeping a smart speaker away from all the windows in house and from all outside doors must be a rule. The location could potentially give anyone from the outside access to device, and they could have access to your other smart home devices. For example, if a door locked with a smart lock, someone outside the apartment can command to open the door simple using mail slot in the door. Placing a device under the TV can trigger a device by with a phrase "Alexa, open.." coming from an talk show or an advertisement and thus device starts recording what's said after the command[10].

One more issue arises when user command with a half open phrase, for example, if user says, "Alexa, set the reminder!" Alexa will respond with "Reminder for what?" and will record the audio to recognize a user's response. It will last at least next 6 seconds for Alexa and one additional reprompt from Google and follow-up 8-second waiting till the running request will be forcefully terminated by the device. All this time everything was told near the device will be recorded and sent into the cloud and/or third party servers.

V. VOICE-BASED REMOTE ATTACKS

As was mentioned in section II-C voice assistant devices are not doing only those tasks they were delivered with. The very interesting and powerful feature is the ability to install voice-driven capabilities from thirds party bringing products and services to life at own voice assistant. Today anyone can publish their skills through Amazon and Apple market[11]. For example, to develop any type of Alexa skill, only Amazon developer account is needed and it is even free. Both individuals and companies can open an Amazon developer account. Developer isn't required to be part of a company to open an account. The research of security risks of voice-controlled third party by Nan et al. [1] discovered that both Amazon and Apple market have only minimum protection in place to regulate the functions submitted: almost nothing from security checking of uploaded content done by Amazon, before authors reported their security leaks findings; and only the basic check is performed on Google to find duplicated invocation names [1].

As skill/action was uploaded to the market it can be invoked by user even if user never downloaded and installed it[1]. If user says a phrase supported by the service in combination with the invocation name for a custom skill to request information, ask a question, or tell to do something this invokes a skill with or without a specific request (intent). Additionally user can indirectly invoke an custom feature through a name-free interaction. In this case, the user asks o perform a task, without naming the skill that should fulfill the request. ("Alexa, what is my horoscope?")[12]. These all means that once a malicious skill is published, it can also be transparently launched by the victim through their voice commands, without being downloaded and installed on their device [1].

It was found by Nan et al. that a name of third party feature is not obligatory unique skill identifier, there are multiple skills with same invocation names are on the Amazon market. For example, Nan et al. pointed that 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 [1]. There are some undisclosed policies about how Alexa will chose what skill will be launched next time when such a common name is spoken. Authors of the paper guessed based on observations in their research that it can be a random choice [1]. Anyway, it's known that longest

[10]https://www.cnet.com/home/smart-home/4-places-to-avoid-putting-your-amazon-echo-in-your-home/

[11]https://developer.amazon.com/en-US/docs/alexa/ask-overviews/what-is-the-alexa-skills-kit.html

[12]https://developer.amazon.com/en-US/docs/alexa/custom-skills/understanding-how-users-invoke-custom-skills.html

string match from user's request is used to find the skill. In sections below the problem will be discussed deeply. This problem is less serious for Google, which does not allow duplicated invocation names [1]. However, it also cannot handle similar names and problem lays in a natural language processing model.

Thus in the research by Nan et al two different types of voice-based attacks were discovered and reported to Amazon and Google. Authors pointed to the fact, up to their knowledge, no protection was in place to defend against the reported threats when reported the issues to Amazon and Google in February 2018. Even today, some months after publication of the report, Amazon can still not defeat VSA (section V-A) and only has limited protection against VMA (section V-B) by detecting empty recordings [1].

A. Voice squatting attack (VSA)

This type of attack based on the nature how voice assistant invokes third party feature in the case several search results fulfil the spoken request. For this the selected third party application that available on market is studied by an attacker with the aim to gain information 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.) [1]. The goal is to develop a a malicious skill with the similar voice command and to cause a system to trigger that malicious skill instead of the one the user intends to use. Since the longest string match used for starting the application, it is easy to add a simple word "please" or"right now" and create a malicious software with the same name as attacked skill ending with "please" or "right now". Nan er al. 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" . Figure 3 summarizes the responses from both Amazon Echo and Google Home users. The results show that more than 85% of them have used natural-language utterances to open a skil nstead of the standard one. This indicates that it is completely realistic for the user to launch a wrong skill whose name is better matched to the utterance than that of the intended skill [1].

	Amazon	Google
Invoke a skill with natural sentences:		
Yes, "open *Sleep Sounds* please"	64%	55%
Yes, "open *Sleep Sounds* for me"	30%	25%
Yes, "open *Sleep Sounds* app"	26%	20%
Yes, "open my *Sleep Sounds*"	29%	20%
Yes, "open the *Sleep Sounds*"	20%	14%
Yes, "play some *Sleep Sounds*"	42%	35%
Yes, "tell me a *Cat Facts*"	36%	24%
No, "open *Sleep Sounds*"	13%	14%
Other (please specify)	3%	4%
Invoke a skill that did not intend to:		
Yes	29%	27%
No	71%	73%

Fig. 3. Survey responses of Amazon Echo and Google Home users [1].

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]. For example, if user says 'Alexa, Open PayPal please", which normally opens the skill PayPal[13], but can trigger a malicious skill PayPal 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 [1].

For their study researches randomly sampled 100 skills each from Alexa and Google assistant markets [1]. As the attack targets then were randomly sampled ten skills from each skill markets selected before. For each skill Nan et al. built four new skills whose invocation names include the target's name together with the terms identified from their survey study, for example, they added words "please"/"app" at the end of the name or "my"/"the" at the beginning. In one attack, the researchers registered an "attack skill" called "Capital Won" that sounded very similar to the legitimate skill Capital One"[14] They showed that when someone asked for the official skill from an Amazon Echo device, they could get the malicious instead.Another experiment showed the reality with the sequence matching algorithm when searching for the name of pronounced skill. There are more opportunities to trigger an evil skill. The adversary who aims at Capital One could register a skill Capital Won, Capitol One, or Captain One. For this, they created an evil skill dubbed "Capital One Please". By strategically placing the word "please" in the application's title the they received succeeded launches of their "fake" skill just because the part of their responders talk to Alexa politely in a natural manner. As it was mentioned above in section III-F the feelings of having a social conversation often arises after some long time being talking to voice assistant, especially by elderly people. Similar techniques were successfully used on Google Home. From the responses in research by Nan et al. was 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]. Additionally, during the study it was 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 [1].

Summarizing the experiment, Nan et al. said that on Alexa, an attack skill with the extended name 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 [1]. Finally,

[13]https://www.amazon.com/-/de/dp/B075764QCX/

[14]https://www.capitalone.com/applications/alexa/

the researchers uploaded four skills to Amazon and one to Google that sounded so similar to legitimate services they could launch instead of the real ones, but they contained no malicious functionality. For the average test participant, they'd launch "fake" skills more than 50% of the time when they tried to call the proper ones, according to a paper published.

The consequences of misusing the weakness by intruder could be very dramatic. The researches claimed that through voice squatting, the attack skill can impersonate another skill and fake the device to collect the information the user only shares with the target skill [1]. This could cause serious leaks to untrusted parties since some Amazon and Google skills request private information from the user to continue proceed the user'r request. Thinking being interacting with an official application user is expected to be asking for their phone number, home address or email and will reveal the personal information to a malicious application. Even more, erroneously invoked skill can also perform a Phishing attack by delivering misleading information through the voice channel to the user e.g., fake customer contact number or website address, when impersonating a reputable one, such as Capital One [1].

B. Voice Masquerading Attack (VMA)

TODO

Sapana navaga

VI. CONCLUSIONS

In this research the most popular voice assistant systems named Alexa and Google Assistant analysed, focusing on the third-party skills deployed to these devices. The most common use cases were clarified and the patterns of usage were discussed. It was found that the owners of smart speakers have less privacy and security concerns about having a listening device in their homes as people who still don't own such device. Interesting are the findings about the time of the day when people give commands to launch their favourite music and places they chosen to place a device in their homes. The usage by children was also touched in this paper since children tend to use new technologies available at home very easy in natural manner learning by doing.

The core part of this paper is the research about security concerns and privacy leaks. Accidental recordings happened during using the voice assistant are mentioned in the work. The problem with placement of device in a human house also discussed. So that, placing the device near windows or external door open potentially a chance for an intruder to perform actions on device such as to open a smart lock, connected to a device, or transfer money from a bank account, linked to a device.

The last part of the work goes deeper into specific voice-controlled attacks based on third party features. By matching the longest string from user's request the malicious skill can be launched even it was not enabled and downloaded by a user. It was found that an adversary can intentionally induce confusion by using the name or similar one of a target skill, to invoke an attack skill when trying to open the target after user gave a command.

All these findings show that despite the fact of high and rising popularity of voice assistants and the fact that they are developed and controlled by a world-known leading companies, the device installed inside the human house can still lead to privacy leaks and security risks. Probably, it should be considered a good manner to report all persons/guests who could enter private area about the presence of continuously listening digital assistant, as it must be done today with traditional video surveillance.

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