# Smart Home Personal Assistants: A Security and Privacy Review

JIDE S. EDU, JOSE M. SUCH, and GUILLERMO SUAREZ-TANGIL, King's College London, UK

Smart Home Personal Assistants (SPA) are an emerging innovation that is changing the means by which home users interact with technology. However, several elements expose these systems to various risks: (i) the open nature of the voice channel they use, (ii) the complexity of their architecture, (iii) the AI features they rely on, and (iv) their use of a wide range of underlying technologies. This article presents an in-depth review of SPA's security and privacy issues, categorizing the most important attack vectors and their countermeasures. Based on this, we discuss open research challenges that can help steer the community to tackle and address current security and privacy issues in SPA. One of our key findings is that even though the attack surface of SPA is conspicuously broad and there has been a significant amount of recent research efforts in this area, research has so far focused on a small part of the attack surface, particularly on issues related to the interaction between the user and the SPA devices. To the best of our knowledge, this is the first article to conduct such a comprehensive review and characterization of the security and privacy issues and countermeasures of SPA.

CCS Concepts: • **Security and privacy → Systems security**; **Security requirements**; **Usability in security and privacy**; Social aspects of security and privacy; • **Human-centered computing → Natural language interfaces**;

Additional Key Words and Phrases: Smart home personal assistants, security and privacy, voice assistants, smart home, Amazon Echo/Alexa, Google Home/assistant, Apple Home Pod/Siri, Microsoft Home Speaker/Cortana

**ACM Reference format:**
Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2020. Smart Home Personal Assistants: A Security and Privacy Review. *ACM Comput. Surv.* 53, 6, Article 116 (December 2020), 36 pages.
https://doi.org/10.1145/3412383



## 1 INTRODUCTION

Human-computer interaction (HCI) has traditionally been conducted in the form of different types of peripheral devices such as the keyboard, mouse, and most recently tactile screens. This has been so because computing devices could not decode the meaning of our word, let alone understand our intent. Over the past few years, however, the paradigm has shifted, as we witnessed the rapid development of voice technology in many computing applications. Since voice is one of the most

The first author thanks the Federal Government of Nigeria through Petroleum Technology Development Fund (PTDF) for funding his Ph.D. at King's College London.
Authors' addresses: J. S. Edu, J. M. Such, and G. Suarez-Tangil, King's College London, Department of Informatics, Faculty of Natural and Mathematical Science, Strand campus, London, UK; emails: {jide.edu, jose.such, guillermo.suarez-tangil}@kcl.ac.uk.

0360-0300/2020/12-ART116 $15.00
https://doi.org/10.1145/3412383

effective and expressive communication tools, voice technology is changing the way in which users interact with devices and the manner they consume services. Currently, the most significant innovations that use voice technology are Smart Home Personal Assistants (SPA). SPA are intelligent assistants that take instructions from users, process them, and perform the corresponding tasks. They offer hands-free and eye-free operations, allowing users to perform diverse activities using voice commands while concentrating elsewhere on other tasks. Besides offering users the benefit of a quick interaction—humans speak faster than they type [118], using voice for HCI can be considered more natural [58] when compared to other interfaces, like keyboard and mouse. Not to mention the stronger social presence offered to users when they hear synthesized speeches very much like their own as responses from this technology [71].

SPA are rapidly becoming common features in homes and are increasingly becoming integrated with other smart devices [111]. It is believed that 10% of the world consumers own SPA devices [97]. According to a recent survey by Voicebot, over 50 million Alexa Echo devices have been sold to date in the U.S. alone [64]. There are a number of features that contribute to the popularity of SPA. SPA are quite different from early voice-activated technologies that could only work with small inbuilt commands and responses. Instead, SPA use Internet services and benefits from recent advances in Natural Language Processing (NLP), which allow them to handle a wide range of commands and questions. They enable a playful interaction, making their use more engaging [74]. They are assigned a name and a gender, which encourages users to personify them and therefore interact with them in a human-like manner [90]. They are used to maintain shopping and to-dos lists, purchase goods, and food, play audio-books, play games, stream music, radio and news, set timers, alarms and reminders [122], get recipe ideas, control large appliances [79], send messages, make calls [59], and many more depending on their usage context [52, 146]. With the continuous proliferation and the rapid growth of SPA, we are now approaching an era when SPA will not only be maneuvering our devices at home but also replacing them in many cases. For instance, many SPA are now able to make phone calls, which positions them as a communicating device, and a likely alternative to landlines phones in the future, and some SPA are also equipped with display interface for watching videos/movies and smart home cameras directly in the SPA devices [5].

As these devices become increasingly popular [97], the most sought-after features expose SPA to various risks. Some of those features are the open nature of the voice channel they use, the complexity of their architecture, the AI features they rely on, and the use of a wide range of different technologies. It is paramount to understand the underlying risks behind their use and fathom how to mitigate them. While most of these devices have incorporated some security and privacy mechanisms in their design, there is still a significant number of security and privacy challenges that need to be addressed. This is all the more important because SPA carry out distinct roles and perform various functions in single and multi-user environments, particularly in an intimate domain like homes. Since users co-locate with this technology, it also has an impact on the changes in their neighboring environment [108]. In fact, there have already been reported security and privacy incidents in the media involving SPA, such as the case of an Amazon Alexa recording an intimate conversation and sending it to an arbitrary contact [146]. Users are concerned about these devices' security and privacy [34, 40]. In the absence of better technical security and privacy controls, users are implementing workarounds like turning off the SPA when they are not using it [1, 69]. Unfortunately, several mitigating techniques proposed in various studies fall short in addressing these risks. For instance, authors of Reference [150] propose a presence-based access control system that does not support an extensive set of use cases. Furthermore, other solutions, such as the one in Reference [35], affect the usability of the SPA.

Despite the fast-growing research on SPA's security and privacy issues, the literature lacks a detailed characterization of these issues. This article offers the first comprehensive review of

existing security and privacy attacks and countermeasures in smart home personal assistants and categorizes them. For this, we first provide an overview and background of the architectural elements of SPA, which is vital to understand both potential weaknesses and countermeasures. In addition, and based on our analysis and categorization of risks, attacks, and countermeasures, this article presents a roadmap of future research directions in this area. We found out that while the attack surface of SPA is distinctly broad, the research community has focused only on a small part of it. In particular, recent works have focused mostly on issues related to the direct interaction between a user and their SPA. While those problems are indeed very important and further research is needed for effective countermeasures, we also found that research is needed to address other issues related to authorization, speech recognition, profiling, and the technologies integrated with SPA (e.g., the cloud, third-party skills, and other smart devices).

### 1.1 Research Questions

We focus on the following main research questions:

- RQ1—What are the main security and privacy issues behind the use of SPA?
- RQ2—What are the features that characterize the known attacks to SPA?
- RQ3—What are the main limitations of the existing countermeasures, and how can they be improved?
- RQ4—What are the main open challenges to address the security and privacy of SPA?

### 1.2 Research Method

We used a systematic literature review (SLR) approach [48, 65] to assess existing literature on the security and privacy of SPA. The primary search process involved searching for keywords related to the study (smart home personal assistants, voice assistants, privacy, security) through databases like ACM Digital Library, Web of Science, IEEE Xplore Digital Library, and ScienceDirect. The secondary search process consisted of searching publications manually in the relevant research area for completeness. Regarding the inclusion and exclusion criteria for the papers we found through the search process above, we included in this review papers that describe research on SPA or research that is of direct relevance or application to SPA. The papers are reviewed with respect to their techniques, years, criteria, metrics, and results. We exclude position papers or short papers that do not describe any results.

### 1.3 Review Structure

The rest of this article is structured as follows: Section 2 offers an introduction to SPA, their architecture. In Section 3, we describe the different security and privacy issues in the SPA. Known attacks on SPA are discussed in Section 4. Section 5 describes existing countermeasures, and Section 6 provides a summary and some discussions on future research directions. Finally, Section 7 draws the conclusion.

## 2 BACKGROUND

SPA have a complex architecture (see details in Section 2.1). As a general introduction, and despite the fact that different SPA across different vendors have a few distinctive characteristics, all SPA perform similar functions and share some common features. In particular, SPA's architectures usually include, together with other architectural elements such as cloud-based processing and interaction with other smart devices, the following: (i) a *voice-based intelligent personal agent* such as Amazon's Alexa, Google's Assistant, Apple's Siri, and Microsoft's Cortana [125]; and (ii) a *smart speaker* such as Amazon's Echo family, Microsoft's home speaker, Google's home Speaker,

and Apple's HomePod. Note that, while we focus on SPA as one full instantiation and ecosystem based on voice-based personal assistants, some of the issues mentioned in this review may apply to other non-SPA voice-based personal assistants, as there are parts of their architecture that may be similar, especially those parts not related to the smart speaker.

SPA decode users' voice input using NLP to understand users' intent. Once the intent is identified, it delegates the requests to a set of *skills*[1] from where it obtains answers and recommendations. Conceptually, skills are similar to mobile apps, which interface with other programs to provide functionality to the user. The entire skills ecosystem provides an environment that offers the user the ability to run more complex functions such as calendar management, shopping, music playback, and other home automation tasks. There are two types of skills, namely, *native skills* and *third-party skills*. The former are skills given by the SPA provider that perform basic functions and leverage providers' strengths in areas such as productivity (Microsoft Cortana), search (Google Assistant), and e-commerce (Amazon Alexa) [145]. The latter are skills built by third-party developers using *skill kits* [4, 45], which are development frameworks with a set of APIs offered by the SPA provider to perform basic operations. There are currently thousands of SPA skills hosted online, although the numbers keep growing daily. For example, Amazon's skill market now has over 70,000 Alexa skills worldwide [62] and the Google Assistant skill market has over 2,000 skills [13]. These skills are classified into different categories such as *home control skills, business and finance skills, health and fitness skills, games and trivia skills, news skills, social skills, sports skills, utilities skills*, and so on. As further support to the skills, SPA often have the ability to learn information about users' preferences such as individual language usages, words, searches, and services using Machine Learning (ML) techniques [89] to make them smarter over time.

### 2.1 Smart Home Personal Assistants Architecture

SPA are Internet-based systems with a regular iteration of updates. One benefit of this is that its capabilities are wide-ranging and dynamic—they will evolve along with the proliferation of new Internet services. Figure 1 shows the key components in the SPA system architecture. Each component is a potential attack point for an adversary. How some of them may be exploited is discussed in Section 4.

Point 1 represents the point of interaction between the users and the SPA devices. SPA devices such as Amazon Echo are equipped with powerful microphones, and the device itself consists of a voice interpreter that records users' utterances. To make use of the SPA, the voice interpreter needs to be activated. Many of the voice interpreters are often pre-activated and run in the background. After the voice interpreter is activated, it then waits for the wake-up word to be triggered [78]. Once it receives the wake-up keyword, it puts the SPA into recording mode. In recording mode, any user utterances are processed and sent through the home router (Point 2) to the SPA cloud (Point 3) [7] for further analysis. Only the wake-up command is executed locally, while all other commands are sent to the cloud. Hence, the SPA must always be online.

The captured utterances are decoded using NLP in the SPA cloud as we detail in Section 2.2 below. It must overcome the issue of background noise, echo, and accent variation in the process of extracting the intent [78]. Once the intent is extracted, it is then used to determine which skill to invoke. There are two ways to invoke a skill. First, they can be explicitly invoked by using their activation name: for example, where a skill name is "Tutor Head," it can be triggered by saying the words: "talk to Tutor Head." Explicit invocation can be extended to use a deep link connection, as detailed here [46] for Google Assistant. For instance, "talk to Tutor Head to find the

[1]Note that, for ease of exposition, we adopt Amazon's terminology of *Skills*, but these may be called differently in other SPA platforms. For instance, in Google's Assistant and Google Home, skills are called *Actions* instead.

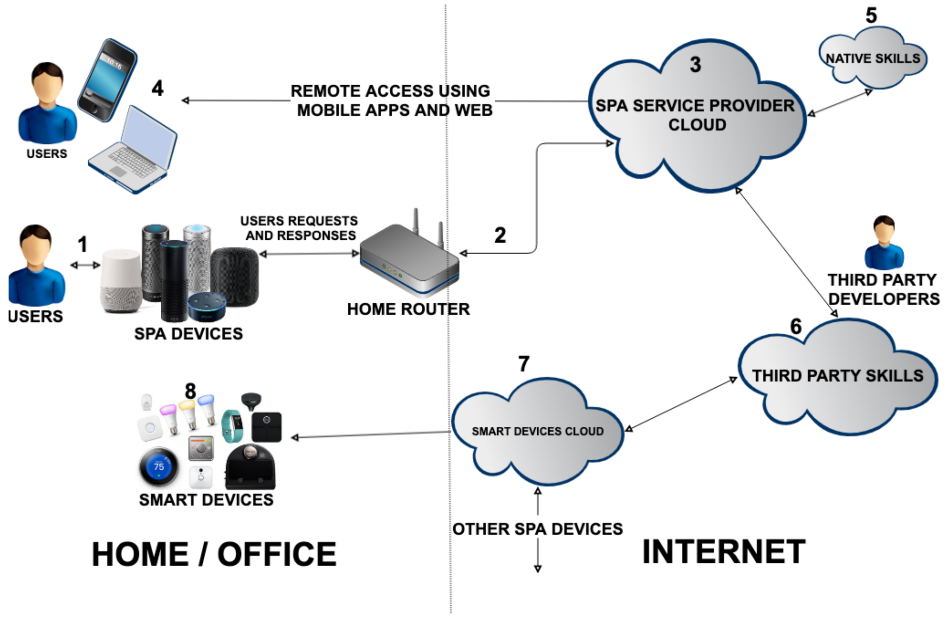


Fig. 1. SPA architecture and its key components [7, 32].

next course" where the next course is a predefined action under the "Tutor Head" skill. Second, skills can be implicitly invoked by an intent's query composition without explicitly using their invocation name. If a query does not directly match with a skill, then the SPA will either inform the user or match the query to another similar skill when appropriate.

By default, the SPA provider will try to find a native skill to process the request invoked by the user [4]. In this case, the SPA cloud service then sends the intent to its native skill, which processes the request in the cloud of the SPA (Point 5) and sends a response back to the SPA device. When there are no native skills available, the request is sent to third-party skills (Point 6). These are typically hosted in a remote web service host controlled by the developer of the third-party skill. Once the request is processed, the third-party skill returns the answers to the SPA cloud service, which sometimes asks for more information before the request is finalized. In the case where the intent is meant to control other smart devices, the relevant information is forwarded to their respective cloud service (at Point 7), and from there, the instructions are relayed to the target smart device (at Point 8).

### 2.2 Natural Language Processing in SPA

SPA benefit from recent advances in Natural Language Processing (NLP), which allow them to handle a wide range of commands and questions. The NLP improvements are attributed to: (i) a number of novel advances in ML, (ii) a better knowledge of the construction and use of the human language, (iii) an increase in the computing power, and (iv) the availability of sizable labeled datasets for training speech engines [50]. Processing user speech includes a complex procedure that involves audio sampling, feature extraction, and speech recognition to transcribe the requests into text. Since humans speak with idioms and acronyms, it takes an extensive analysis of natural language to get correct outputs. For instance, issuing a command to an SPA asking it to remind you about a meeting at a specific time can be done in several ways. While some parts of this command are more specific than others and can easily be understood, such as the day of the week, other

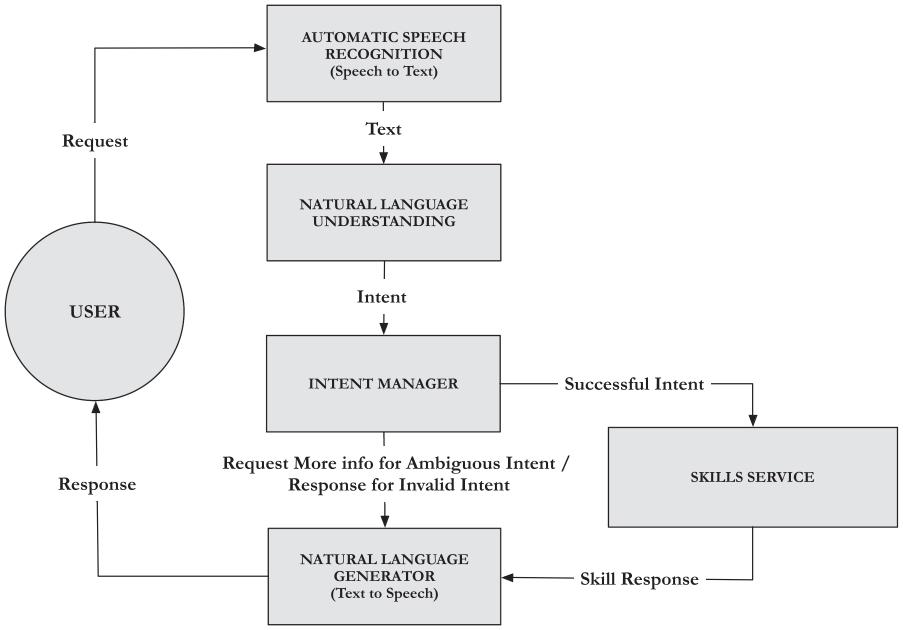


Fig. 2. NLP speech system from *Speech to Text* (ASR) to *Text to Speech*.

words that support them can be dynamic. This implies that understanding an intention as simple as a meeting reminder might require non-trivial interactions. Figure 2 illustrates the process involved in understanding a user's intent and generating responses.

Intent recognition starts with signal processing, which offers the SPA a number of chances to make sense of the audio by cleaning the signal. The idea is to enhance the target signal, which implies recognizing the surrounding noise to reduce it [87]. That is one reason why most SPA devices are equipped with multiple microphones to roughly ascertain where the signal is coming from so that the device can concentrate on it. Once the original signal is identified, acoustic echo cancellation [151] is then used to subtract the noise from the received signal so that only the vital signal remains. Typically, most speech recognition systems work by converting the sound waves from the user's utterances into digital information [41]. This is further analyzed to extract features from the user's speech, such as frequency and pitch. Primarily, Automatic Speech Recognition (ASR) comprises two steps: features extraction and pattern classifiers using ML [73]. There are several feature extraction methods, with Mel frequency cepstral coefficient (MFCC) being one of the most popular, since it is believed to mimic the human auditory system [158]. These features are then fed into an acoustic model trained using ML techniques to match the input audio signal to the correct text [101]. For instance, ML models based on Hidden Markov Model (HMM)[41] often compare each part of the waveform against what comes previously and what comes next, and against a dictionary of waveforms to discover what is being said.

Once the SPA cloud has the text that transcribes what the user has said, it employs Natural Language Understanding (NLU), a key component of Natural Language Processing (NLP), to understand what the user intends to do. This is done using discreet to discreet mapping, with some instances relying on statistical models or ML techniques like deep learning to assume the likely intent. The more data available to the NLP system from regular usage, the better the prediction of the user's intent. After the NLU extracts the intent, the intent manager then decides whether

more information is needed to provide an accurate answer before forwarding the intent to the skill service for processing. After the intent is processed, the generated skill response is sent to the Natural Language Generation (NLG), where it is converted into natural language representation. It is then communicated back to the user, and it is typically (e.g., Amazon Echo) played by a smart speaker.

### 2.3 Assets in the SPA Architecture

Next, we discuss the assets in the SPA architecture from SPA users' point of view, and why users consider the assets to be important or sensitive, to understand what is at stake, and what should be protected.

*2.3.1 SPA and Smart Devices.* The SPA device and other peripheral smart devices are essential assets in this domain. There are different types of SPA devices attending to where the personal assistant interacts with the user. SPA devices can be integrated into smart speakers like Amazon Echo, Google Home, and Apple HomePod. As illustrated in Figure 1, SPA also interact with other smart home devices [1, 140] such as smart heating and cooling devices (e.g., Nest, or Ecobee 4), Smart security (e.g., Scout or Abode), smart lighting devices (e.g., Philip Hue or LIFX), Smart kitchen (e.g., GE+ Geneva) and surveillance cameras (e.g., Cloud Cam, Netgear Arlo Q). All these assets are generally characterized by the hardware they are built on.

*2.3.2 Personal Data.* Personal data is one of the most valuable assets in the SPA ecosystem because of the amount and variety in which personal data is collected, shared, and processed. Therefore, many of the security issues explained below also impact users' privacy, even though this may affect users differently depending on what they value based on their perceptions and preferences [1, 69, 139]. All this may in turn be defined by the user's understanding of the data flows in SPA [1] and what they have experienced in other computing contexts [140]. We give more details below of examples of particular types of personal data in the SPA ecosystem.

*User voice records (audio clips and transcripts).* SPA need to continuously learn from past computations for reliable speech recognition. To achieve this, SPA need a large training dataset of user conversations. Users are known to have concerns about the storage of the recordings of those conversations in some cases and, particularly, about what they may be used for [77].

*User account data.* Users also have data as part of their account with the SPA provider. For instance, in Amazon Alexa, this includes users *location, mobile number, email address, name, device address, payment information, and shopping lists [6].* Note, however, this data is not restricted to the SPA provider, and skills can request permission to access data from the user's account with the SPA provider.

*Skill interaction data.* Skills can potentially ask users for any personal data through their conversations with the users. In fact, there is research evidence that skills collect personal data during voice interaction without asking for any permissions regarding user account data [92]. Birthdate, age, and blood type are examples of the data they may ask for according to this research.

*Smart devices data.* The integration between an SPA and other smart devices brings the smart home into one verbally controlled system and offers the SPA the privilege to manage the services of other connected smart devices. This integration enables access to home sensors that generate valuable personal data.

*Behavioural data.* Apart from the raw data mentioned so far, other sensitive data can be *inferred* from user actions with the SPA or by processing the raw data. This includes predicting

users' behavioral characteristics like user interests, usage patterns and sleeping patterns as shown in References [20, 21], where the authors demonstrate how personal information can be inferred from data stored by the SPA provider by using a forensic toolkit that extracts valuable artifacts from Amazon Alexa by taking advantage of the Alexa unofficial API.

*2.3.3 Other Assets.* There are other assets such as reputation, financial well-being, physical well-being, emotional well-being, and relationships, all of which could be valued differently by users. For instance, if an attacker successfully breaks into an SPA, users could be affected financially if there are unauthorized purchases, emotionally from shame or embarrassment, as well as suffer damage to their reputation if an adversary uses SPA to impersonate them. In fact, some SPA users restrict the use they make of SPA to avoid impacts on these assets, e.g., many SPA users avoid purchasing through SPA, because they do not think the process is secure or trustworthy enough [1].

## 3 SECURITY AND PRIVACY ISSUES

In this section, we present a classification of the main security and privacy issues of SPA. We use this classification to later map current attacks and countermeasures in Sections 4 and 5.

### 3.1 Weak Authentication

Here, we discuss issues related to how SPA verify users and how an adversary can exploit such a process.

*3.1.1 Wake-up Words.* By design, SPA authentication is done using wake-up words that are recognized locally in the device. A user has the option to select a wake-up word from a set of predefined options, having one by default. It is therefore very easy for an attacker to infer the wake-up word of the user. In addition to the wake-up word, SPA have no additional ways of authenticating the user. The device will accept any command succeeding the wake-up keyword. Hence, it is easy for anyone in proximity to issue commands to the SPA. Authors in References [2, 150, 154, 155] have shown how this weak authentication can be used as a proxy to more elaborated security and privacy attacks.

*3.1.2 Always On, Always Listening.* As mentioned, the voice command interpreter constantly listens to the user utterances while waits for the wake-up word. Having a device permanently on and always listening poses important security and privacy concerns. Accidentally saying the wake-up word or any other phonetically similar words will put the assistants to record. Consequently, any conversation that follows is uploaded to the Internet. This issue could affect the users' privacy in a situation where private or confidential conversations are accidentally leaked, or where an attacker can retrieve sensitive information from these devices. Likewise, it could also affect the device security as an adversary can issue an unauthorized command to compromise such devices and use them to target other connected smart devices. Recently, due to this feature, a private conversation of a couple was accidentally recorded and sent to a random contact with the Echo device [51]. This example shows that the users are not in total control of their voice data.

*3.1.3 Synthesized Speech.* SPAs are known to listen to audio playback. Just recently, a Tv commercial by Burger King prompted Google Home to read information to the user from Wikipedia about the Whopper hamburger [147]. While major SPAs like Alexa and Google have now figured out how to filter out background media [98, 113], they are still vulnerable to synthesizing audio that exploits side channels or adversarial examples. For instance, they are vulnerable to inaudible sound reproduced at ultrasonic frequencies [117, 154], and synthesized speech transmitted through electromagnetic radiation. In particular, laser-powered "light commands" [135]. Since the SPA wake-up

word can be readily guessed, and the SPA has no means of detecting if a user is in close proximity, there is little or no limit to which speech can be supplied to them and by whom, provided it is meaningful and can be matched with an intent. Synthesized speech (like ultrasonic/inaudible attacks) could offer an adversary a covert channel to issue a malicious command. An attacker could even distribute these speeches over channels like TV and radio to attack multiple targets at once.

### 3.2 Weak Authorization

In this part, we evaluate the issues regarding how the SPA manages the level of access to data, and the mechanisms users have to control that.

*3.2.1 Multi-user Environment.* The absence of proper functional role separation prevents users from correctly defining what and how resources should be accessed. It is challenging to specify who has access to which resources and how such access should be granted. By default, in a multi-user environment—which many households are, any user can put the SPA into recording mode and issue out instructions to it. Even though the primary user can specify certain access controls for secondary users, the level of granularity is generally coarse and not extensive. For instance, any member of an Amazon household (a feature that allows sharing of contents with family members) can modify the device set-up such as the network connection, sound, and many more without the primary user consent.

*3.2.2 Weak Payment Authorization.* SPA systems are increasingly supporting online ordering. Implementing proper security controls challenges usability. For instance, Amazon Alexa users have the option to set a four-digit PIN code to confirm purchases. At the time of writing, this option is not enabled by default. Even when such an option is turned on, it is vulnerable due to weak lockout[2] implementation [47]. This is because Alexa allows two PIN tries before an ordering process lockout, after which the user has to restart the ordering process from the beginning. However, there is no restriction on how many times a user can try to order after every lockout [47]. Following this, vendors have tried to implement alternative countermeasures against misuse in the ordering process. We next show two cases of this. First, some vendors have prevented changes to the shipping address during ordering. Preventing any change to the shipping address during this process is not enough when dealing with "insiders" (i.e., unauthorized users who have access to the premises where the SPA are installed). The case described in Reference [68] shows how a kid recently made an unauthorized order worth of about $300 using her mother's Amazon account [68]. Second, other vendors have tackled this weak authorization problem by providing prompt notification to the users about orders. This poses a problem to users who do not frequently check their phones or emails, or who may not understand what is happening.

*3.2.3 External Party.* One important concern is how SPA providers, skills developers, developers of integrated smart home devices, and those that have direct access to any of the points of the SPA architecture secure users from external parties that do not have access to any of these points. Like in every other cloud service, the question remains on how data gathered by those involved in the SPA system is shared with third parties, particularly regarding what kind of controls and mechanisms can be implemented to provide more control to users. Informed decisions can sometimes be taken when third parties provide privacy policies and terms of use [144]. However, it is currently uncertain what the scope of those terms might imply and how they are enforced.

[2]Lockout is a security mechanism that locks an application for some time before a reattempt is allowed.

### 3.3 Profiling

Beyond authorization, i.e., deciding who has access to what data, there is also the problem of data inference—traditionally known as information processing [123]. Data inference has a particularly dangerous incarnation in SPA in the form of profiling. Profiling identifies, infers, and derives relevant personal information from data collected from users. Profiled data can be related to the interests, behaviors, and preferences of the targeted users [30]. In this subsection, we look into how SPA data can be used to profile users.

*3.3.1 Traffic Analysis.* A good instance of an en-route type of profiling is traffic analysis. An attacker can take advantage of SPA traffic's improper concealment to profile a user as shown in Reference [11]. In particular, attackers can leverage en-route profiling to infer a user's presence. This can be further used to conduct more sophisticated attacks. En-route profiling attacks can be made even when the network traffic is encrypted. While there are obfuscation techniques that can be used to hinder these types of attacks, they have not been adopted in SPA. In this scenario, the most plausible adversary would be a dishonest or unethical Internet service provider. Governments or other global adversaries with access to the user network traffic can also exploit this weakness. The practicality of this threat to encrypted SPA traffic is shown in Reference [11]. While authors in Reference [11] perform traffic analysis without even needing an in-depth inspection of the network packages, MiTM techniques—such as SSL-stripping [156]—might be used to perform profiling over plain-text.

*3.3.2 Uncontrolled Inferences.* Profiling, in this case, is about inferences made by any of the parties in the SPA ecosystem (third-party skill developers, SPA providers, etc.) from data they collect with the consent of the user. This includes some of the personal data mentioned in Section 2.3 (conversations, account data, interaction data, etc.). That is, the starting point is data about the user that the user may have consented to share. This data is then used to infer *new* data about the user that the user had not shared. An example would be the behavioral data mentioned in Section 2.3. Therefore, the problem is that even when users can choose whether they share some data, they have no control over what the parties can do with the data, or what kind of inferences or aggregations they could make to derive other new personal information about the user, e.g., users' tastes or predilections. Note that in some cases, collusion between the parties might be possible to be able to conduct more powerful inferences. For instance, malicious skills may collude to aggregate personal data from multiple skills similar to what we have seen in smartphone apps [83]. Here, skill connection pairing [63] may be leveraged to create colluding skills aiming at getting more elaborated profiling. Uncontrolled inferences are especially critical as advances in data analysis enable automated techniques to make sense of unstructured data at scale.

### 3.4 Adversarial AI

As described in Section 2.2, for an SPA to fulfill the user's request, it needs to first understand what the user's said, understand what the user wants, and before selecting the best skill to fulfill the request. For these, the speech recognition system uses AI techniques like NLP and ML. However, these techniques can introduce the issues discussed below.

*3.4.1 Adversarial ML.* ML in SPA system is used for many tasks, including speech recognition. Conventionally, ML is designed based on the notion that the environment is safe, and there is no interference during training and testing of the model [99]. However, such an assumption indirectly overlooks cases where adversaries are actively meddling with the learning process [99]. ML is known to be vulnerable to specially-crafted inputs, described as adversarial examples, which are usually derived by slightly modifying legitimate inputs [138]. These perturbations typically remain

unknown to the person supervising the ML task but are wrongly classified by already trained ML models. Examples can be used to manipulate what the SPA system understands from spoken user commands [154]. This could then be used to generate a denial of service attack, invoke an incorrect skill [142], or to reduce the ML model quality and performance [18]. Most ML models that perform the same task tend to be affected by similar adversarial inputs even if they use different architectures and are trained on different datasets [100]. This allows the attacker to easily craft adversarial inputs with little knowledge about the target ML model. Research has also shown that speech recognition models often find it challenging to differentiate words with similar phonemes [67], e.g.: distinguish between "Cat," "Pat," and "Fat," which can come handy when crafting adversarial inputs. Commonly exploited ML vulnerabilities are not the only type of examples that may apply. For instance, to predict the best skill to process the user's request, most SPA continuously learn from the user interactions and regularly retrain their ML models with new data. Attackers could insert adversarial samples into the training dataset to corrupt the ML models (poisoning attack). Another example would be targeting the ML models to extract valuable information (membership inference attack), e.g., the accent of the speakers in speech recognition models [121].

*3.4.2 NLP Vulnerabilities.* Although adversarial ML has a direct effect on the NLP system in SPA as it underpins many NLP tasks used for speech recognition, there are also other parts of the NLP system in SPA that do not directly use ML but that may also be exploited. Following the example of skill invocation given in the previous subsection, the adversarial NLP problem appears once user utterances have already been transcribed into text and the system needs to decide which skill to invoke from the text (note the difference with the problem of translating into text two words with similar pronunciation). In particular, Amazon's Echo and Alexa seem to use the lengthiest string match when deciding which skill is called [155]. For example, the text "talk to *tutor head* for me please" will trigger the skill "*tutor head for me*" rather than the skill "*tutor head*." In a similar way to adversarial ML, an attacker could use such difficulty to trick users into invoking a malicious skill intentionally. This can be achieved by registering a skill with the same name (but longest possible string match) than a legitimate skill. Besides, there is currently no restriction on the number of skills that can be registered, hence, an adversary can register as many skills as possible to increase the possibility of getting their skills called.

### 3.5 Underlying and Integrated Technologies

To broaden SPA capabilities and offer ubiquitous services, SPA rely on skills and other existing infrastructures like cloud services and smart devices. This means that they can potentially inherit or be subject to issues and vulnerabilities present in or arising from these technologies.

*3.5.1 Third-party Skills.* An attacker could take advantage of lax enforcement of the skill implementation policies and exploit the interaction between the user and the SPA system. For example, by faking the hand over process, a malicious skill can pretend to hand over control to another skill and deceive users into thinking that they are interacting with a different skill (Voice Masquerading attack) to eavesdrop on user conversations and collect sensitive information. After all, it is difficult for the user to determine if they are taking to the right skill at a particular period of time because of the vagueness of voice command [91]. Likewise, a malicious skill can fake or ignore the skill termination command and continue to operate stealthily [124]. Furthermore, the existing SPA architecture support only permission-based access control on sensitive data. It is insufficient at controlling how skills use data once they get access [54]. This could create privacy concerns, especially in over-privileged skills as it does not allow users to specify the intended data flow patterns once a skill has permissions to access data. In fact, authorizing a malicious skill to access confidential information may result in leaking sensitive information to unwanted parties. In the

SPA ecosystem, the end-user does not have any kind of access to the skills, which is rather different from the apps in smartphones that will be running in your phone, so protection mechanisms in the smartphone can be used to target apps. In contrast, users don't have a way to install any protection mechanisms beyond those the SPA provider can put in place for skills. A user must rely on the SPA provider to ensure that such services are as secure as they need to be. However, even if the SPA provider would provide a vetting process, related works have shown that they can be successfully evaded [124, 155].

*3.5.2 Smart Home Devices.* While SPA integration with other smart home devices brings the smart home into one verbally controlled system, it also creates a single key point of interest to attackers. Attackers can take advantage of this in two ways. On the one hand, breaching the SPA can allow attackers to control a wide range of connected devices. More so, privacy issues could emerge from data accumulation, data acquisition, and integration as discussed in References [75, 114], where the authors perform a comprehensive review of privacy threats of Information Linkage from data integration in IoT ecosystems. On the other hand, vulnerabilities in connected smart devices could be used as an intermediate step to attack the SPA [31, 116, 127]. Attacks in connected smart home devices have been investigated in numerous works, including: (1) snooping attack where an adversary listens to the smart home traffic to read confidential data [31], (2) privilege escalation where attackers use design and configuration flaws in smart home devices to elevate privileges and access confidential home users information, (3) insecure interactions between apps that are used for controlling peripheral devices and third-party counterpart apps that could open channels for remote attackers, and (4) other direct compromises of various smart home devices [31, 36]. For instance, the API service on Google Home before mid-July 2018 was reported to be vulnerable to DNS rebinding attacks, which allow remote attackers to initiate a denial of service attack, extract information about the Wi-Fi network or accurately locate this device [93]. It is important to note that some of the issues we identify in this review are not specific to SPA alone. They are also present in other smart home and IoT devices, since the SPA and other IoT devices conduct information exchange and communications in a similar way, and are often co-located within the same environment. Nonetheless, the SPA ecosystem is quite unique, e.g., the speech and intent recognition steps, which determine the actual third-party skill that is to serve a user command may lead to specific adversarial AI issues as mentioned above.

*3.5.3 Cloud.* While the cloud offers the advantage of having readily available and virtually unlimited resources, they also present attackers with new opportunities [86]. On the one hand, they are data-rich environments that are centrally located in a single point, and in particular in SPA architectures, they keep most of the personal data mentioned in Section 2.3. If this element is breached, then attackers may get access to valuable and sensitive information. This is the most concrete and frequently mentioned threat by users regarding smart home data [140]. On the other hand, they usually offer multiple remote ways of accessing the data (e.g., web or app-enabled access) and facilitate online configuration, thereby widening the attack surface. The SPA provider cloud (point #3 in Figure 1) is therefore subject to these issues. Most importantly, data in the cloud are subjected to insider attacks (i.e., abuse of authorized access) [104, 126]. For instance, some SPA providers may let employees listen to recorded conversations as they view this process as a critical part of evaluating their SPA speech recognition system [126] and a way of improving customer experience [104]. This is a critical issue when their privacy statements fail to mention this type of usage or whether conversations are used anonymously [112]. Likewise, the SPA provider cloud could also suffer from incomplete data deletion [109]. This situation may enable SPA providers to retain (intentionally or accidentally) private data even after being deleted (assuming users manage to find the way to delete information from the cloud, which is not always easy for them [110]). For

instance, it is known that Amazon could keep transcripts of users' voice interactions with Alexa even after the recordings are deleted [60].

## 4 ATTACKS

This section offers a review of known attacks on the SPA system and examines the vulnerabilities they exploit w.r.t. the issues described in Section 3 and the point they target in the architecture in Section 2. Table 1 shows an overview of the most relevant attacks mapped to the vulnerability(ies) they exploit and the affected points in the architecture. We found that most of the attacks target the following elements of the architecture depicted in Figure 1:

(1) User to SPA device (#1): There is a wide range of attacks targeting this point of the architecture. In particular, we identify related works (i) exploiting weak authentication and (ii) attacking underlying and integrated technologies.
(2) SPA device to SPA service provider cloud (#2): There is an attack reported in the literature that targets this point of the architecture and exploits improper concealment of SPA traffic.
(3) SPA service provider cloud (#3): Several attacks are also found at this point of the architecture targeting the SPA cloud components. We identify works exploiting (i) ML Vulnerabilities and (ii) underlying technologies.
(4) Third-party Web skills (#6): Attacks targeting this point of the architecture exploit user misconceptions about the SPA system, and in particular about the skill. We show related works exploiting NLP subsystem vulnerabilities.

We could not find any attacks targeting architectural elements #4 (remote access via mobile and Web), #5 (native Web skills), #7 (smart device cloud), and #8 (connected smart devices). However, this does not mean that attacks targeting those architectural elements are not possible. In fact, some of the threats outlined in Reference [31] and the attacks demonstrated by researchers in Reference [107] could possibly exploit #8. Besides, some of the vulnerabilities that exist in #3 might also be found in #7 as they are both cloud technology. Likewise, attacks targeting #6, such as voice squatting and voice masquerading [155], might also be possible in #5, since both are skill services. Nevertheless, as far as we know, they have not been exploited yet. We discuss this more in detail later on in Section 6.

We next describe the attacks we found in related literature by types (or categories) of attacks, particularly looking at the vulnerabilities (described in Section 3) that they exploit and the assumptions they make on the environment.

### 4.1 Side Channel Attacks

This includes attacks that are based on information gained from the way an SPA is implemented, rather than vulnerabilities in the SPA itself. The *always on, always listening*, and the *lack of arbitrary wake-up words* within the *weak authentication* category are the most exploited vulnerabilities in this class of attack.

Lei Xinyu et al. [150] look at issues in single-factor authentication methods based on a wake-up word, and the lack of a mechanism that can be used to figure out if a user is close-by or not. Using Amazon's Echo device, the authors perform a home burglary attack to manipulate a connected door lock. Likewise, they successfully make a purchase using the compromised device. Authors in Reference [135] also exploit the lack of proper user authentication and vulnerable microphones to inject voice commands into SPA. By simply modulating the amplitude of laser light, the authors successfully use light-injected voice commands to unlock a connected smart lock integrated with the SPA, and to locate, unlock, and start cars (including Ford and Tesla) provided they are linked with the target's Google account. However, unlike in other classes of attacks where attackers are

Table 1. Categorization of Attacks Found in Previous Studies Based on Vulnerabilities Exploited and Attack Point

| Attack Class | Studies | Weak Authentication | | | Weak Authorization | | | Profiling | | Adversarial AI | | Integrated Techs. | | | Attack Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Wakeup Word | Always Listening | Synthesized Speech | Payment Auth. | Multiuser Environ. | External Party | Traffic Analysis | Uncont. infer. | Adv ML | NLP Vul | Skills | Cloud | Smart Devices | |
| Side Channel | Lei Xinyu et al. [150] | ✓ | ✓ | | ✓ | | | | | | | | | ✓ | 1 |
| | Zhang et al. [154] | ✓ | ✓ | ✓ | | | | | | | | | | | 1 |
| | Segawara et al. [135] | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | 1 |
| | Roy et al. [117] | ✓ | ✓ | ✓ | | | | | | | | | | | 1 |
| Behavioral Profiling | Apthorpe et. al. [11] | | | | | | | ✓ | | | | | | | 2 |
| Attacks on Voice Models using Adversarial samples | Gong & Poellabaeur [42] | ✓ | ✓ | | | | | | | ✓ | | | | | 1, 3 |
| | Schönherr et al. [120] | ✓ | ✓ | | | | | | | ✓ | | | | | 1, 3 |
| | Carlini and Wagner [18] | ✓ | ✓ | | | | | | | ✓ | | | | | 1, 3 |
| | Vaidya et al. [142] | ✓ | ✓ | ✓ | | | | | | ✓ | | | | | 3 |
| | Carlini et al. [16] | ✓ | ✓ | ✓ | | | | | | ✓ | | | | | 3 |
| Skill Squatting & Masquerading | Zhang et al. [155] | | | | | | | | | ✓ | ✓ | ✓ | | | 3, 6 |
| | Kumar et al. [67] | | | | | | | | | ✓ | ✓ | ✓ | | | 3, 6 |
| | Security Research Labs [124] | | | | | | | | | | | ✓ | | | 3, 6 |

restricted by distance due to the use of sound for signal injection, attackers here are only limited by their capabilities to carefully aim the laser beam on the devices' microphones. Since light does not accurately penetrate through an opaque object, this attack requires a line of sight to the targeted SPA devices.

The non-linearity in the Micro-Electro-Mechanical Systems (MEMS) microphone over ultrasound is exploited by Zhang et al. [154]. Non-linearity is described as hardware features that cause signals with high-frequency triggers at high power to be shifted to low frequencies by microphones (and speakers) [117]. Even though microphones are designed to be a linear system, they exhibit non-linearity in higher frequencies. By synthesizing high-frequency sounds that are not within the human hearing range but are still intelligible to SPA devices, the authors are able to activate and control the voice of the SPA. This technique is called the dolphin attack as it uses ultrasonic frequencies like what Dolphins use to communicate among themselves. This attack was confirmed on seven popular voice intelligent assistants (Siri, Cortana, Huawei Hi Voice, Google Now, Samsung S Voice, and Alexa) over a range of different voice platform. On the downside, this attack cannot be conducted above a distance of 5ft from the targeted device. Likewise, it requires specialized hardware to synthesize and play the ultrasonic signal, making it unrealistic for a real-world attack.

In a different study, Roy et al. [117] develop a long-range version of the dolphin attack. They achieved a range of 25ft from their target. By exploiting the non-linearity inside the microphone, like in Reference [154], they generated long-range high-frequency signals that are inaudible to human but intelligible to SPA. As in the previous study, they control and issue commands to SPA devices with the assumption that the adversary can synthesize a legitimate voice signal. However, rather than using a single ultrasound speaker as done in Reference [154] to play the synthesized signal, the authors used multiple speakers that are physically separated in space. They employ spectrum splicing to optimally slice voice command frequencies and play each slice on independent speakers in a way that the total speaker output is inaudible. Nevertheless, the attack is only feasible in an open environment. This is because high frequencies are more susceptible to interference, which is a limiting factor to the distance [53]. Likewise, this attack requires multiple ultrasound speakers, making it more challenging to implement in a real-world attack.

### 4.2 Behavioral Profiling

At point #2 of the architecture where SPA devices exchange information with the SPA cloud provider, authors in Reference [11] identify privacy vulnerabilities with SPA by passively analyzing encrypted smart home traffic. Their study indicates that encryption alone does not offer all the necessary privacy protection requirements. The authors profile users' interaction with Amazon Echo devices by plotting send/receive rates of the stream even with encrypted traffic. This poses a severe privacy implication to smart home users as an attacker can use this to infer their lifestyle and the best time to conduct an attack undetected, as discussed in Section 3.3.1. However, the method used in this study might not apply to a situation where different IoT devices communicate with the same domain because of the difficulty of labeling streams by device type.

### 4.3 Attacks on Voice Models using Adversarial Samples

Here, we discuss attacks on speech recognition and processing system using adversarial inputs.

Looking at where data-driven ML models operate, authors of Reference [42] show a new end-to-end scheme that creates adversarial inputs by perturbing the raw waveform of an audio recording. With their end-to-end perturbation scheme, the authors crafted adversarial inputs that mislead the ML model. Note that this is widely used in para-linguistic applications. Their adversarial perturbation has a negligible effect on the audio quality and leads to a vital drop in the efficiency

of the state-of-the-art deep neural network approach. On the downside, such an attack needs to be embedded in a legitimate audio signal to make them truly obscure. While this attack was not evaluated on a real SPA, it was successful against paralinguistic tasks, which are clearly relevant to SPA. In particular, speaker recognition task for performing voice matching [3, 44] to predict the identity of the speaker.

More recently, Schönherr et al. [120] have proposed an adversarial example based on psychoacoustic hiding to exploit the characteristics of Deep Neural Network (DNN)-based ASR systems. The attack extended the initial DNN analysis process by adding a back-propagation step to study the level of freedom of an adversarial perturbation in the input signal. It uses forced alignment to identify the best temporal fitting alignment between the maliciously intended transcription and the benign audio sample. It is also used to reduce the perceptibility of the perturbations. The attack is performed against Kaldi,[3] where it obtained up to 98% success rate with a computational effort for a 10-secs sound file in less than 2 min. However, like in Reference [42], this attack also needs to be embedded in another audio file, which significantly influences the quality of the adversarial example.

Another important study conducted by Carlini and Wagner in Reference [18] proposes an attack on speech recognition systems using Connectionist Temporal Classification (CTC) loss. They demonstrated how a carefully designed loss function could be used to generate a better lower-distortion adversarial input. This attack works with a gradient-descent-based optimization [43] and replaces the loss function with the CTC-loss, which is optimized for time sequences. However, the audio adversarial examples generated when played over-the-air cease to be adversarial, making it unrealistic for a real-world attack.

Similarly, Vaidya et al. [142] perform an attack on speech recognition systems using unintelligible sound. This is done by modifying the Mel-Frequency Cepstral Coefficients (MFCC)—feature of the voice command. The attack is performed in two steps: first, altering the input voice signal through feature extraction with adjusted MFCC parameters, and then regenerating an audio signal by applying a reverse MFCC to the extracted features. When put together, this attack is able to craft a well designed adversarial input. The MFCC values are selected in a way that they can create a distorted audio output with least sufficient acoustic information. This audio output can still achieve the desired classification outcome and is correctly interpreted by the SPA while unintelligible to human listeners. Although this attack successfully exploits the differences between how computers and humans decode speech, it could, however, be detected if a user is in proximity—provided that they hear unsolicited SPA responses. The attack presented by Vaidya et al. [142] is extended in the work of Carlini et al. [16], where the authors test the attack effectiveness under a more realistic scenario and craft an adversarial example completely imperceptible to humans by leveraging the knowledge of the target speech recognition system.

### 4.4 Skill Squatting and Masquerading Attacks

In this section, we discuss attacks that exploit how skills are invoked and the way skills interact with each other.

Authors of Reference [155] target the interaction between third-party skills and the SPA service. Specifically, they analyze two basic threats in Amazon's Alexa and Google's Assistant SPA services: voice squatting and voice masquerading. Voice squatting allows an attacker to use a malicious skill with longest matching skill name, similar phonemes, or paraphrased name to hijack the voice command of another skill as described in Section 3.4.2. In five randomly sampled vulnerable target skills, the authors successfully "hijacked" the skill name of over 50% of them. The feasibility of this

[3]A widely adopted open-source toolkit written in C++ that offers a wide range of modern algorithms for ASR.

type of attack is high, particularly in SPA, such as Alexa that allows multiple skills with the same invocation name. This attack can be used to damage the reputation of a legitimate skill as any poor service of the malicious skill will be blamed on it.

Equally, in voice masquerading attack, a malicious skill pretends to invoke another skill or fake termination. Then, the skill keeps recording the user's utterances. This attack could be used to snoop on the conversations of the user. While voice squatting attacks exploit the weaknesses in the skill's invocation method, voice masquerading targets user's misconceptions about how SPA skill-switch services work. With some skills requesting for private information, an adversary could use these attacks to obtain sensitive information and cause crucial information leakage to unwanted parties. Voice squatting attack is also shown in the work of Kumar et al. [67]. But unlike what was done in Reference [155], Kumar et al. use the intrinsic errors in NLP algorithms and words that are often misinterpreted to craft malicious skills and exploit the ambiguity in the invocation name method.

## 5 COUNTERMEASURES

In mitigating the identified risks and attacks, there have been a number of studies proposing various countermeasures. In this section, we summarise research on countermeasures, highlighting limitations and deficiencies. We give a summary of these in Table 2. We mapped the proposed countermeasures to the vulnerabilities discussed in Section 3. The current mitigation level in the table (last row of Table 2) aims to provide a quick indication of the extent the issues identified have been resolved by the countermeasures proposed by the existing publications analyzed to date. In some cases, a combination of countermeasures is enough to address a specific concern, while others will require new countermeasures to effectively address them. The table also has a column called "Usability Impact" to indicate whether usability is considered or not by the countermeasure. We use the symbol "!" where there is "potential usability impact" such as where users are required to put on extra wearable devices (sacrificing user convenience) [35, 61], or the solution might restrict the SPA capability [26], and "?" for the rest, which means "usability not explicitly considered," as we did not find enough information in the papers to make any claims (positive or negative) about usability. Finally, we also map these countermeasures to the elements of the architecture depicted in Figure 1 to describe the points at which the mitigations would be applied. Most countermeasures map to:

(1) User to SPA device (#1): There is a wide range of countermeasures proposed to mitigate attacks at this point of the architecture. In particular, we found many related works mitigating *weak authentication* vulnerabilities.
(2) SPA device to SPA service provider cloud (#2): At this point of the infrastructure, we found studies proposing different mitigation techniques to obfuscating traffic between the SPA device and the SPA service provider cloud, with the aim to mitigate *en-route* vulnerabilities within the *profiling* category.
(3) SPA service provider cloud (#3): Few of the existing countermeasures also focused on the *Adversarial AI* vulnerabilities that are found at this point of the architecture and recommended measures aim to mitigate the risks associated with them.
(4) New Architecture: Countermeasures in this category modify to some extent the existing SPA architecture as part of the mitigation and/or mitigate vulnerabilities that cut across multiple points of the infrastructure. We mapped these countermeasures to multiple architecture elements to signal the points where the mitigations apply or the points that would change as part of an architecture modification.

Table 2. Categorization of Countermeasures Found in Related Studies

| Class | Studies | Weak Authentication | | | Weak Authorization | | | Profiling | | Adversarial AI | | Integrated Techs. | | | Mitigating Point | Usability Impact |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Wakeup Word | Always Listening | Synthesized Speech | Payment Auth. | Multiuser Environ. | External Party | Traffic Analysis | Uncont. Inf. | ML Vul. | NLP Vul. | Skills | Cloud | Smart Devices | | |
| Voice Auth. | Voice Match / Profiles [3, 44] | | | ✓ | ✓ | | | | | | | | | | 1 | ? |
| | Kepuska and Bohouta [61] | ✓ | ✓ | ✓ | | | | | | | | | | | 1 | ! |
| | Huan et al. [35] | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | 1 | ! |
| | Chen et al. [19] | | | ✓ | | | | | | | | | | | 1 | ? |
| Location Verification | Lei Xinyu et al. [150] | ✓ | ✓ | | | | | | | | | | | | 1 | ? |
| Spectral Analysis & Frequency Filtering | Roy et al. [117] | | | ✓ | | | | | | | | | | | 1 | ? |
| | Zhang et al. [154] | | | ✓ | | | | | | | | | | | 1 | ? |
| | Lavrentyeva et al. [70] | | | ✓ | | | | | | | | | | | 3 | ? |
| | Malik et al. [76] | | | ✓ | | | | | | | | | | | 3 | ? |
| Traffic Shaping | Liu et al. [72] | | | | | | | ✓ | | | | | | | 2 | ? |
| | Park et al. [102] | | | | | | | ✓ | | | | | | | 2 | ? |
| | Apthorpe et al. [10] | | | | | | | ✓ | | | | | | | 2 | ? |
| Command & Phonetic Analysis | Zhang et al. [155] | | | | | | | | | | | ✓ | | | 3 | ? |
| | Kumar et al. [67] | | | | | | | | | | ✓ | ✓ | | | 3 | ? |
| New Architecture | Coucke et al. [26] | ✓ | | | | | ✓ | ✓ | ✓ | | | ✓ | ✓ | | New Arch. | ! |
| | Current Mitigation Level | ◐ | ◐ | ● | ◐ | ○ | ◐ | ◐ | ◐ | ◐ | ◐ | ◐ | ◐ | ○ | | |

### 5.1 Voice Authentication

One of the defense that has been put in place against weak authentication is voice authentication. With this defense, the SPA can tell apart individual users when they speak. For instance, some SPA such as Google and Amazon perform speaker verification through voice authentication, known as *voice match* [44] and *voice profiles* [3], respectively. However, none of these mechanisms is enabled by default, and it is left to the users to first realize about their existence and then decide whether they would like to activate them or not. Even when these mechanisms are activated, they are still open to attack as an attacker can still trick the system with a collected or synthesized voice sample of the legitimate user [19]. Collecting voice samples is an easy task, since the human voice is open to the public. Unlike passwords that can easily be changed if compromised, a human voice is a feature that is difficult to replace.

Another important voice authentication method is proposed in Reference [35]. In this study, the authors present a continuous authentication VAuth system that aims to ensure that the SPA works only on legitimate users' commands. The solution consists of a wearable security token that repeatedly correlates the utterances received by the spa with the body-surface vibrations it acquires from the legitimate user. The solution was said to achieve close to 0.1% false positive and 97% detection accuracy and works regardless of differences in accents, languages, and mobility. Even though this system achieves a high detection accuracy, the need to wear devices such as eyeglasses, headset, and necklaces would introduce a potentially unbearable burden and inconvenience to the users.

Kepuska and Bohouta [61] also proposed a multi-modal dialogue system that combines more than one of voice, video, manual gestures, touch, graphics, gaze, and head and body movement for secure SPA authentication. Even though this system might be able to solve the authentication and voice impersonation challenges earlier discussed, the authors have only been able to test the individual components of the system and not the entire system as a whole.

Finally, Chen et al. [19] propose a software-only impersonation defensive system. The system is developed based on the notion that most synthesized speech needs a loudspeaker to play the sound to an SPA device. As conventional loudspeakers generate a magnetic field when broadcasting a sound, the system monitors the magnetometer reading, which is used to distinguish between voice commands from a human speaker and a loudspeaker. In a situation where the magnetic field emitted is too small to be detected, the system uses the channel size of the sound source to develop a means of authenticating the sound source. However, the effectiveness of the system depends heavily on the environmental magnetic interference. Likewise, the sound source needs to be at a distance of more than 2.3 in (6 cm) to their system to prevent the magnetic field from interfering with the magnetometer's reading. In addition, the system has a high false acceptance rate when the sound source distance to their system is greater than 4 in (10 cm) in a situation where the loudspeaker magnetic field is un-shielded and less than about 3 in (8 cm) when shielded.

### 5.2 Location Verification

Another important measure implemented against weak authentication is presence-based access control system. This system allows an SPA to verify if a user is truly nearby before accepting any voice commands. Lei Xinyu et al. [150] propose a solution that uses the channel states information of the router Wi-Fi technology to detect human motions. Interestingly, it eliminates the need for some wearable devices and introduces no added development cost as it uses the existing home Wi-Fi infrastructure. The solution has an advantage over the traditional voice biometrics recognition, i.e.: that becomes ineffective as users age, become tired, or ill. However, the system's effectiveness depends on selecting the best location for the Wi-Fi devices and setting the right parameters for

the detection. Besides, it only supports commands that come from the same room where the SPA device is deployed: in their case, an Amazon Echo. Likewise, the system is situational as it works best if there is no structural change to the location where the devices are deployed.

### 5.3 Frequency Filtering & Spectral Analysis

Another category of countermeasures aims to enhance authentication, particularly protecting the SPA against synthesized speech using frequency filtering and spectral analysis.

In the work of Roy et al. [117], the authors propose a system nicknamed *lip read* that is based on the assumption that some of the features of voice signals–basic frequencies and pitch–is preserved when it passes through non-linearity. It was reported that this system obtains a precision rate of 98% and a recall rate of 99% in a situation where the adversary does not influence the attack command. However, there is no formal guarantee of this countermeasure as they are unable to model the frequency and phase responses for general voice commands. Likewise, their defense only considers inaudible voice attack ignoring finding the true trace of non-linearity. Similarly, Zhang et al. [154] propose another set of countermeasures against synthesized speech attacks. The authors recommend two hardware-based mitigating measures—the first one aims to enhance the microphones use by the SPA devices. In contrast, the latter hardware-based defense is intended to cancel any unwanted baseband signal. Enhancing the microphone approach entails designing an improved microphone similar to the one found in Apple iPhone 6 plus that can subdue any ultrasonic sound. However, canceling the unwanted baseband signal of the inaudible voice command solution entails introducing a module before the low pass filter in the subsystem used for voice capturing to identify and cancel the inaudible voice commands baseband signal. Likewise, the software-based countermeasure relies on the principle that a demodulated attack signal can be distinguished from legitimate ones using a machine-based learning classifier.

In another study, Malik et al. [76] proposed a countermeasure based on higher-order spectral analysis (HOSA) features to detect replay attacks on SPA. The authors show that replay attacks introduce non-linearity, which can be a parameter to detect it. Lavrentyeva et al. [70] also explore different countermeasures to defend against voice replay attacks. Even though the countermeasure is implemented at #3 of the architecture, because it needs extensive computational power, it aims to secure #1. The researchers use a reduced version of Light Convolutional Neural Network architecture (LCNN) based on the Max-Feature-Map activation (MFM). The LCNN approach with Fast Fourier Transform (FFT)-based features obtained an equal error rate of 7.34% on the ASVspoof 2017 dataset compared with the spoofing detection method in Reference [141] with an error rate of 30.74%. The authors further utilized Support Vector Machine (SVM) classifier to offer valuable input into their system's efficiency. Consequently, their primary system based on systems scores fusion of LCNN (with FFT-based features), SVM (i-vector approach), recurrent neural network (RNN), and convolutional neural network (with FFT-based features) shows a better equal error rate of 6.73% on their evaluation dataset.

### 5.4 Traffic Shaping

To defend against profiling, Liu et al. [72] propose a countermeasure to mitigate traffic analysis vulnerabilities (part of the *profiling* category). The authors present a solution that protects the smart home against traffic analysis—a community-based "differential privacy framework." The framework route traffic between different gateway routers of multiple cooperating smart homes before sending it to the Internet. This masks the source of the traffic with little bandwidth overhead. Nevertheless, this approach requires cooperation from multiple homes, which makes it difficult to implement. In addition, it could result in long network latency if the homes are not geographically close.

Other approaches can leverage traffic shaping to prevent profiling. For instance, in Reference [102], Park et al. conceal smart home traffic patterns using dummy activities that have a high likelihood of occurrence. This is done considering the behavior of the inhabitants of that environment during the time of measurement. While this technique is energy efficient and supports low latency transmission of real data, its implementation requires the participation of many devices and can not shape traffic from genuine user activities. In another study [10], Apthorpe et al. propose a traffic shaping algorithm to make it challenging for an adversary to effectively distinguish dummy traffic patterns generated to mimic genuine user activities from the actual genuine traffic. However, this method only works against a passive network adversary and protects only traffic rate metadata such as packet times and sizes. This approach needs to be used with other methods to protect the categorical metadata such as protocol, IP address, and DNS hostnames. Likewise, the bandwidth overheads required to reduce the adversary confidence varies with respect to the type of device being protected. In fact, most of the existing traffic shaping techniques depend on effectively mimicking and realistic timing fake user activities.

### 5.5 Command and Phonetic Analysis

Here, we discuss countermeasures aiming at mitigating the issues of malicious skills. In particular, the skill vulnerabilities exploiting the interaction between the user and the third-party skill services.

Zhang et al. [155] present a system that examines the skill's response and the user's utterance to detect malicious skills that pretend to hand over control to another skill and deceive users into thinking that they are interacting with a different skill. The system relies on a User Intention Classifier (UIC) and a Skills Response Checker (SRC). The SRC semantically analyzes the skill response and compares it against utterances from a black-list of malicious skill responses to flag off any malicious response. While the user UIC, however, protects the user by checking their utterances to correctly determine their intents of context switches.[4] This is done by matching the meaning of what the user says to the context of the skill the user is presently interacting with and also that of the system commands. They also consider the link between what the user says and the skill that they are currently using. UIC complements the SRC, and their system reports an overall detection precision rate of 95.60%. Nevertheless, one key shortcoming of this system is the difficulty in implementing a generic UIC due to variation in Natural language-based command and how to distinguish legitimate commands.

In a similar study, Kumar et al. [67] suggests performing phonetic and text analysis for every new skill's invocation name to mitigate voice squatting attacks. They check whether the new skill's invocation name can be mistaken with an existing one, vetting then the creation of the clashing skill. Their solution is similar to what is currently being implemented during domain registration, where registrars do not register domain names that resemble that of popular domains.

### 5.6 New Architecture

In this section, we discuss a countermeasure that proposes a novel architecture for SPA, different from the one described in Section 2.1. In particular, we discuss the work proposed by Coucke et al. [26], which proposes changes to the architecture, particularly in terms of the speech recognition functionality. Coucke et al. [26] present a *privacy by design spoken Language Understanding platform* that does not send user queries to the cloud for processing. The speech recognition and the intent extraction are done locally on the SPA devices themselves using a partially trained model with *crowd-sourced* data and using *semi-supervised* learning. Many use cases do not need Internet access. However, when the use case requires internet access, such as when data needs

[4]This is examining the intents of changing from one task to the other.

to be retrieved or transmitted to an Internet service, then the system processes the data within the SPA device where it was generated rather than in the cloud. This makes it hard for an adversary to perform a mass attack as they can only target a single user or device at once. With such an infrastructure, issues related to *always on always listening*, *cloud*, and *third-party access*, have limited impact, since the data is processed locally. Besides, it allows personalizing the wake-up word, mitigating the wake-up word vulnerability introduced in Section 3.1.1. However, the platform requires a user to specify the skills on which their assistant will be trained on. Hence, such an assistant can only work within predefined scopes of the selected skills on which their model was trained, thereby restricting their capabilities to only those skills used for their training. It is important to also note that, although this infrastructure modifies the existing SPA architecture so that speech recognition and intent identification is conducted locally, it does not completely eliminate data transmission to other devices or cloud services. The SPA still communicates with other connected devices or cloud services depending on the context of use. This means that attacks like the one described in Reference [116] may still be possible.

## 6 DISCUSSION AND OPEN CHALLENGES

Building on the analysis and categorization of the related literature studied in the previous sections, we then offer a synthesis and summary of this review and suggest future research areas.

One can easily observe in Table 1 that vulnerabilities related to weak authentication are the most exploited flaws. The *wake-up word* and the *always listening features* are typically combined and can be described as the gateway of synthesized speech attacks. No related works currently exploit the multiuser environment and external party access. We also observed that the majority of the attacks target point #1 of the architecture: the point of interaction between the users and the SPA devices as it requires an attacker with lower capabilities. Although few attacks exploit more than one point of the architecture—e.g. [42, 67, 155], none is observed at point #5, point #7 and #8 even though attacks targeting those architectural elements seem possible as discussed in Section 4. Similarly, Table 2 shows that countermeasures for *weak authentication* vulnerabilities, and in particular countermeasures toward mitigating synthesized speech have received wide attention in the literature. Taking both Tables 1 and 2, we can see a concentration of research efforts toward one particular part of the whole SPA architecture, the direct interaction between the user and the smart speaker—or point #1 of the architecture. While indeed, this is an important part of the architecture, SPA should consider security in a holistic manner. This shows that despite the growing research efforts in security and privacy in SPA, we, as a community, also need to recognise and tackle SPA problems that go beyond that point of the architecture. Based on our findings, we suggest a number of open challenges in SPA. These include: (i) a practical evaluation of existing attacks and countermeasures, (ii) making authentication and authorization stronger as well as smarter, (iii) building secure and privacy-aware speech recognition, (iv) conducting systematic security and privacy assessments to understand the SPA eco-system and associated risks better, (v) increasing user awareness and the usability of security and privacy mechanisms in SPA, and (vi) understanding better profiling risks and potential countermeasures. All of which are discussed below in the following subsections.

### 6.1 Practical Evaluation of Existing Attacks and Countermeasures

We observed that many of the attacks target the underlying hardware of the voice infrastructure. For instance, References [117, 154] use high frequencies signal to attack the non-linearity in SPA devices microphones. While some of these attacks synthesize speech in a way that may be intelligible to humans and easily noticed by users in proximity [150], other attacks synthesize speech in a way that is unintelligible to the users [117, 154]. Thus, one could argue that the

second type of attack is more likely to be successful in practice than the first type. Our study also revealed that many of the attacks require different domain-specific knowledge to be successful, which might not always be available. For example, attacks conducted in References [42, 117, 142] need knowledge of the machine classifiers, while the one demonstrated in Reference [155] requires the understanding of the SPA skills invocation model. In some cases, this knowledge is available or can be reverse-engineered from interactions with the SPA and their architecture. However, beyond these observations that we can derive from a literature review, some important questions remain unanswered, such as: (1) What is the severity of the existing attacks? (2) What is the likelihood of success of these attacks in practice? (3) What is the cost associated with existing attacks and countermeasures? (4) What is the effectiveness of these countermeasures? and (5) How usable are these countermeasures?

### 6.2 Making Authentication Stronger

Despite receiving most of the attention in terms of countermeasures, with some of the issues and attacks having a counterpart countermeasure, weak authentication issues have not yet been completely addressed. As discussed earlier, many of the attacks targeting the SPA system exploit its weak authentication, especially the *always on, always listening features*. This attack is usually combined with other vulnerabilities. Although one could say that the *always on, always listening features* improve the responsiveness of the devices by making resources available to the user before they start uttering commands, the security and privacy risks may outweigh the benefit. Several independent input variables such as voice, video, manual gestures, touch, graphics, gaze, and others like the solution proposed in Reference [61] could be combined to make authentication stronger. However, most SPA are designed without environmental sensors. The lack of environmental sensors makes it challenging to implement context-aware authentication systems that could sense the physical environment, and leverage such information to adjust the security parameters accordingly. Also, there may be privacy issues and concerns when using even more personal information (e.g., video). Likewise, current authentication mechanisms in integrated technologies like other smart home devices are decentralized. Each integrated technology has its own authentication mechanism. By implementing a centralized mechanism, potentially in an SPA, a user could access multiple integrated technologies by authenticating only once. This would enhance usability by lessening the authentication burden on users and improving security as it would ensure consistent authentication across smart home devices. However, this needs to be implemented carefully so as not to create a single point of failure.

Future research can also consider how communication protocols may improve current authentication mechanisms in SPA. There are examples of how these mechanisms can be used in other systems such as remote car unlocking and contactless payment, where they are becoming an effective way to verify users' presence [14]. Popular among them are the distance-bounding protocols, which can be used to authenticate the user and access their location. These protocols have proven to be practicable especially in a system that is susceptible to distance-based frauds. Distance-bounding protocols are based on timing the delay between when a verifier sends a challenge to the moment the response is received. This allows the verifier to detect a third-party interference as any sudden delay in the proper response, which is considered to be the result of a delay due to a long-distance transmissions [14, 143]. Nevertheless, the effectiveness of this protocol depends on getting the correct propagation time.

### 6.3 Enhanced Authorization Models and Mechanisms

More flexible access control and authorization models and mechanisms are needed. These mechanisms should be able to dynamically authorize and adapt permissions to users based on the current

context and their preferences. According to a recent study, users preferred authorization policies in smart homes are affected by some distinct factors [49]: (i) the capabilities within a single device, (ii) who is trying to use that capability, (iii) and the context of use. Hence, designing authorization models that consider SPA capabilities and the context of use may help create authorization rules that adequately balance security, privacy, and functionality. In fact, similar models have already been implemented successfully in other domains like smartphones [94]. Furthermore, we have observed that SPA requires more fine-grained authorization mechanisms. This not only applies to the voice of the user itself but also to the data that can be obtained from how users interact with the devices. In particular, these interactions can be used to infer, for instance, the sleeping patterns of a user, as discussed earlier.

Novel authorization models and mechanisms for SPA should consider not only single users but also multiple users. However, there are no security and privacy mechanisms for SPA that considers *multi-user environment* issues. This is important, as even if SPA would support multiple accounts, it is a common practice to share accounts between multiple users [80] (especially if one of the accounts has more privileges). The lack of proper authorization can prompt insider misuse, e.g., members of the household spying on their partners [39], which can be particularly problematic in the case of intimate partner abuse [81]. Moreover, smart home data is relational and it usually refers to a group of people collectively [103], e.g., if there is a way to infer whether there is someone at home or not, then this already gives information that can be sensitive to everyone living there. Some general-purpose smart home privacy-enhancing IoT infrastructures like the Databox [103] recognize the multiuser problem but no solution has been proposed yet in general for smart homes or in particular for multiuser sharing management in SPA. A great deal of research on methods and tools to help users manage data sharing in multiuser and multiparty scenarios have been proposed for social media (see Reference [131] for a survey), and particular methods for detecting and resolving multiuser data sharing conflicts, such as Reference [130], could be adapted from there or used to inspire multiuser solutions for the SPA case.

Furthermore, the existing SPA architecture supports only permission-based access control on sensitive data, which is insufficient at controlling how third-party skills use data once they get access. Future research should study how to implement a framework that allows users to pronounce their intended data flow patterns. Similar frameworks [37, 54] have been successfully applied in smartphones for IoT apps. Also, there is a lack of authorization frameworks for data generated during user interactions with a third-party skill, which is one of the personal data assets mentioned in Section 2.3. Novel authorization mechanisms that allow users to specify, monitor and control what data can be shared with those that have no direct access to the SPA architecture, under what condition should the data be shared (reason), how it should be shared (means) and what it can be used for (purpose) could also help address the issue of external parties.

### 6.4 Secure and Privacy-aware Speech Recognition

NLP and ML models are used in conjunction for speech recognition. Protecting these models against manipulation, e.g., through well-crafted adversarial inputs as pointed out in Section 3.4, becomes paramount. It is apparent from Tables 1 and 2 above that there are many attacks exploiting adversarial ML and NLP issues, and there are substantially more attacks than defenses studied in the related literature. SPA providers need to consider adversarial examples when developing their speech recognition models. However, that is not an easy task, and more research is required in this direction. Some existing countermeasures used in other domains such as adversarial training and distillation could help to develop robust ML models for speech recognition in SPA, but they can be defeated using black-box attacks or attacks that are constructed on iterative optimization [17]. Also, validating the input and reprocessing it to eliminate possible adversarial

manipulations before it is fed to the model is a countermeasure that greatly depends on the domain, and is subjected to environmental factors [99]. Likewise, testing is not enough to secure ML, as an adversary can use a different input from those used for the testing process [43].

Furthermore, the performance of the current speech recognition system still deserves improvement as shown earlier—recall that these systems often find it difficult to (i) understand words with similar phonemes [67], (ii) understand different but similar words, and (iii) resolve variation in natural language-based command words [154]. Since the word error rate (WER) is the common metric used for evaluating the performance of automatic speech recognition systems [24], it may be easy for an adversary to craft an adversarial input that could maximize the WER of the speech recognition system by exploiting the NLP framework and the ML techniques. This is shown in Reference [154], where the speech recognition system is exploited to manipulate the intent the system understands from the user's command.

Beyond security, obtaining valuable information from big data while still protecting user's privacy has become interesting research in data analysis. While SPA providers let users review and delete their voice recordings, a recent study shows that users are unaware (or do not use) those privacy controls [69]. It is also unclear how effective these controls actually are even if used; e.g., these controls allow the user to delete particular raw utterances but they cannot delete what could be inferred from them (i.e., the model) [60]. In light of this, SPA vendors need to understand the privacy challenges of machine learning. For instance, although most existing SPA providers aim to ensure privacy while processing users' voice in the cloud, that is a difficult endeavor with current SPA architectures. With edge computing gradually coming into the limelight, data can now be processed locally, where it is generated, rather than being transmitted to a centralized data processing centre [115]. This helps to reduce the current dependency on the Internet and eliminate the necessity of putting sensitive data into the cloud. While related work [26] addresses this direction with a decentralized voice processing platform, it is challenging to build a general-purpose SPA using such platforms. This is because SPA developed with such platforms can only work within predefined scopes of the selected skills on which their model was trained. Therefore, there is a need for future efforts on how to make voice processing privacy-preserving without hindering SPA's capabilities effectively.

### 6.5 AI-based Security and Privacy

In addition to using AI techniques for SPA functionality, e.g., speech recognition, they could also be used to make SPA more secure and aid users in managing their privacy as they see fit. AI techniques would include not only data-driven techniques like ML but also knowledge-based techniques such as normative systems and argumentation, which have been successfully used to develop intelligent security and privacy methods in other domains [128, 132]. AI techniques could be used to address the issue of *always on always listening* and*synthesized speech* under the weak authentication vulnerabilities. For instance, it could be applied to detect malicious commands being spoken to the SPA devices (i.e., to make authentication stronger and more resilient to attacks). Likewise, it could be used to solve the issue of *multi-user authorization and over-privileged skills* by applying it to help primary users configure the permissions they grant to other users and third-parties skills, respectively. Similar research has already been shown to detect intrusions [25] and to help users in other domains like mobile App permission management [96] and Social Media privacy settings and data sharing [84]. As for the speech recognition, these ML-based methods need to be engineered considering adversarial cases [43].

Examples of the use of knowledge-based AI techniques include the use of norms, which have been widely explored in recent years, especially to reduce the autonomy of autonomous and intelligent systems to conform to decent behaviors [27]. Norms are usually delineated formally using

deontic logic to state what is permissible, obligatory, and prohibited, providing a rich framework to express context-dependent policies, e.g., based on Contextual Integrity [95], and they can be defined, verified, and monitored for socio-technical systems like SPA [29, 56]. For instance, norms would be beneficial to avoid issues like the case discussed in Reference [146], where a private conversation is recorded by an Alexa and forwarded to a random contact, as a norm could specify the type of conversations that may or may not be shared with particular contacts and that norm could be verified and monitored for compliance. Another example is norms that govern multiuser interactions with the SPA as discussed in Section 6.3. Norms for SPA could be elicited automatically as in Reference [28] or by crowd-sourcing the acceptable flows of information as in Reference [38]. Another knowledge-based AI technique like automated negotiation [15, 134] could be used to help SPA users navigate the trade-offs and negotiate consent in the very complex SPA ecosystem, including third-party skills and smart devices. For instance, instead of having the user manually inspecting and approving every permission for the many third-party skills that may request them (as it happens now in SPA ecosystems like Amazon Alexa and Google Home), the SPA could automatically negotiate those permissions with the third-party skills. This can be done, however, always in a way in which consent could be revocable and access patterns apparent to the user on-demand, allowing reactive and dynamic data sharing adjustment. Finally, other AI techniques like computational trust [105] could be used to choose and only share data with third-party skills and smart devices that are privacy-respecting and trustworthy.

### 6.6 Systematic Security and Privacy Assessments

SPA are a type of cyber-physical system. Previous research looked at how the assurance techniques and testing methodologies most commonly used in conventional IT systems [106] apply to cyber-physical systems, including penetration testing, static and dynamic analysis, fuzzing, and formal verification. However, it is still unclear how these security testing techniques apply to the SPA system and what are the practices used by third-party developers in this ecosystem. Assurance techniques are known to have different cost-effectiveness in practice [133], and that cost-effectiveness for one very same assurance technique has been shown to vary across different cyber-physical systems [12], such as Industrial Control Systems [66]. Therefore, a direction for future research is to study and evaluate how these assurance techniques will perform for the case of SPA and whether or not SPA's unique features like voice recognition and its integration with other technologies like the cloud and other smart devices require novel techniques or methodologies. For instance, the known potential to have composite vulnerabilities that exploit both the physical and the IT part of cyber-physical systems [22, 23] has already been shown to also apply to SPA, e.g., Reference [117]. Additionally, authors in Reference [154] show that physical properties can be used to compromise the SPA by using high frequencies signals to attack the non-linearity in SPA devices microphones as detailed above in Section 4.1. A set of key research questions to answer revolve around which assurance techniques can be used to improve security in SPA systems (see Appendix A in Reference [66]). In particular: (1) Can a review of standards and procedures be used to mitigate security risks in SPA systems? (2) Can we run dynamic analysis techniques over components of the SPA architecture? and (3) Can we devise a methodology to provide an independent validation when many components of an SPA system are hosted in the cloud?

Future work should also look at the best and most systematic way to conduct privacy assessments in SPA [149]. However, it remains unclear how many privacy violations there are in the wild of the third-party ecosystem and what is the extent of such violations. Measuring privacy violations systematically is particularly challenging as privacy policies are usually unstructured data. Thus, it is hard to infer properties from them automatically. Of particular interest might be to study the (extent of) traceability between the actions of the data specified in privacy policies, such as

those in the privacy policies of the third-party skills developers in SPA, and the related data operations obvious to users via SPA and/or associated smartphone interfaces, which will also be crucial to help tackle the current *weak authorization* and *profiling* issues of SPA. One important research question is whether related works could be adopted to measure policy traceability in the SPA domain. Methodologies could be adapted from the social media [9] and smartphone apps [85], which already showed the extent of traceability in these domains, together with methods to help developers automatically map traceability between policies and operationalized controls and maintain it through the development cycle [8]. As real breaches happen (e.g., Reference [146]), methods to study whether there are gaps in security and privacy policies, such as Reference [57] applied to SPA, would also be helpful. Thus, a systematic study could measure how many privacy policies are complete and how many are broken for the third-party SPA ecosystem. Likewise, a longitudinal study is required to comprehend the SPA skill's ecosystem to understand the type of skills available, the capabilities they have, how they are being used, and who is behind them (number of third-party developers, etc.). This will further ensure a better understanding of the different risks that the ecosystem presents and aid in formulating appropriate security and privacy policies for the users.

### 6.7 Increasing User Awareness

Although implementing a technical defensive measure might go a long way in mitigating some of the identified risks, effective countermeasures will be difficult without better user awareness. Research shows that the lack of awareness about data practices in smart home devices affect users' security and privacy practices [140]. Some SPA users are not very concerned when it comes to the security and privacy issues in SPA [152], as they believe they are not valuable targets for attackers [140], or they simply exhibit inaccurate and incomplete mental models of the SPA ecosystem [1]. Therefore, it is essential that users understand the risks and threats present in the SPA ecosystem, including the assets that can be compromised and why they need protection for better risk management. Users should be well informed to adopt best practices and even understand what key steps they have to take when either their security or privacy is breached. One crucial way of keeping SPA users informed is to design usable privacy notice that helps them understand and manage their data in SPA, accompanied with usable security and privacy mechanisms (as discussed below in Section 6.8). Privacy notices must be relevant, actionable and understandable as discussed in Reference [119], and their design should be considered along four main dimensions: (1) timing, when should a privacy notice be presented; (2) channel, how should the privacy notice be delivered; (3) modality, how the information should be conveyed; and (4) control, how choice options are integrated into the privacy notice. Another example would be leveraging the already discussed assessments in Section 6.6, to produce a white (or black) list of third-party skills based on the level of security and/or privacy they offer considering the results of the assessments.

### 6.8 Usability of SPA Security and Privacy Mechanisms

While users' awareness is crucial in understanding the system's risks, awareness without usable security and privacy controls mechanisms may not be effective in mitigating these risks. For instance, some SPA users, while aware of some risks, do not know how they can protect themselves [1]. In addition to knowing the mechanisms they could use to protect themselves (such as those to achieve a basic level of cyber hygiene [129] but in the SPA domain), users should be able to utilize any SPA security and privacy mechanisms in a convenient manner that does not affect usability or functionality of SPA. This is because convenience and connectivity are important concerns for smart home users, influence their perceptions and opinions, and their attitude toward external entities that design and regulate SPA [157]. Nonetheless, these measures' primary concern is that they have an important impact on usability, as they clash with the sought "hands-free"

experience when interacting with SPA. In some other cases where non-technical coping strategies may not be available, SPA users are merely avoiding the SPA functionality they perceive to be risky, e.g., some SPA users only create shopping lists through the SPA but buy the items using the traditional web interface as they perceive buying through the SPA as risky and do not know how to protect themselves [1].

From all the technical countermeasures that we surveyed in this article (see Section 5), the vast majority of them did not explicitly consider usability. What is worse, there were cases in which some potentially negative usability impacts introduced by the countermeasures were clearly apparent such as where users need to use a wearable device like in References [35, 61], and where the SPA capability might be restricted [26]. Future work should conduct rigorous and systematic studies of the usability of the countermeasures already proposed to assess how usable they really are. Beyond these usability studies of existing countermeasures, future work on SPA security and privacy mechanisms should also consider usability from the onset, not as an afterthought. For instance, novel SPA security and privacy mechanisms should avoid requiring extensive user involvement. Otherwise, it has been shown they may not be used [153]. A potential avenue to explore as future work regarding this example could be the AI-based techniques discussed in Section 6.5, which could be leveraged to predict user's preferences and help users set security and privacy controls much easier and with less involvement.

### 6.9 Profiling Attacks and Defences

Regarding profiling, we can clearly see in Table 1 that few attacks have been reported on this. Some of these attacks make some hard assumptions, like having access to all cloud data about a user through their user account. We believe that further research is needed to assess whether other types of more sophisticated profiling could be conducted with access to less information. Furthermore, the community needs to understand whether tracking, which is pervasive across the web [82], could also apply and be feasible across the SPA ecosystem. In terms of defenses, we can also see in Table 2 the lack of work in this area. Some of the challenges we mentioned before would indeed help alleviate profiling such as user awareness and usable controls (Sections 6.7 and 6.8), systematic privacy assessments (Section 6.6), and knowledge-based AI techniques to express/verify norms about how data are collected and use of data across the SPA ecosystem (Section 6.5). However, other open challenges would remain, and profiling-specific countermeasures are also needed. For instance, SPA traffic needs to be properly obfuscated and masked to encode user's interaction with the devices in addition to the existing encryption mechanisms already in place. Note that current encryption mechanisms are not sufficient to avoid traffic profiling as shown in Reference [11]. Beyond differential-private approaches like the countermeasure introduced earlier [72], one possible avenue would be to adapt existing mechanisms to the case of SPA, such as traffic morphing techniques [148] to prevent statistical traffic analysis. This can be done by altering one category of traffic to look like another one. However, this and most other existing traffic analysis countermeasures are vulnerable as they only obfuscate exposed features of the traffic by muffling this features and adding dummy packets. Thus, they are unable to prevent the leakage of many identifying information [33]. Another avenue could be based on mix networks [137] and/or onion routing [88]. However, both of them may also be vulnerable to attack. For instance, mixing is susceptible to long term correlation and sleeper attacks [137], and onion routing is susceptible to an adversary correlating the traffic [136] and to misconfigured and malicious relays [55].

## 7 CONCLUSIONS

This article analyzes and classifies the security and privacy issues associated with SPA and how a range of malicious actors can exploit them to harm the security and privacy of end-users. We

have shown that the attack surface of this increasingly popular technology is vast. We have noted that the interaction between the users and the SPA devices is currently the weakest link. However, we have identified a wide range of attacks that can put users at stake. In as much as there is no single panacea solution for all security issues, the proper understanding of security pitfalls will go a long way in enabling manufacturers, researchers, and developers to design and implement robust security control measures. Although there is already very active research on securing intelligent assistants, few of the approaches consider the whole picture of the complex architecture SPA have. We particularly highlighted open challenges for future research that we deem of critical importance, including making authentication stronger, enhancing authorization models and mechanisms, building secure and privacy-aware speech recognition, conducting systematic security and privacy assessments, developing AI-based security and privacy countermeasures, improving user awareness and usability, and studying further profiling attacks and defenses.

As future work, we would like to keep on expanding our understanding of the different open challenges presented above. While we included all available articles at the time obtained through the method stated earlier, SPA security and privacy is a fast-moving field still in its infancy. We hope this survey serves researchers to help prioritize the most promising areas to improve our understanding of attacks on SPA and to devise usable ways to counter them. Also, most of the literature we found focused on the two most popular SPAs—Amazon Alexa and Google Assistant. However, there are many other SPAs (e.g., Microsoft Cortana). Even though they may have a similar architecture, there may be specific issues with them not covered in this article, so expanding our current article in this regard would also be an exciting line of future work.

## REFERENCES


[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Proceedings of the 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, Santa Clara, CA. Retrieved from https://www.usenix.org/conference/soups2019/presentation/abdi.

[2] Efthimios Alepis and Constantinos Patsakis. 2017. Monkey says, monkey does: Security and privacy on voice assistants. *IEEE Access* 5 (2017), 17841–17851. DOI: https://doi.org/10.1109/access.2017.2747626

[3] Amazon. 2017. About Alexa Voice Profiles. Retrieved from https://www.amazon.com/gp/help/customer/display.html?nodeId=202199440.

[4] Amazon. 2018. The Alexa Skill Store for France is a Fast Growing Land of Opportunity. Retrieved from https://developer.amazon.com/docs/ask-overviews/understanding-the-different-types-of-skills.html.

[5] Amazon. 2019. All-new Echo Show (2nd Gen). Retrieved from https://www.amazon.com/All-new-Echo-Show-2nd-Gen/dp/B077SXWSRP.

[6] Amazon. 2019. Configure Permissions for Customer Information in Your Skill. Retrieved from https://developer.amazon.com/en-US/docs/alexa/custom-skills/configure-permissions-for-customer-information-in-your-skill.html.

[7] Amazon. n.d. Understand How Users Interact with Skills. Retrieved from https://developer.amazon.com/en-GB/docs/alexa/ask-overviews/understanding-how-users-interact-with-skills.html.

[8] Pauline Anthonysamy, Matthew Edwards, Chris Weichel, and Awais Rashid. 2016. Inferring semantic mapping between policies and code: The clue is in the language. In *Proceedings of the International Symposium on Engineering Secure Software and Systems*. Springer, 233–250.

[9] Pauline Anthonysamy, Phil Greenwood, and Awais Rashid. 2013. Social networking privacy: Understanding the disconnect from policy to controls. *Computer* 46, 6 (2013), 60–67.

[10] Noah Apthorpe, Danny Yuxing Huang, Dillon Reisman, Arvind Narayanan, and Nick Feamster. 2019. Keeping the smart home private with smart(er) IoT traffic shaping. *Proc. Privacy Enhanc. Technol.* 2019, 3 (2019), 128–148. DOI: https://doi.org/10.2478/popets-2019-0040

[11] Noah Apthorpe, Dillon Reisman, and Nick Feamster. 2017. A smart home is no castle: Privacy vulnerabilities of encrypted IoT traffic. Retrieved from http://arxiv.org/abs/1705.06805.

[12] Sara Abbaspour Asadollah, Rafia Inam, and Hans Hansson. 2015. A survey on testing for cyber physical system. In *Proceedings of the IFIP International Conference on Testing Software and Systems*. Springer, 194–207.

[13] Ava Mutchler. 2018. Google Assistant App Total Reaches Nearly 2400. Retrieved from https://voicebot.ai/2018/01/24/google-assistant-app-total-reaches-nearly-2400-thats-not-real.

[14] Gildas Avoine, Muhammed Ali Bingol, Ioana Boureanu, Srdjan Capkun, Gerhard Hancke, Suleyman Kardas, Chong Hee Kim, Cedric Lauradoux, Benjamin Martin, Jorge Munilla, Alberto Peinado, Kasper Bonne Rasmussen, Dave Singelee, Aslan Tchamkerten, Rolando Trujillo-Rasua, and Serge Vaudenay. 2018. Security of distance-bounding: A survey. *ACM Comput. Surv.* 51, 5, Article 94 (Sept. 2018), 33 pages. DOI: https://doi.org/10.1145/3264628

[15] T. Baarslag, A. T. Alan, R. C. Gomer, I. Liccardi, H. Marreiros, E. Gerding, et al. 2016. Negotiation as an interaction mechanism for deciding app permissions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'16), Extended Abstracts*. 2012–2019.

[16] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *Proceedings of the 25th USENIX Security Symposium (USENIXSecurity'16)*. USENIX Association, Austin, TX, 513–530. Retrieved from www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini.

[17] Nicholas Carlini and David A. Wagner. 2016. Towards evaluating the robustness of neural networks. Retrieved from http://arxiv.org/abs/1608.04644.

[18] Nicholas Carlini and David A. Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *Proceedings of the IEEE Security and Privacy Workshops (SP'18)*. 1–7. https://doi.org/10.1109/SPW.2018.00009

[19] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems (ICDCS'17)*. DOI: https://doi.org/10.1109/icdcs.2017.133

[20] Hyunji Chung and Sangjin Lee. 2018. Intelligent virtual assistant knows your life. *CoRR* abs/1803.00466 (2018). arxiv:1803.00466

[21] Hyunji Chung, Jungheum Park, and Sangjin Lee. 2017. Digital forensic approaches for Amazon Alexa ecosystem. *Digital Investigat.* 22 (2017), S15 to S25. DOI: https://doi.org/10.1016/j.diin.2017.06.010

[22] Pierre Ciholas, Aidan Lennie, Parvin Sadigova, and Jose M. Such. 2019. The security of smart buildings: A systematic literature review. *arXiv preprint arXiv:1901.05837*. https://arxiv.org/pdf/1901.05837.pdf.

[23] Pierre Ciholas and Jose M Such. 2016. Composite vulnerabilities in Cyber Physical Systems. In *Proceedings of the First International Workshop held on 06 April 2016 in conjunction with the International Symposium on Engineering Secure Software and Systems, London, UK*. https://eprints.lancs.ac.uk/id/eprint/79052/4/Proceedings_serecin_2016.pdf.

[24] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. 2017. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*. 6980–6990.

[25] Emilio Corchado and Álvaro Herrero. 2011. Neural visualization of network traffic data for intrusion detection. *Appl. Soft Comput.* 11, 2 (2011), 2042–2056.

[26] Alice Coucke, Alaa Saade, Adrien Ball, Theodore Bluche, Alexandre Caulier, David Leroy, Clement Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Mael Primet, and Joseph Dureau. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *CoRR* abs/1805.10190 (2018).

[27] Natalia Criado, Estefania Argente, and V. Botti. 2011. Open issues for normative multi-agent systems. *AI Commun.* 24, 3 (2011), 233–264.

[28] Natalia Criado and Jose M. Such. 2015. Implicit contextual integrity in online social networks. *Info. Sci.* 325 (2015), 48–69.

[29] Natalia Criado and Jose M. Such. 2016. Selective norm monitoring. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16)*. 208–214.

[30] Ayse Cufoglu. 2014. User profiling-a short review. *Int. J. Comput. Appl.* 108, 3 (2014). Retrieved from https://research.ijcaonline.org/volume108/number3/pxc3900179.pdf.

[31] Tamara Denning, Tadayoshi Kohno, and Henry M. Levy. 2013. Computer security and the modern home. *Commun. ACM* 56, 1 (Jan. 2013), 94–103. DOI: https://doi.org/10.1145/2398356.2398377

[32] Google Developer. 2019. Developer Preview of Local Home SDK. Retrieved from https://developers.googleblog.com/2019/07/developer-preview-of-local-home-sdk.html.

[33] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton. 2012. Peek-a-Boo, I still see you: Why efficient traffic analysis countermeasures fail. In *Proceedings of the IEEE Symposium on Security and Privacy*. 332–346. DOI: https://doi.org/10.1109/SP.2012.28

[34] Aarthi Easwara Moorthy and L. Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *Int. J. Hum. Comput. Interact.* 31, 4 (Apr. 2015), 307 to 335. DOI: https://doi.org/10.1080/10447318.2014.986642

[35] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous authentication for voice assistants. *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom'17)*. DOI: https://doi.org/10.1145/3117811.3117823

[36] E. Fernandes, J. Jung, and A. Prakash. 2016. Security analysis of emerging smart home applications. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'16)*. 636–654. DOI: https://doi.org/10.1109/SP.2016.44

[37] Earlence Fernandes, Justin Paupore, Amir Rahmati, Daniel Simionato, Mauro Conti, and Atul Prakash. 2016. FlowFence: Practical data protection for emerging IoT application frameworks. In *Proceedings of the 25th USENIX Security Symposium (USENIXSecurity'16)*. USENIX Association, Austin, TX, 531–548. Retrieved from https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/fernandes.

[38] R. Fogues, P. K. Murukannaiah, J. M. Such, and M. P. Singh. 2017. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Trans. Comput.-Hum. Interact.* 24, 1 (2017), 5.

[39] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. A stalker's paradise: How intimate partner abusers exploit technology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 667.

[40] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer attitudes towards privacy and security in home assistants(CHI EA '18). ACM, New York, NY, Article LBW050, 6 pages. DOI: https://doi.org/10.1145/3170427.3188448

[41] Sri Garimella, Arindam Mandal, Nikko Strom, Björn Hoffmeister, Spyridon Matsoukas, and Sree Hari Krishnan Parthasarathi. 2015. Robust i-vector based adaptation of DNN acoustic model for speech recognition. In *Proceedings of the INTERSPEECH Conference*.

[42] Yuan Gong and Christian Poellabauer. 2017. Crafting adversarial examples for speech paralinguistics applications. Retrieved from http://arxiv.org/abs/1711.03280.

[43] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. *Commun. ACM* 61, 7 (2018), 56–66.

[44] Google. 2017. Set up Multiple Users for Your Speaker or Smart Display. Retrieved from https://support.google.com/assistant/answer/9071681.

[45] Google. 2018. Actions on Google. Retrieved from https://developers.google.com/actions/samples/.

[46] Google. 2018. Invocation and Discovery. Retrieved from https://developers.google.com/actions/sdk/invocation-and-discovery.

[47] William Haack, Madeleine Severance, Michael Wallace, and Jeremy Wohlwend. 2017. Security analysis of Amazon Echo. Retrieved from https://courses.csail.mit.edu/6.857/2017/project/8.pdf

[48] Jo E. Hannay, Dag I. K. Sjoberg, and Tore Dyba. 2007. A systematic review of theory use in software engineering experiments. *IEEE Trans. Softw. Eng.* 33, 2 (2007), 87–107. DOI: https://doi.org/10.1109/tse.2007.12

[49] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Durmuth, Earlence Fernandes, and Blase Ur. 2018. Rethinking access control and authentication for the home Internet of Things (IoT). In *Proceedings of the 27th USENIX Conference on Security Symposium (SEC'18)*. USENIX Association, Berkeley, CA, 255–272.

[50] J. Hirschberg and C. D. Manning. 2015. Advances in natural language processing. *Science* 349, 6245 (2015), 261–266. DOI: https://doi.org/10.1126/science.aaa8685

[51] Helena Horton. 2018. Amazon Alexa Recorded Owner's Conversation and Sent to "random" Contact, Couple Complains. Retrieved from www.telegraph.co.uk/news/2018/05/25/amazon-alexa-recorded-owners-conversation-sent-random-contact/.

[52] Matthew B. Hoy. 2018. Alexa, siri, cortana, and more: An introduction to voice assistants. *Med. Ref. Serv. Quart.* 37, 1 (2018), 81–88. DOI: https://doi.org/10.1080/02763869.2018.1404391

[53] Texas Instruments. 2013. AN-1973 Benefits and Challenges of High-Frequency Regulators. Retrieved from http://www.ti.com/lit/an/snva399a/snva399a.pdf.

[54] Yunhan Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Earlence Fernandes, Zhuoqing Mao, and Atul Prakash. 2017. ContexIoT: Towards providing contextual integrity to appified IoT platforms. In *Proceedings 2017 Network and Distributed System Security Symposium*. DOI: http://dx.doi.org/10.14722/ndss.2017.23051

[55] George Kadianakis, Claudia V. Roberts, Laura M. Roberts, and Philipp Winter. 2017. Anomalous keys in Tor relays. Retrieved from http://arxiv.org/abs/1704.00792.

[56] Özgür Kafali, Nirav Ajmeri, and Munindar P. Singh. 2016. Revani: Revising and verifying normative specifications for privacy. *IEEE Intell. Syst.* 31, 5 (2016), 8–15.

[57] Özgür Kafali, Jasmine Jones, Megan Petruso, Laurie Williams, and Munindar P. Singh. 2017. How good is a security policy against real breaches? A HIPAA case study. In *Proceedings of the IEEE/ACM 39th International Conference on Software Engineering (ICSE'17)*. IEEE, 530–540.

[58] Candace Kamm. 1995. User interfaces for voice applications. *Colloq. Paper* 92 (1995), 10031–10037.

[59] Heather Kelly. 2017. Apple's HomePod is Coming. Here's What You Need to Know About Smart Speakers. Retrieved from http://money.cnn.com/2017/06/08/technology/gadgets/apple-homepod-smart-speaker-faq/index.html.

[60] Makena Kelly and Nick Statt. 2019. Amazon Confirms it Holds on to Alexa Data Even if You Delete Audio Files. Retrieved from https://www.theverge.com/2019/7/3/20681423/amazon-alexa-echo-chris-coons-data-transcripts-recording-privacy.

[61] Veton Kepuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *Proceedings of the IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC'18)*, 99–103. DOI: https://doi.org/10.1109/ccwc.2018.8301638

[62] Bret Kinsella. 2018. Alexa Skill Store for France is a Fast Growing Land of Opportunity. Retrieved from https://voicebot.ai/2018/11/03/the-alexa-skill-store-for-france-is-a-fast-growing-land-of-opportunity/.

[63] Bret Kinsella. 2018. Amazon Introduces Skill Connections so Alexa Skills Can Work Together. Retrieved from https://voicebot.ai/2018/10/04/amazon-introduces-skill-connections-so-alexa-skills-can/.

[64] Bret Kinsella. 2018. The Information Says Alexa Struggles with Voice Commerce But Has 50 Million Devices Sold. Retrieved from https://voicebot.ai/2018/08/06/the-information-says-alexa-struggles-with.

[65] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering A systematic literature review. *Info. Softw. Technol.* 51, 1 (2009), 7–15. DOI: https://doi.org/10.1016/j.infsof.2008.09.009

[66] William Knowles, Jose M. Such, Antonios Gouglidis, Gaurav Misra, and Awais Rashid. 2015. Assurance techniques for industrial control systems (ics). In *Proceedings of the 1st ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy*. ACM, 101–112.

[67] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on amazon Alexa. In *Proceedings of the 27th USENIX Security Symposium (USENIXSecurity'18)*. USENIX Association, Baltimore, MD, 33–47.

[68] Angelica Lai. 2018. Sneaky Kid Orders $350 Worth of Toys on Her Mom's Amazon Account. Retrieved from https://mom.me/news/271144-sneaky-kid-orders-350-worth-toys-her-moms-amazon-account/.

[69] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. DOI: https://doi.org/10.1145/3274371

[70] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. 2017. Audio-replay attack detection countermeasures. Retrieved from http://arxiv.org/abs/1705.08858.

[71] Kwan-Min Lee and Clifford Nass. 2005. Social-psychological origins of feelings of presence: Creating social presence with machine generated voices. *Media Psychol.* 7, 1 (2005), 31–45. DOI: https://doi.org/10.1207/S1532785XMEP0701_2

[72] J. Liu, C. Zhang, and Y. Fang. 2018. EPIC: A differential privacy framework to defend smart homes against internet traffic analysis. *IEEE Internet Things J.* 5, 2 (Apr. 2018), 1206–1217. DOI: https://doi.org/10.1109/JIOT.2018.2799820

[73] N. D. Londhe, M. K. Ahirwal, and P. Lodha. 2016. Machine learning paradigms for speech recognition of an Indian dialect. In *Proceedings of the International Conference on Communication and Signal Processing (ICCSP'16)*. DOI: https://doi.org/10.1109/iccsp.2016.7754251

[74] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, New York, NY, 5286–5297. DOI: https://doi.org/10.1145/2858036.2858288

[75] Nishtha Madaan, Mohd Abdul Ahad, and Sunil M. Sastry. 2018. Data integration in IoT ecosystem: Information linkage as a privacy threat. *Comput. Law Secur. Rev.* 34, 1 (2018), 125–133. DOI: https://doi.org/10.1016/j.clsr.2017.06.007

[76] K. M. Malik, H. Malik, and R. Baumann. 2019. Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'19)*. 523–528.

[77] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy attitudes of smart speaker users. *Proc. Privacy Enhanc. Technol.* 2019, 4 (2019), 250–271.

[78] Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Bjorn Hoffmeister, and Arindam Mandal. 2018. Monophone-based background modeling for two-stage on-device wake word detection. Retrieved from http://sigport.org/2800.

[79] Taylor Martin. 2018. 12 Reasons to use Alexa in the Kitchen. Retrieved from https://www.cnet.com/how-to/how-to-use-alexa-in-the-kitchen/.

[80] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. 2016. She'll just grab any device that's closer: A study of everyday device and account sharing in households. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 5921–5932.

[81] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. [n.d.]. Security and privacy experiences and practices of survivors of intimate partner abuse. *IEEE Secur. Privacy* 5, 76–81.

[82] Jonathan R. Mayer and John C. Mitchell. 2012. Third-party web tracking: Policy and technology. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 413–427.

[83] Atif M. Memon and Ali Anwar. 2015. Colluding apps: Tomorrow's mobile malware threat. *IEEE Secur. Privacy* 13, 6 (2015), 77–81.

[84] Gaurav Misra and Jose M. Such. 2017. PACMAN: Personal agent for access control in social media. *IEEE Internet Comput.* 21, 6 (2017), 18–26.

[85] Gaurav Misra, Jose M. Such, and Lauren Gill. 2017. A privacy assessment of social media aggregators. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 561–568.

[86] Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Avi Patel, and Muttukrishnan Rajarajan. 2012. A survey on security issues and solutions at different layers of Cloud computing. *J. Supercomput.* 63, 2 (2012), 561–592. DOI: https://doi.org/10.1007/s11227-012-0831-5

[87] Ladislav Mosner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Bjorn Hoffmeister. 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. DOI: https://doi.org/10.1109/icassp.2019.8683422

[88] Steven J. Murdoch and Piotr Zielinski. 2007. Sampled traffic analysis by internet-exchange-level adversaries. In *Proceedings of the 7th International Conference on Privacy Enhancing Technologies (PET'07)*. Springer-Verlag, Berlin, Heidelberg, 167–183. Retrieved from http://dl.acm.org/citation.cfm?id=1779330.1779341.

[89] Chetan Naik, Arpit Gupta, Hancheng Ge, Mathias Lambert, and Ruhi Sarikaya. 2018. Contextual slot carryover for disparate schemas. In *Proceedings of the Interspeech Conference*. 596–600. DOI: https://doi.org/10.21437/Interspeech.2018-1035

[90] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are people polite to computers? Responses to computer-based interviewing systems1. *J. Appl. Soc. Psychol.* 29, 5 (1999), 1093–1109. DOI: https://doi.org/10.1111/j.1559-1816.1999.tb00142.x

[91] Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, and Tatsuya Mori. 2019. Poster. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS'19)*. DOI: https://doi.org/10.1145/3319535.3363274

[92] Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, and Tatsuya Mori. 2019. Poster: A first look at the privacy risks of voice assistant apps. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS'19)*. Association for Computing Machinery, New York, NY, 2633–2635. DOI: https://doi.org/10.1145/3319535.3363274

[93] Lily Hay Newman. 2018. Millions of Streaming Devices Are Vulnerable to a Retro Web Attack. Retrieved from https://www.wired.com/story/chromecast-roku-sonos-dns-rebinding-vulnerability/.

[94] Xudong Ni, Zhimin Yang, Xiaole Bai, A. C. Champion, and D. Xuan. 2009. DiffUser: Differentiated user access control on smartphones. In *Proceedings of the IEEE 6th International Conference on Mobile Adhoc and Sensor Systems*. 1012–1017. DOI: https://doi.org/10.1109/MOBHOC.2009.5337017

[95] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Rev.* 79 (2004), 119.

[96] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kevin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. 2017. Smarper: Context-aware and automatic runtime-permissions for mobile devices. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 1058–1076.

[97] OVUM. 2017. Virtual digital assistants to overtake world population by 2021. Retrieved from https://ovum.informa.com/resources/product-content/virtual-digital-assistants-to-overtake-world-population-by-2021.

[98] Constantinos Papayiannis, Justice Amoh, Viktor Rozgic, Shiva Sundaram, and Chao Wang. 2018. Detecting media sound presence in acoustic scenes. In *Proceedings of the Interspeech Conference*. 1363–1367. DOI: https://doi.org/10.21437/Interspeech.2018-2559

[99] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2018. Towards the science of security and privacy in machine learning. In *Proceedings of the 3rd IEEE European Symposium on Security and Privacy*.

[100] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. 2016. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. Retrieved from http://arxiv.org/abs/1605.07277.

[101] Ankur P. Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*. Retrieved from https://arxiv.org/abs/1606.01933.

[102] Homin Park, Can Basaran, Taejoon Park, and Sang Son. 2014. Energy-efficient privacy protection for smart home environments using behavioral semantics. *Sensors* 14, 9 (2014), 16235–16257. DOI: https://doi.org/10.3390/s140916235

[103] Charith Perera, Susan Y. L. Wakenshaw, Tim Baarslag, Hamed Haddadi, Arosha K Bandara, Richard Mortier, Andy Crabtree, Irene C. L. Ng, Derek McAuley, and Jon Crowcroft. 2016. Valorising the IoT databox: Creating value for everyone. *Trans. Emerg. Telecommun. Technol.* 28, 1 (2016), e3125.

[104] Aimee Picchi. 2019. Amazon Workers are Listening to What You Tell Alexa. Retrieved from https://www.cbsnews.com/news/amazon-workers-are-listening-to-what-you-tell-alexa/.

[105] I. Pinyol and J. Sabater-Mir. 2013. Computational trust and reputation models for open multi-agent systems: A review. *Artific. Intell. Rev.* 40, 1 (2013), 1–25.

[106] Marco Prandini and Marco Ramilli. 2010. Towards a practical and effective security testing methodology. In *Proceedings of the IEEE Symposium on Computers and Communications (ISCC'10)*. IEEE, 320–325.

[107] Priyanka Chouhan and Rajendra Singh. 2016. Security attacks on cloud computing with possible solution. *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)* 6, 1 (Jan. 2016), 92–96.

[108] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHIEA'17)*. ACM, New York, NY, 2853–2859. DOI: https://doi.org/10.1145/3027063.3053246

[109] Kopo M. Ramokapane, Awais Rashid, and Jose M. Such. 2016. Assured deletion in the cloud: Requirements, challenges and future directions. In *Proceedings of the ACM Cloud Computing Security Workshop*. 97–108.

[110] Kopo M. Ramokapane, Awais Rashid, and Jose M. Such. 2017. "I feel stupid I can't delete...": A study of users' cloud deletion practices and coping strategies. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 241–256.

[111] Toni Reid. 2018. Everything Alexa Learned in 2018. Retrieved from https://blog.aboutamazon.com/devices/everything-alexa-learned-in-2018.

[112] Mary Lou Roberts. 2019. Are Your Voice Assistants Always Listening? The Simplistic Answer is "Yes." Retrieved from http://www.capecodtoday.com/article/2019/08/11/248280-Are-Your-Voice-Assistants-Always-Listening.

[113] Mike Rodehorst. 2019. Why Alexa Won't Wake Up When She Hears Her Name in Amazon's Super Bowl Ad. Retrieved from http://web.archive.org/web/20190211063816/https://developer.amazon.com/blogs/alexa/post/37857f29-dd82-4cf4-9ebd-6ebe632f74d3/why-alexa-won-t-wake-up-when-she-hears-her-name-in-amazon-s-super-bowl-ad.

[114] Rodrigo Roman, Javier Lopez, and Stefanos Gritzalis. 2018. Evolution and trends in the security of the Internet of Things. *IEEE Comput.* 51 (July 2018), 16–25. DOI: https://doi.org/10.1109/MC.2018.3011051

[115] Rodrigo Roman, Ruben Rios, Jose A. Onieva, and Javier Lopez. 2019. Immune system for the Internet of Things using edge technologies. *IEEE Internet Things J.* 6, 3 (June 2019), 4774–4781. DOI: https://doi.org/10.1109/JIOT.2018.2867613

[116] E. Ronen, A. Shamir, A. Weingarten, and C. O Flynn. 2018. IoT goes nuclear: Creating a zigbee chain reaction. *IEEE Secur. Priv.* 16, 1 (Jan. 2018), 54 to 62. DOI: https://doi.org/10.1109/MSP.2018.1331033

[117] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible voice commands: The long-range attack and defense. In *Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI'18)*. USENIX Association, Renton, WA, 547–560.

[118] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proc. ACM Interact. Mobile, Wear. Ubiq. Technol.* 1, 4 (2018), 1–23. DOI: https://doi.org/10.1145/3161187

[119] F. Schaub, R. Balebako, and L. F. Cranor. 2017. Designing effective privacy notices and controls. *IEEE Internet Comput.* 21, 3 (2017), 70–77.

[120] Lea Schonherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2018. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. Retrieved from http://arxiv.org/abs/1808.05665.

[121] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. 3–18.

[122] Micah Singleton. 2017. Alexa Can Now Set Reminders for You. Retrieved from https://www.theverge.com/circuitbreaker/2017/6/1/15724474/alexa-echo-amazon-reminders-named-timers.

[123] D. J. Solove. 2006. A taxonomy of privacy. *U. Penn. Law Rev.* 154, 3 (2006), 477–560.

[124] SRLabs. 2019. Smart Spies: Alexa and Google Home Expose Users to Vishing and Eavesdropping. Retrieved from https://srlabs.de/bites/smart-spies/.

[125] Statista. 2018. Worldwide Intelligent/Digital Assistant Market Share in 2017 and 2020, by Product. Retrieved from https://www.statista.com/statistics/789633/worldwide-digital-assistant-market-share/.

[126] Nick Statt. 2019. Google Defends Letting Human Workers Listen to Assistant Voice Conversations. Retrieved from https://www.theverge.com/2019/7/11/20691021/google-assistant-ai-training-controversy-human-workers-listening-privacy.

[127] Guillermo Suarez-Tangil, Juan E. Tapiador, Pedro Peris-Lopez, and Arturo Ribagorda. 2014. Evolution, detection and analysis of malware in smart devices. *IEEE Commun. Surveys Tutor.* 16, 2 (2014), 961–987.

[128] Jose M. Such. 2017. Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, 4761–4767.

[129] Jose M. Such, Pierre Ciholas, Awais Rashid, John Vidler, and Timothy Seabrook. 2019. Basic cyber hygiene: Does it work? *Computer* 52, 4 (2019), 21–31.

[130] J. M. Such and N. Criado. 2016. Resolving multi-party privacy conflicts in social media. *IEEE Trans. Knowl. Data Eng.* 28, 7 (2016), 1851–1863.

[131] J. M. Such and N. Criado. 2018. Multiparty privacy in social media. *Commun. ACM* 61, 8 (2018), 74–81.

[132] Jose M. Such, Natalia Criado, Laurent Vercouter, and Martin Rehak. 2016. Intelligent cybersecurity agents. *IEEE Intell. Syst.* 31, 5 (2016), 3–7.

[133] Jose M. Such, Antonios Gouglidis, William Knowles, Misra Gaurav, and Rashid Awais. 2016. Information assurance techniques: Perceived cost effectiveness. *Comput. Secur.* 60 (2016), 117–133. DOI: https://doi.org/10.1016/j.cose.2016.03.009

[134] J. M. Such and M. Rovatsos. 2016. Privacy policy negotiation in social media. *ACM Trans. Auton. Adapt. Syst.* 11, 1 (2016), 4.

[135] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light commands: Laser-based audio injection attacks on voice-controllable systems. In *29th USENIX Security Symposium (USENIX Security'20)*. USENIX Association, 2631–2648. https://www.usenix.org/system/files/sec20-sugawara.pdf.

[136] Paul Syverson. 2009. Why I'm not an Entropist. In *Proceedings of the 17th International Security Protocols Workshop*. Retrieved from https://www.freehaven.net/anonbib/cache/entropist.pdf.

[137] Paul Syverson. 2011. Sleeping dogs lie on a bed of onions but wake when mixed. In *Proceedings of the HotPETS Conference*. Retrieved from https://petsymposium.org/2011/papers/hotpets11-final10Syverson.pdf.

[138] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199.

[139] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating users' preferences and expectations for always-listening voice assistants. *Proc. ACM Interact. Mobile Wear. Ubiq. Technol.* 3, 4 (2019), 1–23.

[140] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. 2019. "I don't own the data": End user perceptions of smart home device data practices and risks. In *Proceedings of the 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, Santa Clara, CA. Retrieved from https://www.usenix.org/conference/soups2019/presentation/tabassum.

[141] Massimiliano Todisco, Hctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients. *Comput. Speech Lang.* 45, C (Sep. 2017), 516–535. DOI: https://doi.org/10.1016/j.csl.2017.01.001

[142] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine noodles: Exploiting the gap between human and machine speech recognition. In *Proceedings of the 9th USENIX Workshop on Offensive Technologies (WOOT'15)*. USENIX Association, Washington, D.C.

[143] Xueou Wang, Xiaolu Hou, Ruben Rios, Per Hallgren, Nils Ole Tippenhauer, and Martin Ochoa. 2018. Location proximity attacks against mobile targets. In *Proceedings of the 23rd European Symposium on Research in Computer Security (ESORICS'18) (LNCS)*, Vol. 11099. Springer, Barcelona, 373–392. DOI: https://doi.org/10.1007/978-3-319-98989-1

[144] Human Right Watch. 2017. China: Voice biometric collection threatens privacy. Retrieved from https://www.hrw.org/news/2017/10/22/china-voice-biometric-collection-threatens-privacy.

[145] Ryen W. White. 2018. Skill discovery in virtual assistants. *Commun. ACM* 61, 11 (Oct. 2018), 106–113. DOI: https://doi.org/10.1145/3185336

[146] Sam Wolfson. 2018. Amazon's Alexa Recorded Private Conversation and Sent it to Random Contact. Retrieved from www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation.

[147] Venessa Wong. 2017. Burger King's New Ad Will Hijack Your Google Home. Retrieved from https://www.cnbc.com/2017/04/12/burger-kings-new-ad-will-hijack-your-google-home.html.

[148] Charles Wright, Scott Coull, and Fabian Monrose. 2009. Traffic morphing: An efficient defense against statistical traffic analysis. In *Proceedings of the Network and Distributed Security Symposium*. IEEE.

[149] David Wright and Paul De Hert. 2012. Introduction to privacy impact assessment. In *Privacy Impact Assessment*. Springer, 3–32.

[150] Lei Xinyu, Tu Guan Hua, Alex X. Liu, Li Chi Yu, and Tian Xie. 2017. The insecurity of home digital voice assistants: Amazon Alexa as a case study. Retrieved from https://arxiv.org/pdf/1712.03327.pdf.

[151] Jun Yang. 2018. Multilayer adaptation based complex echo cancellation and voice enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. DOI: https://doi.org/10.1109/icassp.2018.8461354

[152] Eric Zeng, Shrirang Mare, and Franziska Roesner. 2017. End user security and privacy concerns with smart homes. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security (SOUPS'17)*. USENIX Association, USA, 65–80.

[153] Eric Zeng and Franziska Roesner. 2019. Understanding and improving security and privacy in multi-user smart homes: A design exploration and in-home user study. In *Proceedings of the 28th USENIX Security Symposium (USENIXSecurity'19)*.

[154] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS'17)*. DOI: https://doi.org/10.1145/3133956.3134052

[155] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian. 2019. Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. 1381–1396.

[156] Sendong Zhao, Wu Yang, Ding Wang, and Wenzhen Qiu. 2012. A new scheme with secure cookie against SSLStrip attack. In *Web Information Systems and Mining*, Fu Lee Wang, Jingsheng Lei, Zhiguo Gong, and Xiangfeng Luo (Eds.). Springer, Berlin, 214–221.

[157] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User perceptions of smart home IoT privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 200 (Nov. 2018), 20 pages. DOI: https://doi.org/10.1145/3274469

[158] Andras Zolnay, Daniil Kocharov, Ralf Schluter, and Hermann Ney. 2007. Using multiple acoustic feature sets for speech recognition. *Speech Commun.* 49, 6 (2007), 514–525. DOI: https://doi.org/10.1016/j.specom.2007.04.005

Received March 2019; revised April 2020; accepted July 2020