Shouting Through Letterboxes: A study on attack susceptibility of voice assistants

Andrew McCarthy
Computer Science Research Centre
University of the West of England
Bristol, United Kingdom
Andrew6.McCarthy@uwe.ac.uk

Benedict R. Gaster
Computer Science Research Centre
University of the West of England
Bristol, United Kingdom
Benedict.Gaster@uwe.ac.uk

Phil Legg
Computer Science Research Centre
University of the West of England
Bristol, United Kingdom
Phil.Legg@uwe.ac.uk

***Abstract*—Voice assistants such as Amazon Echo and Google Home have become increasingly popular for many home users, for home automation, entertainment, and convenience. These devices process speech commands from a user to execute some action, such as playing music, making online purchases, or triggering home automation such as lights or security locks. The process of mapping speech input to a text command is performed using a machine learning model. In this study, we explore the concept of how voice assistants could be exploited, where genuine audio commands are manipulated such that an attacker could trigger alternative responses from the voice assistant. We present a small-scale study to examine mis-interpretations made by voice assistants. We also study user perception of how secure their voice devices are, and their approach to security and privacy.**

I. Introduction

The rise in popularity of smart home voice assistants such as Amazon Alexa and Google Home bring both new functionality and convenience for home users, along with new attack vectors and security risks. The use of such devices has caused debate [1] due to the properties associated with an 'always-on' microphone, that can potentially introduce privacy risks in the home. Smart home assistants introduce a wealth of functionality designed for user convenience and to improve seamless experience and interaction. Typical tasks may range from playing music and calling friends, through to home automation (e.g., heating a kettle, unlocking a door), location tracking of family members, managing finances, or making online purchases. Such voice interactions could well open up further potential security risks [2] [3]. Whilst voice interactions may introduce great convenience, it is naturally an insecure medium due to the need to speak aloud. Imitation attacks could be conducted to mimic an authorised user, and replay attacks could instigate previously-executed commands. Most home assistants do not require authentication and so any nearby user can easily execute a command, which could include a form of adversary. More recently, high-frequency audio attacks (referred to as a 'DolphinAttack') have shown how a command could be executed without the command being audible to a human target [4]. Similarly, there may be further properties of the voice assistant that can be exploited by an attacker, such as how machine learning techniques may be used for performing speech-to-text translation or speaker recognition [5].

In this paper, we explore the potential security risks that can be introduced through the usage of smart voice assistants. In particular, we focus on how voice assistants may be triggered to execute commands under covert means, for example, where a careful-crafted audio sample may seem inaudible or nonsensical as far as a human is concerned, but that may be recognised and executed as some form of command by the device. Our study consists of two parts. Firstly, we conduct a short user study to gauge user opinions on the security and safety of voice assistants. Secondly, we conduct a small practical study to demonstrate mis-classification of audio to trigger alternative actions as a means of covert behaviour. Our work contributes towards this relatively new research area by understanding user perceptions for home voice assistant security, and by providing a proof-of-concept that shows how such systems could potentially be exploited by an adversary.

The paper is structured as follows: Section II provides background to the subject, including security and privacy concerns associated with voice assistants. Section III provides details of our two research studies. Section IV shows the findings from our questionnaire regarding user perceptions of security in voice assistants. Section V describes attack vectors for introducing compromise in a text-to-speech system. Section VI presents the results for our proof-of-concept demonstration for attack injection. Section VII provides a discussion of our findings, and future directions. Finally, Section VIII concludes our study.

II. Background

In this section we introduce intelligent voice assistants and detail potential vulnerabilities, attack methods, and privacy concerns. Danaher [6] provides a generous definition of an intelligent personal assistant as *"any computer-coded software system/program that can act in a goal-directed manner"*. Turing's [7] seminal work in AI gave great importance to human-like quality of responses, however common *faux-pas* that occur in responses from systems such as Alexa and Siri often reveal the current limitations of the machine. We define Intelligent Voice Assistants (IVAs) as: "computers capable of interacting with users through voice, performing goal-based tasks or services on behalf of users". Current IVAs implement a two-stage model [8]. First is the activation stage where the

system will continually listen for a specific *wake-word*, such as "OK Google" or "Alexa". Following this, stage two listens for the spoken command, which is then converted into a text-based representation for processing. Where appropriate, an audible response is returned for the user, either to answer the query or to confirm some desired action.

A. Security of Voice Assistants

Modern speech recognition systems rely on common machine learning approaches such as neural networks, however their use can be compromised through adversarial examples [9]. Attacks against machine learning systems may include mis-classifications that are designed to compromise normal system operations, and query-based attacks to reveal confidential information about the model and/or its users.

Given a classification task, a system may learn a suitable decision boundary for classifying two or more instances (e.g., classifying different words). An adversary may compromise performance by subtle manipulation of input features, such that a sample appears to mis-classify. Adversarial attacks have been shown on facial recognition [10], road sign recognition [11], and network intrustion detection [12]. Biggio and Roli [13] declare that machine learning is not a cyber security panacea, due to such distinct exploitable vulnerabilities that could potentially impact on larger connected systems.

Users that choose to adopt such smart home devices implicitly accept the increased level of risk to their home security and privacy. It is for users to decide on how they balance the trade-off between convenience and operation, with the issue of security and privacy risks [14]. However, voice as a means of interaction is becoming more common for causal users, replacing or augmenting command-lines and keyboards used for textual input [15]. Alepis and Patsakis go further by arguing that voice assistants are replacing traditional user interfaces, changing how we access Applications, data and the internet [16], and therefore, they require a broad set of permissions and privileges. Moreover, often they belong to the Operating System (OS) or device manufacturers, and generally have relatively elevated privileges. Thus, voice assistants offer rich rewards if they are exploited, and are hence an attractive target for adversaries. As adoption increases, the threat posed by voice assistants is one that could potentially be more dangerous than is generally recognised by many users.

B. Adversarial Attacks on Voice interfaces of voice assistants

Advances in Text-To-Speech (TTS) research mean impersonation of speakers is easier than ever. For example, VoCo software system [17] enables an editor to insert or replace words in a recording, distorting its original meaning. Results are worryingly accurate and simple to produce.

In August 2019 it was reported that a British energy firm fell victim to a voice mimic fraud with around £200,000 paid to fraudsters [18]. Insurance firm Euler Hermes disclosing the fraud confirmed that AI software accurately mimicked the executive's accent and style of speaking [18].

The situation is even more alarming for voice assistants that perform little authentication. An adversary need not wait for an authorized user to utter coveted phrases, to perform a replay attack. Nor spend the moments required to produce a convincing replica of a user's command using VoCo. An attacker only needs speak commands themselves! Moreover, many voice assistants respond to synthesised voices, adversaries can direct computers to vocalize a synthesized command using TTS technologies, our experiments show voice assistants respond to TTS generated commands.

A wide variety of voice-controlled home Internet of Things (IoT) devices are available, ranging from smart kettles, door locks, security alarms and heating systems. Through controlling more devices via voice assistants the systems inevitably gain complexity and the attack surface increases, presenting greater risk. Nascent use of voice assistants for triggering multiple high wattage home IoT devices could impact on power distribution grids [19], causing power outages that could significantly harm critical sectors such as transport and health.

Rajaratnam and Kalita [20] discuss how noise flooding, a form of availability attack, could be used to detect adversarial attacks on speech recognition. Further availability attacks include electromagnetic attacks that are commonly used for radio jamming or denial of service attacks; however, using intentional electromagnetic interference (IEMI) against mobile phone voice assistants is possible. Headphones connected to a mobile phone act as efficient antennae for IEMI by which adversaries can inject voice commands actioned by voice assistants [21].

C. Hidden Voice

Recent research shows viable attacks to voice-controlled systems with hidden commands unnoticed or unintelligible to humans but recognized by voice assistants [22]. The severity of such attacks largely depends on the commands a target device accepts. Attacks could lead to information leakage, unavailability of service, or act as a foothold toward further attacks. The reach of an attack could be increased by embedding into trending social media videos, or broadcasting through a loud speaker at popular events.

The adversarial attack *CommanderSong* [23] inserts voice commands into music videos or audio files, with minor perturbations, resulting in normal sounding audio; however, voice assistants recognize embedded commands and action them. Likewise, the *DolphinAttack* [4] has been shown to compromise voice assistants by modulating voice command into ultrasonic frequencies rendering them inaudible to humans.

D. Privacy Concerns

Endersley defined three levels of situational awareness: perception, comprehension, projection [24]. Those unable to perceive risks are unconcerned because they are unaware [25]. Further, users vary in risk appetite, perhaps accepting privacy and security risks in return for personalised services or convenience. The introduction of the General Data Protection Regulation (GDPR) has seen privacy become a recognised concern

by users in terms of how their data is used. Increasing numbers of voice assistants present on-going security concerns, and privacy laws are ill-suited to the task. Technical changes to help bolster privacy protections for users are suggested by [26]: 1) Shorten the length of time recordings are retained; 2) Clearly notify users how their recordings are used; 3) Devices should have clear visual indicators showing collection and transmission of data; 4) Companies should make remote conversion of devices technically impossible; and 5) As much processing as possible should be done on the device and not in the cloud.

Companies such as Google and Amazon deal in personal information. Technology companies take rights to define what is private and use "digital exhaust" or behavioural surplus of users' interactions. Our "digital exhaust" is information leaked through our actions. For example, places we visit, purchases we make, likes we give on Facebook are used to profile us. Companies use this to predict future behaviour, selling these predictions to advertisers. Their investment in voice assistants is likely partly based on the extraction of behavioral surplus from human experience and its predictive value realized in future-behaviour markets [27].

People often lack awareness of information they disclose, potentially disclosing in error [28]. Further, actions often discord with stated preferences; moreover people often reveal more information when 'chatting' about themselves. [29]; We posit users are less security and privacy conscious around personal voice assistants, opening further risks.

The NHS and Amazon have reportedly joined forces, allowing Alexa users to access reliable NHS health information through their Alexa device [30]. This is trumpeted as a boon for visually impaired users; however, provides more encouragement to disclose our private health information to a huge technology corporation, who ultimately advertise health products to us. This is enormously unsettling. Users must control their personal information, and preserve their privacy. Thus, retaining a modicum of independence from behemoth technology companies.

Constantly listening devices raise possibilities of information leakage from which personal or contextual information could be gleaned. For example, household sounds such as laughter, children, or activities may imply emotions, mood, or other information about a user's life.

Voice assistant use on phones opens avenues for extracting private data, even from locked devices. Forensic data retrieval from mobile phones is normally a complex task; however, voice commands eliciting a response from Siri enable adversaries and police forensics alike to recover call logs, SMS, Contacts, Maps, Calendar and device information [31].

III. Methodology

Our study consists of two main contributions. Firstly, we conduct a survey studying user attitudes towards security issues of voice assistants using a structured questionnaire. Our survey of user behaviour and habits provides insight into voice assistant use in homes. A selection of questions were asked, collecting and collating the results, in order to identify common themes and trends in the use of voice assistants, investigating whether user's security practices for voice assistants are laxer compared to general security practices. Topics of interest are: what room is the assistant placed in, how people use their voice assistants, awareness and use of voice assistant security features , do they mute the microphone, number and type of other proximate devices.

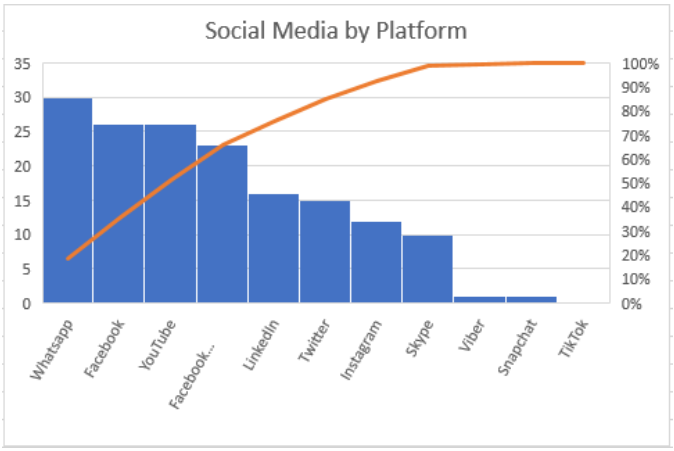

Fig. 1. Pareto chart to show the most used Social Media Platforms

Following the initial survey, we present a proof-of-concept demonstration on how audio samples can be deliberately manipulated to trigger an IVA to execute some other command to that heard by humans A selection of exploratory experiments were conducted determining the ease and seriousness of attacks. The researcher's voice and synthesized TTS voices were used to produce recordings of speech commands. Further experiments were conducted playing audio from a laptop computer in close proximity to the voice assistant, observing whether the command was executed. Effects of adversarial noise are explored using dense white noise, determining how an adversary can affect speech recognition.

IV. Survey: Online Questionnaire

Through our survey, we hope to obtain insight into how and why IVAs are used by consumers. We also aim to investigate the resulting trade-off between security and usability. We believe that the personal nature of voice assistants lowers user's circumspection. Our survey aims to test the following null (H_0) and alternative (H_1) hypotheses statements:

- H_0**:** Users adopt the **same** security posture when using voice assistants as they do when engaging with other technology and online communications.
- H_1**:** Users adopt a **weaker** security posture when using voice assistants as they do when engaging with other technology and online communications.

A. Data Collection

An anonymized online survey in the form of a questionnaire was designed and conducted to gather qualitative responses

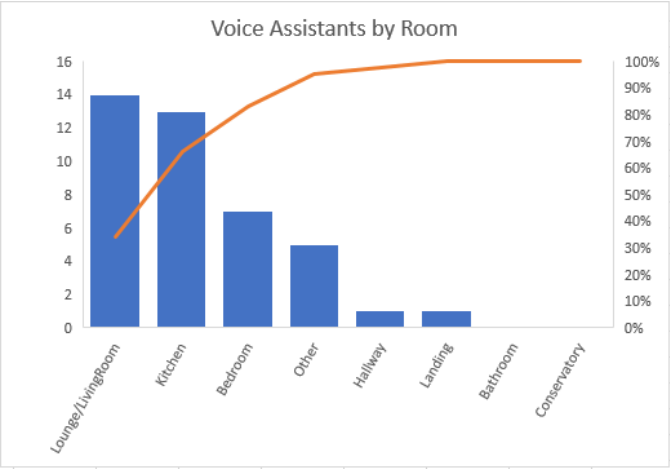

Fig. 2. Pareto chart to show the most popular locations of voice assistants

from participants. Particular focus was given to topics of behaviour, attitudes to security, and privacy. As part of our survey, we describe related scenarios and ask participants how concerned they would be, using a recognised approach for measuring privacy [32]. The full details of the questionnaire are available in [33]. Fourty adult participants contributed towards the presented survey on a voluntary basis. Invitations were sent out to the wider academic community and promoted through our personal networks.

B. Survey Results

The results of our survey are analysed in four themes: Personal traits, Voice assistant ownership, General Security, and Voice Assistant Usage.

1) Personal Traits: The majority of participants (70.28%) report themselves as early adopters of technology. Unsurprisingly most participants tend to report security is very important. Likewise, with privacy; however, seven participants reported having a voice assistant in a bedroom, seemingly contradictory. The majority of participants (62.16%) reported being very or extremely computer literate. Factors influencing this might be: a limited sample of participants over fifty-five years old, older generations may be less familiar and comfortable with computers. We distributed invitations to students on the M.Sc. Cyber Security, possibly skewing results with more technically capable people. The researcher's friends and family may have been influenced by knowing the researcher, or have gained knowledge over time from the researcher. Thus, through reporting bias under reporting of perceived bad behaviours is possible. The majority of participants (72.97%) report they are likely to use smart devices for home automation; contrasting with the finding that controlling the home in this way was not a popular aspiration. Most participants report not connecting things to their voice assistant despite, companies assertively marketing voice-controlled home IoT devices.

Moreover, participants (75.68%) tended to strongly agree they are always online, with most spending between 1-3 hours on social media daily, the most popular social media platforms were Whatsapp, Facebook, and YouTube as in Figure 1. Thus suggesting videos or audio files containing voice commands could be a viable attack vector when shared on social media.

2) Ownership: The majority (72.97%) of participants own a voice assistant. Potentially reflecting a sample bias, those choosing to complete the survey are likely more interested. Interestingly, voice assistants are becoming ubiquitous, integrated into many devices. Consequently, some participants who claim they do not have a voice assistant do, regardless whether the functionality is enabled, or used.

Figure 2 shows that voice assistants are commonly placed in the Lounge/Livingroom, Kitchen, and Bedrooms (traditionally private spaces). Some report voice assistants in hallways, bolstering thoughts that voice attacks could be executed through letterboxes. Five participants chose 'other', but could not specify where. Thus 'other' could refer to another room, for example a garden shed; Although some participants may recognize that the voice assistant on their mobile phone is not tied to one room and instead is mobile. Further research into users' views of mobile voice assistants might yield interesting results. For example, are voice assistants perceived as present only when being interacted with?

All participants confirmed at least one other device is in the room with a voice assistant. This adds weight to chaining voice attacks by playing audio through nearby devices. Laptops, televisions, and telephone/answering machines are particularly interesting attack vectors.

The vast majority (82.05%) of participants reported keeping their phone in their bedroom at night; significant for privacy assuming a voice assistant is enabled on a phone.

3) General Security: Most participants (83.79%) report security as very or extremely important; however, security is subjective; most participants claimed they were likely to use a password manager (56.76%), likely to reuse passwords (56.76%), and likely to change default passwords (89.19%); This seems at odds with how participants use voice assistants: most participants do not mute the microphone when they are not using it (58.06%), and the majority have not trained their voice assistant to uniquely recognize their voice (55.18%).

4) Usage: Most participants find using their voice assistants relatively enjoyable, listening to music is the most popular activity.

Participants tended to disagree with the statement "I have trained the voice assistant to only recognize my voice". Corroborating our view that voice assistants generally accept commands from unauthorized users, and strengthens arguments for user authentication. Participants tend to somewhat agree they are concerned about having deeply private conversations in front of a voice assistant, with 65.71% of participants somewhat agree, agree, or strongly agree. However, most participants do not know how to erase recordings made by a voice assistant and do not mute their voice assistant when not in use. People report being concerned about privacy, but are ignorant of features to take control of their privacy. Privacy controls are not widely reported by technology companies, and

often difficult to find, which could perhaps align with corporate interests.

Considering security and privacy across the domains: personal, family, company, nation. Participants were least concerned about company security and privacy, with national security and privacy third in people's minds. Participants may not consider privacy and security in their job function, perhaps they consider national security and privacy is entrusted to the security services. Our literature review uncovered potential attacks on national infrastructure, facilitated through smaller local attacks. Therefore, awareness at all levels is required. Participants report less concern with security; although implications of a successful attack could include information theft, financial loss or physical harm. Privacy eclipses security in every category except the national level, where security and privacy are rated equal. Privacy is probably forefront in people's minds because of always on microphones. Most participants (88.88%) report they are unlikely to conduct banking through their voice assistant; whereas they are more open to shopping through a voice assistant. Participants agree (63.88%) they are interested in home automation.

C. Summary

Our two hypotheses statements were:

- H_0**:** Users adopt the **same** security posture when using voice assistants as they do when engaging with other technology and online communications.
- H_1**:** Users adopt a **weaker** security posture when using voice assistants as they do when engaging with other technology and online communications.

Our survey results show general security is important to people, and with the exception of reusing passwords, people take sensible precautions. This is at odds with participants use of voice assistants where most have not trained assistants to uniquely identify their voice, and do not mute voice assistants. We conclude that hypothesis H_1 gains credibility from the survey results. Thus we posit that people are less concerned by security and privacy risks of voice assistants due to the personal nature of these devices. Survey and results have been made available for the wider research community in our repository [33].

V. Empirical Study

Voice is difficult to secure as all sound travels through air, effectively an open channel. Voice is vulnerable to eavesdropping, where an adversary can learn information through listening to exchanges between the user and the voice assistant. This section explores a range of adversarial attacks conducted against voice assistants in the researcher's home.

A. Attacking a Voice Assistant

To better understand possible attack scenarios, a selection of experimental approaches were taken to ascertain vulnerabilities. Experiments utilised either the Python speech recognition module [34] or an Amazon Echo device [35], justifiably chosen due to its popularity and convenience. Our survey indicates that Amazon Alexa is the most populace voice assistant.

B. Shouting Through Letterboxes

The Amazon Echo relies on single factor authentication. There are four wake words: "Alexa, Amazon, Computer, and Echo". In order to be certain of the correct wake word an adversary need only cycle through them. Alexa is controllable by anyone proximate to the device. An Amazon Alexa was placed in the hallway three metres from the door. In order to determine how vulnerable a voice assistant might be to adversaries outside the home, a voice command was spoken through the letterbox: "Alexa turn the light on". Alexa was heard to say "OK", and the light was switched on, indicating a successful attack. The potential for adversaries to abuse IVAs in close proximity to a letterbox or open window is clear. Should a smart door lock be controllable by the voice assistant, a burglar could potentially just ask Alexa to unlock the door!

C. Replay Attack

We conducted a simple replay attack, the researcher recorded his voice saying "Alexa, shuffle my music" using a Recorder application on a Huawei mobile phone running android. The recording successfully triggers the voice assistant when played.

D. Answerphone

We conducted a test to determine susceptibility of attack through telephone answering machines. An Amazon Echo is placed in the hallway near the land line telephone/answering machine. A mobile telephone was taken outside the residence and a telephone call initiated with the land-line inside. A message was left saying "Alexa, shuffle my music". On returning a few moments later the Amazon Echo was playing music. Further, playing the answerphone message triggers the voice assistant. This successful attack suggests that voice assistants placed near answering machines may be susceptible to remote command. Only a telephone and knowledge of the telephone number near the voice assistant is required. One could conceive of speculative, wide spread attacks where an adversary uses an automatic dialler to play pre-recorded messages in order to control voice assistants within range.

E. TV, Laptop, Radio, and other devices

Testing the hypothesis that voice assistants can be triggered by nearby devices some recordings were played from a laptop computer near an Amazon Echo, approximately two metres apart. Audio from a radio programme podcast [36] was played, successfully triggering the Amazon Echo. Further, video of a South Park episode containing voice commands [37] was played, successfully triggering the Amazon Echo.

F. Camouflaged attacks

Attacks so far identified are easily detected by anyone nearby. On hearing commands, it is evident that someone is trying to control the voice assistant. Adversaries would benefit from camouflaging their true intentions. This is possible

by hiding commands in a larger context. We envisage an audiobook could contain the admittedly banal prose: “Bertie watched the cars on the main road. The traffic lights turned red. A Lexus stopped.”. This appears part of a story; however, causes an Amazon Echo to close its current interaction or dismiss a notification, as observed. Our experiment played the audio through a laptop speaker, successfully causing Alexa to stop playing music. These attacks are harder to detect; although attentive listeners may detect something untoward.

G. Adding Adversarial noise

In this section we describe our work to attack a voice assistant through adversarial noise. Different colours of noise, refer to the power spectrum of a noise signal created by a stochastic, or random process.

We chose to use white noise in our experiment. White noise is less disruptive to voice, and suited our experiments. Pink noise has better speech masking properties than white noise, and is more effective than white noise when blocking or jamming voice assistants with an availability attack. Attacks on voice assistants are possible by adding adversarial noise to the voice assistant’s input, causing the TTS transcription to differ to what has been heard.

A large corpus of TTS audio files was gathered. A text file containing a list of words and phrases was passed through a TTS tool [38]. This experiment introduced adversarial noise to recordings to determine how easily noise could be used to confuse speech recognition. Results were collected by adding noise to a wave file before passing it to the speech recognition API.

A Python script using the SpeechRecognition module [34], was written, taking a word list passing each word to a TTS engine. The resulting audio file is then passed to the Google speech recognition API gaining a text transcription of the initial audio. Subsequently dense random noise is added, manipulating each sample in the audio. The result is passed again to the Google speech recognition API. Resulting text transcriptions were compared indicating the effect of added noise. Initial experiments gave encouraging anecdotal results. We successfully deceived the Google speech recognition API by adding noise, resulting in different transcriptions. Thus, a proof of concept was achieved: it is possible to affect audio transcription by adding noise. Speech recognition of single words spoken by our TTS system was, however, unreliable often either transcribed incorrectly, or not at all. Hence, we adopted human speech in our later experiments.

Using a list of one hundred words [39] human voice samples for each of one hundred common English words were recorded using a laptop microphone in a quiet room. Once recorded audio files were passed through the Google speech recognition API. Noise was added to them, and they were re-passed through the API, as indicated in Figure 3. Transcriptions of clean and noisy examples were initially compared. This indicates Google speech recognition system can be influenced through adversarial noise. Adding white noise to a recording of the word “people” regularly, resulted in a file that whilst

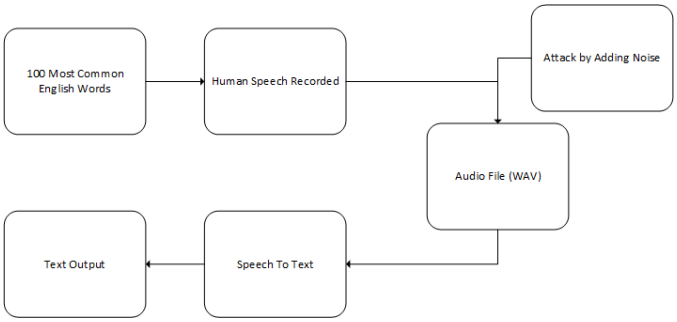

Fig. 3. Adversarial noise applied to recorded human speech

still intelligible to a human is not recognized by the speech recognition system. Adding white noise to an audio file of the word “those” reliably caused resulting files to be recognized as “though” by the speech recognition system. We also hear the word “though”. The noise obliterated the phoneme responsible for the “s” sound.

Given the large input vector that makes up an audio sample, it is difficult to assess the overall variability between samples when influential parameters that could change the classification are changed. To examine this further, we use Principle Component Analysis (PCA) to perform dimensionality reduction, allowing for a more intuitive approach to compare samples, and changes to those samples, in a 2-dimensional space, as in Figure 4. It can be seen that adding adversarial noise to audio samples shifts their position in a 2-dimensional space, closer to a decision boundary.

H. Future Work

We show a misclassification not compromising normal operation of a voice assistant. Next steps would cause a misclassification compromising system behaviour. For example, to execute a specific command, triggered through an adversarial example.

Relevant commands could be used with sophisticated adversarial noise applied to shift samples over a decision boundary. Inverse PCA could be utilised to calculate requisite input for a given position. Further work would focus on generating sophisticated adversarial examples. For example, implementing sparse perturbations, such that few samples in the audio are modified. A genetic algorithm could be devised to optimize the perturbations, evaluating examples against a fitness function, until a suitable adversarial example is found.

VI. Results

Our results show IVAs are potentially vulnerable to attacks by unauthorized users and adversaries. Voice assistants have been shown to be triggered by nearby devices. The effect of adding adversarial noise is shown and plotted using PCA, indicating noise can shift audio files closer to a decision boundary. Through experimentation, we discovered voice assistants are vulnerable to simplistic attacks where adversaries can control a voice assistant. Sound travels through air, an open channel. Thus, voice is difficult to secure.

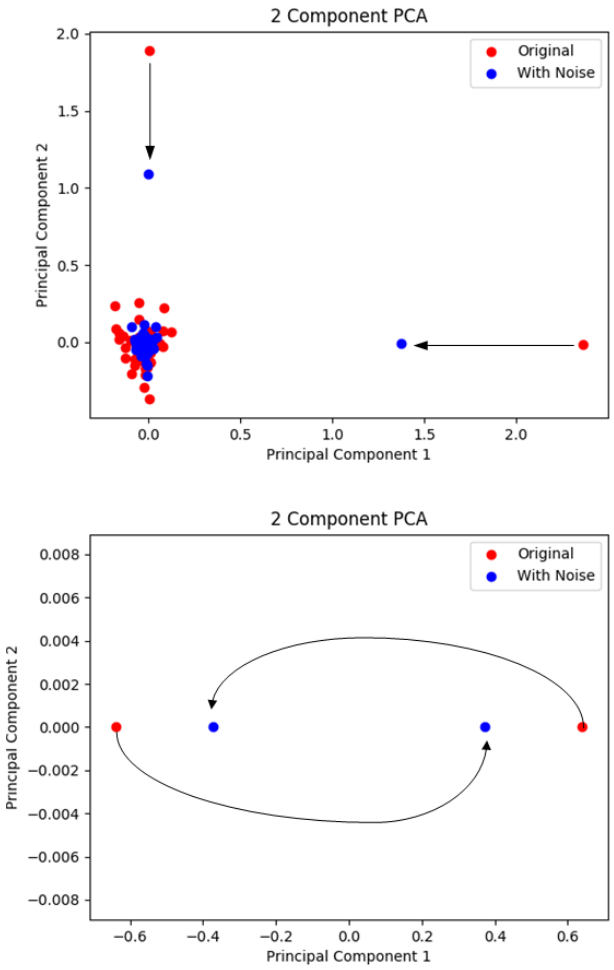

Fig. 4. (a) Principal Component Analysis of two hundred audio files (b) Principal Component Analysis of four audio files relating to the words: Think and Thing

Adversaries can easily wake and command voice assistants, commands can be given by shouting through a letterbox! Use of technology increases distance between adversaries and target voice assistants. We show commands can be issued through answerphone or telephone screening devices. Ordinary broadcast media like television and radio programmes can command voice assistants. social media use and video sharing sites like YouTube and Facebook could open possibilities of large-scale indiscriminate attacks world-wide. Moreover, we show some simple commands can be obfuscated or hidden. Thus, human victims may remain unaware that their voice assistant was attacked. In particular sound files affected by adversarial noise can appear normal; however, be shifted closer to a decision boundary. Thus, classified differently.

VII. Discussion & Analysis

Voice is an insecure user interface. Unauthorized users near voice assistants can issue commands to turn lights on, play music, control attached devices, purchase items, or add items to shopping lists. In general, anyone near a voice assistant can issue any commands. Replay attacks are straightforward though often unnecessary. Moreover, imitation attacks can be performed mimicking authorized voices. Nearby devices can be commandeered to play voice commands to voice assistants. Experiments showed commands played from a laptop computer successfully trigger the voice assistant. The presence of other devices such as a radio, television or laptop computer nearby increases the attack surface available to adversaries wishing mischief or malice.

Evidently adversaries could control voice assistants by gaining access to other home devices. There is no guarantee that they will be content with mischief and likely some are intent on theft or malice. Radio and television broadcasts similarly trigger voice assistants, as can YouTube videos. Extrapolating on our experiments, malware could be written to infect laptops and issue voice commands Broadcast media could unduly influence voice assistants as seen when fast food restaurant Burger King exploited viewers Google Home voice assistants using an advert appearing on YouTube and later appearing on American television.

Further mischief was had by writers of satirical animated sitcom South Park who poked fun at voice assistant users. Characters play with an Amazon Echo and add some peculiar items to their shopping list; however, voice assistants within proximity of the television worryingly obey the commands, adding items to shopping lists [37].

Consumer sound recognition systems like Shazam have been around since the late nineties. Sound recognition is available on voice assistants. For example, Amazon Guard, recognizes sounds of breaking glass or smoke alarms and send alerts to your phone [40]. Adversaries could remotely play sounds of breaking glass triggering Amazon Guard, potentially causing alarm and fear.

Adversarial noise subverts audio, shifting it closer to a decision boundary. Extrapolating, audio without voice commands could be shifted over a decision boundary, and thus recognized by speech classifiers as commands. Generalizing, stricken users may be oblivious to commands their voice assistants are executing. Common voice commands allow activating attached devices, change thermostat settings, make purchases, call or SMS contacts, donate money to charities, switch off security systems, or lock or unlock doors. In General, users are vulnerable to financial loss, and perhaps physical harm through attached devices. It is unproven what could be possible in large-scale attacks. For example, adding Pepsi to millions of users shopping lists. We conjecture could cause shops to sell out, and perhaps even affect stock market valuations. We further conjecture that voice attacks could open malicious websites, perhaps infecting devices with malware through drive-by downloads. Commands could manipulate devices into opening propaganda or political websites, aiming to change users' political preferences. Ultimately the severity of voice attacks depends on the voice assistant and commands it accepts [22], examples include though are not limited to: posting to

social media, activating airplane mode, or opening malicious websites, unlocking doors, and activating devices.

VIII. Conclusion

We explored the nature of voice assistants and devices susceptibility to attacks based on falsified audio. From our small study, we show how a system could manipulate an audio sample, such that the difference is inaudible but cause the output from a machine learning model to differ significantly. We also surveyed home users to understand attitudes and perceptions around home security and voice assistants.

Our future work investigates the nature of adversarial attacks in machine learning, and how greater protections can be developed combating threats, by attempting to identify false inputs before they are accepted as input to learning models.

References

[1] A. Hern, "Apple contractors 'regularly hear confidential details' on siri recordings," *The Guardian*, vol. 26, 2019.

[2] J. Austerjost, M. Porr, N. Riedel, D. Geier, T. Becker, T. Scheper, D. Marquard, P. Lindner, and S. Beutel, "Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments," *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 23, no. 5, pp. 476–482, 2018.

[3] J. Miehle, D. Ostler, N. Gerstenlauer, and W. Minker, "The next step: intelligent digital assistance for clinical operating rooms," *Innovative Surgical Sciences*, vol. 2, no. 3, pp. 159–161, 2017.

[4] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 103–117.

[5] F. Alegre, G. Soldi, N. Evans, B. Fauve, and J. Liu, "Evasion and obfuscation in speaker recognition surveillance and forensics," in *2nd International Workshop on Biometrics and Forensics*, 2014, pp. 1–6.

[6] J. Danaher, "Toward an ethics of ai assistants: an initial framework," *Philosophy & Technology*, vol. 31, no. 4, pp. 629–653, 2018.

[7] A. M. Turing, "Computing machinery and intelligence-am turing," *Mind*, vol. 59, no. 236, p. 433, 1950.

[8] M. Wu, S. Panchapagesan, M. Sun, J. Gu, R. Thomas, S. N. P. Vitaladevuni, B. Hoffmeister, and A. Mandal, "Monophone-based background modeling for two-stage on-device wake word detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5494–5498.

[9] Z. Zhou and C. Firestone, "Humans can decipher adversarial images," *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[10] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[11] H. Kwon, H. Yoon, and D. Choi, "Restricted evasion attack: Generation of restricted-area adversarial example," *IEEE Access*, vol. 7, pp. 60 908–60 919, 2019.

[12] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, 2020.

[13] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[14] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.

[15] A. K. Bhowmik, "Natural and intuitive user interfaces with perceptual computing technologies," *Information Display*, vol. 29, no. 4, pp. 6–10, 2013.

[16] E. Alepis and C. Patsakis, "Monkey says, monkey does: security and privacy on voice assistants," *IEEE Access*, vol. 5, pp. 17 841–17 851, 2017.

[17] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein, "Voco: text-based insertion and replacement in audio narration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.

[18] J. Titcomb, "Cyber criminals cons energy company in artificial intelligence scam," *Daily Telegraph*, p. 33, 2019.

[19] S. Soltan, P. Mittal, and H. V. Poor, "Blackiot: Iot botnet of high wattage devices can disrupt the power grid," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 15–32.

[20] K. Rajaratnam and J. Kalita, "Noise flooding for detecting audio adversarial examples against automatic speech recognition," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 197–201.

[21] C. Kasmi and J. L. Esteves, "Iemi threats for information security: Remote command injection on modern smartphones," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 6, pp. 1752–1755, 2015.

[22] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 513–530.

[23] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 49–64.

[24] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[25] R. S. Shaw, C. C. Chen, A. L. Harris, and H.-J. Huang, "The impact of information richness on information security awareness training effectiveness," *Computers & Education*, vol. 52, no. 1, pp. 92–100, 2009.

[26] A. Pfeifle, "Alexa, what should we do about privacy: Protecting privacy for users of voice-activated devices," *Wash. L. Rev.*, vol. 93, p. 421, 2018.

[27] S. Zuboff, "Surveillance capitalism and the challenge of collective action," in *New labor forum*, vol. 28. SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 10–29.

[28] S. Ahern, D. Eckles, N. S. Good, S. King, M. Naaman, and R. Nair, "Over-exposed? privacy patterns and considerations in online and mobile photo sharing," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 357–366.

[29] S. Spiekermann, J. Grossklags, and B. Berendt, "E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior," in *Proceedings of the 3rd ACM conference on Electronic Commerce*, 2001, pp. 38–47.

[30] H. Siddique, "Nhs teams up with amazon to bring alexa to patients," https://www.theguardian.com/society/2019/jul/10/nhs-teams-up-with-amazon-to-bring-alexa-to-patients, Jul 10 2019.

[31] G. Horsman, "Loose-lipped mobile device intelligent personal assistants: A discussion of information gleaned from siri on locked ios devices," *Journal of forensic sciences*, vol. 64, no. 1, pp. 231–235, 2019.

[32] S. Preibusch, "Guide to measuring privacy concern: Review of survey and observational instruments," *International Journal of Human-Computer Studies*, vol. 71, no. 12, pp. 1133–1143, 2013.

[33] A. McCarthy, B. R. Gaster, and P. Legg. (2019) Voice assistant survey report. https://mccarthya-s3-bucket.s3.eu-west-2.amazonaws.com/publications/ShoutingThroughLetterboxes/index.html.

[34] A. Zhang, "Speechrecognition 3.8.1," https://pypi.org/project/SpeechRecognition/, 2017.

[35] Amazon, "Amazon echo (2nd gen) - smart speaker with alexa - heather grey fabric," https://www.amazon.co.uk/d/Amazon-Echo-Devices/Amazon-Echo-2nd-Generation-Heather-Grey-Fabric/B0749YXKYZ, 2019.

[36] S. Mills, "Scott mills daily - alexa, who is scott mills?" https://www.bbc.co.uk/programmes/p0717jrd, 2019.

[37] A. V. Limited, "South park: White people renovating houses," 2017.

[38] J. Duddington, "espeak text to speech," http://espeak.sourceforge.net/index.html, 2007.

[39] E. First, "100 most common words in english — learn english," http://www.ef.com/wwen/english-resources/english-vocabulary/top-100-words/, 2019.

[40] J. Graham, "Alexa guard can now listen for alarms - or, perhaps, a cheating spouse?" *USA Today*, 2019.