A Taxonomy of Attacks via the Speech Interface

Mary K. Bispham, Ioannis Agrafiotis, Michael Goldsmith
Department of Computer Science
University of Oxford, United Kingdom
Email: {mary.bispham, ioannis.agrafiotis, michael.goldsmith}@cs.ox.ac.uk

***Abstract*—This paper investigates the security of human-computer interaction via a speech interface. The use of speech interfaces for human-computer interaction is becoming more widespread, particularly in the form of voice-controlled digital assistants. We argue that this development represents new security vulnerabilities which have yet to be comprehensively investigated and addressed. This paper presents a comprehensive review of prior work in this area to date. Based on this review, we propose a high level taxonomy of attacks via the speech interface. Our taxonomy systematises the prior work on the security of voice-controlled digital assistants, and identifies new categories of potential attacks which have yet to be investigated and thus represent a focus for future research.**

Keywords*–*cyber security; human-computer interaction; voice-controlled digital assistants; speech interface.

I. Introduction

The introduction of a speech interface represents a potential expansion of a system's attack surface. With regard to voice-controlled digital assistants, there are clearly serious security concerns arising from an increasingly pervasive presence of such agents. Voice-controlled digital assistants are being used to perform an increasing range of tasks, including Web searching and question answering, diary management, sending emails, and posting to social media. Such 'assistants' are intended to act as brokers between users and the vastly complex, often intimidating cyber world. Their functionalities are being expanded from personal to business use [1]. Sarikaya [2] refers to personal digital assistants as a "metalayer of intelligence" between the user and various different services and actions. With the advent of assistants, such as Amazon's Alexa, which can be used to control smart home devices, control of systems via a speech interface has furthermore extended beyond purely virtual environments to include also cyber-physical systems. Pogue [3] describes voice control as a "breakthrough in convenience" for the Internet of Things. Speech interfaces may eventually be used in time-sensitive and even life-critical contexts, such as hospitals, transport and the military [4] [5]. There is some speculation that communication with computers via natural language represents the next major development in computing technology [6].

Notwithstanding its potential benefits, security concerns associated with such a development have yet to be comprehensively addressed. There has been a considerable amount of debate on the threat to privacy from 'listening' devices, highlighted perhaps most dramatically in a recent request for speech data from Amazon's Alexa as a 'witness' in a murder inquiry [7]. By comparison, the security issues associated with voice-controlled assistants have to date received relatively little attention. Such security issues are however significant. A speech interface potentially enables an attacker to gain access to a victim's system without needing to obtain physical or internet access to their device. Thus, the human-like digital personas intended to give users a sense of familiarity and control in interactions with their systems may in reality be exposing users to additional risks. Internet security company AVG pointed out in 2014 the danger of the speech interface being exploited as a new attack surface, demonstrating how smart TVs and voice assistants might respond to synthesised speech commands crafted by an attacker as well as to their users' voices [8]. The reality of this possibility was recently illustrated by a TV advertisement which contained spoken commands for activation of Google Home on listeners' phones for product promotion purposes. The advert was criticised as a potential violation of computer misuse legislation in gaining unauthorised access to listeners' systems [9]. Another example was an instance in which it was shown to be possible to open a house door from the outside by shouting a command to digital assistant Siri (as discussed by Hoy [10]).

This paper provides a review of the research which has been done to date on attacks via the speech interface, and identifies the gaps in this prior work. Based on this review, we propose a new taxonomy of attacks via the speech interface, and make suggestions for further work. The scope of this taxonomy is limited to attacks which gain unauthorised access to a system by sound. It is possible to attack a voice-controlled system other than by sound - in a security analysis of Amazon's Echo, for example, Haack et al. [11] identify three means of attack on such systems. In addition to sound-based attacks, the paper identifies network attacks (e.g., sniffing of speech data in transmission between an individual user's device and a provider's servers) and API-based attacks (which might involve hacking a voice-controlled assistant's API e.g., to change the default wake-up word). However, such attacks not based on sound are not within scope of the taxonomy presented here.

The remainder of the paper is structured as follows. Section II provides general background on human-computer interaction by speech with reference to the current generation of voice-controlled digital assistants. Section III contains a review of prior work relevant to the security of voice-controlled digital assistants as well as some indirectly relevant work in related areas of research. Section IV proposes a new high-level taxonomy of attacks via the speech interface, including attacks which have been demonstrated in prior work as well as attacks which may be possible in the future. Section V concludes the paper and contains some suggestions for future research.

II. Background

Speech interfaces which facilitate the execution of particular actions in response to voice commands are referred to as 'task-based' speech dialogue systems, as distinct from 'chatbots', whose purpose is simply to hold a conversation

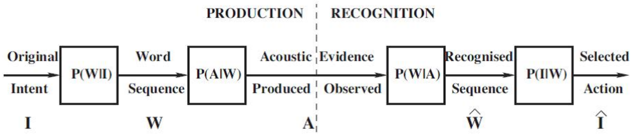

Figure 1. An example of integrated speech and language processing: personal assistance seen as information transmission across a noisy channel [13]

with the user without executing any actions. Current task-based dialogue systems have some similarity with chatbots in that they are often anthropomorphosised, with systems being given the persona of a friendly digital assistant in order to create a sense of communication with a human-like conversation partner. The first voice-controlled digital assistant to be released commercially was Apple's Siri in 2011. Siri was based on an earlier system named Cognitive Assistant that Learns and Organizes (CALO), which had been developed with US defence funding. Siri was followed by the release of Amazon's Alexa in 2014, Microsoft's Cortana in 2015, and most recently in 2016 by Google Assistant [12].

Input to a speech dialogue system is provided by a microphone which captures speech sounds and converts these from analog to digital form. Bellegarda and Monz [13] describe the task of the speech recognition component as the task of extracting from a set of acoustic features the words which generated them, and the task of the natural language understanding component as the task of extracting from a string of words a semantic representation of the user intent behind them. The paper by Bellegarda and Monz conceptualises the process of a user's communication of intent to a speech dialogue system as information transmission across a noisy channel, whereby the user first formulates their intent in words and then vocalises these words as speech, and the dialogue system subsequently extracts from the user's speech the words which generated the speech and then extracts from the words a semantic representation of the intent which generated them. This process is illustrated in the diagram in Figure 1, copied from Bellegarda and Monz's paper.

The typical architecture of a generic speech dialogue system consists of components for speech recognition, natural language understanding, dialogue management, response generation and speech synthesis (see Lison and Meena [14]). The speech recognition and natural language understanding components are the components most likely to be targeted in an attack via the speech interface. Speech recognition is typically performed using Hidden Markov Models (HMMs). HMMs calculate the most likely word sequence for a segment of speech according to Bayes' rule as the product of the likelihood of acoustic features present in the speech segment and the probability of the occurrence of particular words in the sentence context (see for example Juang and Rabiner [15]). HMM-based systems for speech recognition originally used Gaussian Mixture Models (GMMs) for the acoustic modelling and n-grams for the language modelling. In recent years, a shift in modelling methods has been seen with the advent of deep learning. Huang et al. [16] describe recent developments in which Deep Neural Networks (DNNs) have replaced GMMs to extract acoustic model probabilities, and Recurrent Neural Networks (RNNs), a particular type of DNN, have replaced n-grams to extract language model probabilities. Speech recognition technology has become quite advanced. In 2016, Microsoft Research reported that its automatic speech recognition capability had for the first time matched the performance of professional human transcriptionists, achieving a word error rate of 5.9 per cent on the Switchboard dataset of conversational speech produced by the National Institute of Standards and Technology (NIST) in the US (see Xiong et al. [17]).

Natural language understanding in the context of a voice-controlled system is the task of extracting from a user's request a computational representation of its meaning which can be used by the system to trigger an action. The task of mapping a string of words to a representation of their meaning is known as semantic parsing. Liang [18] gives as an example of semantic parsing the instance where a request to cancel a meeting is mapped to a logical form which can be executed by a calendar API. The process of semantic parsing may include syntactic analysis as an intermediate step. Methods of syntactic analysis used in voice-controlled systems include dependency parsing, which is the task of determining syntactic relationships within a sentence, such as verb-object connections (see for example McTear [19]). Current speech dialogue systems typically use semantic representations known as semantic frames (see Sarikaya et al. [20]). Semantic frames provide a structure for representing the meaning of utterances which requires firstly identification of the general domain or concept which a user request relates to (such as travel), secondly determination of the user intent (such as to book a flight), and thirdly slot-filling which involves identifying specific information relevant to the particular request (such as destination city). Sarikaya [2] state that the tasks of domain identification and intent determination in semantic parsing to frames are often performed using support vector machines, whereas slot-fitting is commonly performed using Conditional Random Fields (CRFs). Some recent research has indicated that traditional machine learning methods are now being out-performed in the semantic parsing task for spoken dialogue systems by neural networks, similar to the replacement of n-gram-based systems for language modelling in speech recognition by RNNs. Mesnil et al. [21], for example, present results showing superior performance by RNNs on the slot-filling task for the Air Travel Information System (ATIS) dataset in comparison to the performance of CRFs on the same task. Despite such efforts, it is clear that, unlike in the case of speech recognition, the state-of-the-art in natural language understanding remains far from parity with human capabilities. This is evident in the occasional failure of voice assistants to correctly interpret the meaning of a word in context, despite the correct word or meaning being obvious to any human listener. Stolk et al. [22] give the examples of Apple's assistant Siri mistaking the word 'bank' in the sense of 'river bank' for a financial institution, and of Siri giving directions to a casino when asked about a gambling problem.

Modern voice-controlled digital assistants implement the generic components of speech dialogue systems in the context of a cloud-based service which enables users to interact by voice with smartphones and laptop/desktop computers, as well as to control smart home devices by voice using bespoke hardware. The speech recognition and natural language understanding functionalities of these systems are performed in the

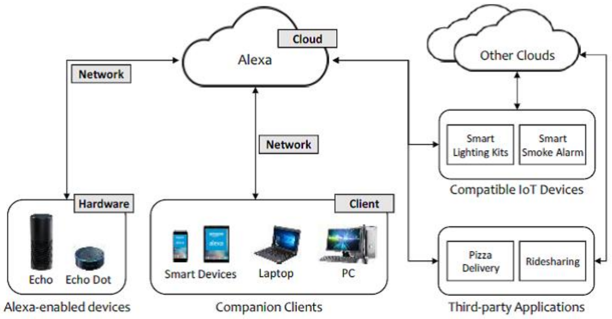

Figure 2. Amazon Alexa Ecosystem [23]

provider's cloud. Chung et al. [23] provide an overview of the typical ecosystem of modern voice-controlled digital assistants in the example of Amazon's Alexa (see Figure 2).

In order to control streaming of audio data to the cloud, current voice-controlled digital assistants include, in addition to the generic speech dialogue system components, an activation component consisting a of wake-up word, which, when spoken by the user, triggers streaming of the subsequent speech audio data to the provider's cloud for processing. Examples of wake-up words include 'Ok Google' for Google Assistant and 'Alexa' for Amazon's Alexa. Wake-up word recognition is the only speech processing capability on users' individual devices, and consists of a short 'buffer' of audio data from the device's environment which is continuously recorded and deleted [24]. Wake-up word activation can be triggered by false positives. Chung et al. [25], for example, refer anecdotally to accidental activation of the Alexa assistant by a sentence containing the phrase 'a Lexus' (see also Michaely et al. [26]), and Vaidya et al. [27] refer to the misrecognition of the phrase "Cocaine Noodles" as "OK Google". False positives in wake-up word recognition may result from misrecognition of a word as the wake-up word, as in the example given by Chung et al., or else from use of a wake-up word in the context of speech not intended to activate a voice assistant, for example the use of the word 'Alexa' as the name of a person in a conversation. Këpuska and Bohouta [28] discuss the latter problem of distinguishing between an 'alerting' and a 'referential' context in wake-up word recognition. It is also possible for voice assistants to be activated by background noise which has frequencies overlapping with those of human speech (see Islam et al. [29]).

The current generation of voice-controlled digital assistants have also introduced platforms for the development of third-party voice applications which can be incorporated in the provider's cloud and made available to users via the assistant's speech interface. Examples of such third-party applications are Alexa Skills and Google Conversation Actions. Third-party applications in systems such as Google Assistant can be accessed by users by asking to 'speak' to the voice app (as named by the developer) [30]. Such apps can be used for example to enable users to access information services or to purchase products.

III. Prior Work

There has been a limited amount of prior work on the security of speech interfaces and voice-controlled digital assistants, as well as some prior work in related areas of research. A review of prior work relevant to attacks via the speech interface of voice-controlled digital assistants is presented, and summarised in Table 1. Whilst the review is concerned with sound-based attacks only, it is recognised that attacks by sound are only a subset of the potential attacks which might be targeted at a voice-controlled digital assistant. The review does not analyse the specific aims of the attacks described in prior work beyond the general goal of gaining unauthorized access to a system via a speech interface.

Several researchers have investigated the ways in which voice-controlled digital assistants might be exploited simply by using standard voice commands. This possibility arises out of the inherently open nature of natural speech. Such potential vulnerabilities associated with speech-controlled systems have been highlighted for example by Dhanjani [31], who describes a security vulnerability identified in Windows Vista which allowed an attacker to delete files on a victim's computer by playing an audio file hosted on a malicious website or sent to the victim as an email attachment. Dhanjani speculates that the potential for such attacks is magnified with the increasing use of speech recognition technology in the Internet of Things. He postulates a hypothetical attack on Amazon's Echo, a device designed to be used for voice control of home appliances via digital assistant 'Alexa', which would potentially cause psychological or physical harm to the victim by controlling their smart home environment. This hypothetical attack involves a piece of malware consisting of JavaScript code which plays an audio file giving a command to Alexa if there has been no user activity on the mouse or keyboard after a certain period of time (thus aiming to play the file at a time when the user may be away from their computer and therefore will not hear the audio command being played). Diao et al. [32] investigate possibilities for gaining unauthorised access to a smartphone via a malicious Android app which uses the smartphone's own speakers to play an audio file containing voice commands. The attacks proposed by the authors include an attack in which the smartphone is manipulated to dial a phone number which connects to a recording device, and then to disclose information such as the victim's calendar schedule by synthesised speech which is recorded by the device. Diao et al. envisage such attacks being executed whilst the victim is asleep and therefore unable to hear the malicious voice command. Such an attack might in fact be executed whilst the victim is neither away from their phone or asleep, but their attention is merely directed elsewhere.

Kasmi and Esteves describe a different type of attack in which voice commands are transmitted silently to a victim's phone via electromagnetic interference using the phone's headphones as an antenna [33]. Unlike plain-speech attacks, this attack is not detectable even if the victim is consciously present at the time of the attack, although for technical reasons the attack can only be performed if the attacker is in close proximity to the victim's device. The types of attack envisaged by Kasmi and Esteves include controlling transmissions from a smartphone by activating or deactivating Wifi, Bluetooth, or airplane mode, and browsing to a malicious website to effect drive-by-download of malware. Young et al. [34] also describe a 'silent' attack on smartphones via the voice command interface which enables an attacker to perform actions such as calling fee-paying phone numbers, posting to Facebook

TABLE I. SUMMARY OF PRIOR WORK RELEVANT TO ATTACKS VIA THE SPEECH INTERFACE

Paper	Attack Target	Attack Category
Dhanjani [31]	speech interface in PC (Windows Vista)	plain speech (overt)
Diao et al. [32]	speech interface in voice-controlled digital assistant (Google Voice Search)	plain speech (overt)
Kasmi and Esteves [33]	voice capture in voice-controlled digital assistant (Google Now, Siri)	silence (covert)
Young et al. [34]	voice capture in voice-controlled digital assistant (Siri)	silence (covert)
Zhang et al. [35]	voice capture in voice-controlled digital assistant (Apple Siri, Amazon Alexa, Microsoft Cortana and others)	silence (covert)
Song and Mittal [36]	voice capture in voice-controlled digital assistant (Google Now, Amazon Alexa)	silence (covert)
Vaidya et al. [27]	speech recognition in voice-controlled digital assistant (Google Now)	noise (covert)
Carlini et al. [37]	speech recognition in voice-controlled digital assistant (Google Now) / speech recognition (CMU Sphinx)	noise (covert)
Iter et al. [38]	speech recognition in speech transcription system (WaveNet)	missense (covert)
Cisse et al. [39]	speech recognition in voice-controlled digital assistant (Google Voice)	missense (covert)
Alzantot et al. [40]	speech recognition in speech transcription system (TensorFlow)	missense (covert)
Carlini and Wagner [41]	speech recognition in speech transcription system (DeepSpeech)	music/missense (covert)
Yuan et al. [42]	speech recognition in speech transcription system (Kaldi)	music (covert)
Papernot et al. [43]	natural language understanding in sentiment analysis system	nonsense (covert)
Liang et al. [44]	natural language understanding in text classification system	missense (covert)
Jia and Liang [45]	natural language understanding in question answering system	missense (covert)

in the victim's name to damage their reputation, accessing email messages, and changing website passwords from the victim's phone. The attack requires a short period of time during which an attacker has unsupervised physical access to the phone in order to to attach a Raspberry Pi-based tool which is recognised by the phone as headphones with a microphone. Zhang et al. [35] and Song and Mittal [36] present methods for injecting voice commands to voice-controlled digital assistants at inaudible frequencies by exploiting non-linearities in the processing of sounds by current microphone technology, which can lead voice-controlled systems to detect a command as having been issued within the human audible frequency range, despite the sound not having been perceptible to humans in reality. Silent attacks such as these target the 'voice capture' stage of voice control, i.e., the process of conversion of speech sounds by the microphone from analog to digital form prior to speech recognition.

There has also been some prior work towards using adversarial machine learning in attacks on voice-controlled digital assistants. The aim of adversarial machine learning is to identify instances in which a machine learning-based system classifies input in a way that a human would regard as erroneous. This is done by some form of systematic exploration of the system's input space with the aim of discovering 'adversarial examples' within that space. Some adversarial machine learning methods involve manipulating inputs based on knowledge of calculations within the classifier (such 'white-box' methods include approaches such as the Fast Gradient Sign Method and the Jacobian-based Saliency Map Approach for altering input to a DNN, as described for example in Goodfellow et al. [46]). Other methods seek to manipulate input on a 'black-box' basis i.e., without knowledge of the inner workings of a target system. McDaniel et al. [47] explain that processes of adversarial machine learning rely on identifying 'adversarial regions' in a classification category which have not been covered by training examples. The exact reasons for the effectiveness of particular adversarial examples are difficult to determine, as the decision-making process in a neural network cannot be precisely reverse-engineered (see for example Castelvecchi [48]). In this sense, whilst some adversarial learning methods require more knowledge of the target network than others, all attacks on DNN-based systems are of necessity 'black-box' attacks, although attacks requiring detailed knowledge of the system's functionality are referred to here as white-box in order to distinguish them from attacks not requiring such detailed knowledge.

Adversarial learning to attack DNN-based systems was first demonstrated in image classification (see for example Szegedy et al. [49]), but has recently also been applied to speech recognition. One example is the work presented by Vaidya et al. [27], who used audio mangling to distort commands issued to precursor to Google Assistant Google Now (this 'mangling' involved reverse MFCC, where MFCC features extracted from a speech sound were used to generate a mangled version of the sound). The mangled commands included commands to open a malicious website, make a phone-call and send a text, in addition to the Google Now wake-up command 'Ok Google'. The work showed that the distorted commands continued to be recognised by the speech recognition system despite being no longer recognisable by humans, who perceived them instead as mere noise. Thus, the distorted commands represented adversarial examples for the target system. The work by Vaidya et al. was expanded by Carlini et al. [37], who also proved the possibility of prompting Google Now to execute mangled commands which had been shown to be unintelligible to humans in an experiment using Amazon Mechanical Turk. The attacks by Vaidya et al. and Carlini et al. on Google Now were 'black-box' attacks i.e., they were constructed without knowledge of the inner workings of the speech recognition system. Carlini et al. additionally conducted a successful 'white-box' attack on Carnegie Mellon University's SPHINX speech recognition system (based on GMMs rather than DNNs), in which 'mangled' adversarial commands were crafted with knowledge of the workings of the system.

Other work on adversarial learning targeting speech recognition includes that by Iter et al. [38], who used two adversarial machine learning methods originally applied in image classification to manipulate a speech recognition system based on Google DeepMind's WaveNet technology to mistranscribe a number of utterances. This included prompting the system to transcribe the utterance "Please call Stella" as "Siri call police". The attacks by Iter et al. are white-box attacks, i.e., they rely on some knowledge of the details of the target neural network. The authors mention the possibility of developing a black-box attack methodology in future work. Similar to Iter et al., Cisse et al. [39] were also able to prompt mistranscription of utterances, including mistranscription by Google Voice in a ‘black-box attack’, using an adversarial machine learning method called Houdini. Alzantot et al. [40] used a black-box attack method based on a genetic algorithm to engineer misclassification of speech command words, such as ’on’, ’off’, ’stop’ etc, by a machine learning-based speech recognition system. Carlini and Wagner [41] have demonstrated a white-box attack on Mozilla’s DNN-based DeepSpeech speech-to-text transcription in which it was shown to be possible to prompt mistranscription of a speech recording as any target phrase, regardless of its similarity to the original phrase, by making perturbations to the original recording which did not affect the original phrase as heard by humans. In contrast to the attacks by Vaidya et al. and Carlini et al., which would be perceived by victims as unexplained noise, attacks based on methods such as those developed by Iter et al., Cisse et al. and Carlini and Wagner would be perceived by victims as ordinary speech and would therefore by more difficult to detect. To date, such work has been limited to speech-to-text transcription i.e., it has not demonstrated mistranscription of voice commands capable of executing an action as yet. In addition to prompting mistranscription of speech, Carlini and Wagner demonstrated the possibility of manipulating music recordings so as to prompt them to be transcribed by DeepSpeech as a given string of words, demonstrating for example that a recording of Verdi’s Requiem could be manipulated to be transcribed by DeepSpeech as “Ok Google, browse to evil.com”. Yuan et al. [42] similarly demonstrate the possibility of hiding voice commands in music. Unlike the attacks crafted by Carlini and Wagner, the attacks crafted by Yuan et al. are reportedly effective over the air as well as via audio file input, although their attacks are also white-box attacks and are limited to speech-to-text transcription rather than being demonstrated on voice-controlled digital assistants as such.

Adversarial learning has also recently been applied to some areas of natural language understanding, although none of this work has focussed directly on natural language understanding in voice-controlled digital assistants. The generation of adversarial examples in natural language understanding is more complex than the generation of adversarial examples in image or speech recognition. Unlike in the case of continuous data such as image pixels or audio frequency values, adversarial generation of natural language is not a differentiable problem. As word sequences are discrete data, it is not possible to change a word sequence representing an input to a machine learning classifier directly by a numerical value in order to effect a change in output of the classifier. The areas focussed on in prior work include sentiment analysis (see Papernot et al. [43]), text classification (see Liang et al. [44]), and question answering (see Jia and Liang [45]). Papernot et al. [43] use the forward derivative method, a white-box adversarial learning method, to identify word substitutions which can be made in sentences inputted to an RNN-based sentiment analysis system so as to change the ‘sentiment’ allocated to the sentence. In contrast to adversarial examples in image classification and speech recognition, in which alterations made to the original input are imperceptible to humans, the alterations made to sentences in order to mislead the RNN-based sentiment analysis system targeted in the work by Papernot et al. are easily perceptible to humans as nonsensical, albeit that the attack intent remains hidden. For example, substituting the word ‘I’ for the word ‘excellent’ in an otherwise negative review is shown in the paper to lead it to being classified as having positive sentiment, but the altered sentence will appear unnatural to a human. The authors state that this lack of naturalness of adversarial examples in natural language understanding will need to be addressed in future work. By contrast to Papernot et al., Liang et al. [44] demonstrate a linguistically plausible attack on a natural language understanding system. The authors adapt the Fast Gradient Sign Method from adversarial learning in image classification to make human-undetectable alterations to a text passage (by adding, modifying and/or removing words) so as to change the category which is allocated to the passage by a DNN-based text classification system. The attack by Liang et al. is white-box, requiring details of the calculations inside the network. Jia and Liang [45] also demonstrate a linguistically plausible attack in the context of question answering. Their work involves misleading a number of question answering systems by adding apparently inconsequential sentences to text passages from which the systems extract answers to questions. The method works by first choosing a target wrong answer to a given question, and then crafting a sentence containing information leading to this wrong answer which can be inserted into the original passage without noticeably changing its overall import. The attack method proposed by Jia and Liang is a black-box method, not requiring knowledge of the internal details of the target network.

IV. Taxonomy

Reflecting on the review of prior work above, we propose a high-level taxonomy of categories of attacks via the speech interface. This taxonomy is presented in Figure 3. Table 1 shows the categorization of each of the attacks from prior work in terms of the high-level taxonomy. The taxonomy covers attack types which have been demonstrated in prior work as well as attack types for which potential is implied by related work in other areas. The principle behind the taxonomy is to identify the various categories of non-speech and speech sounds which humans are capable of perceiving, and to group attacks via the speech interface according to these categories, including both attacks which have been demonstrated in prior work as well as attacks which may be possible in the future as implied by prior work in related areas. By applying this principle, the taxonomy fulfils the dual purpose of systematising prior work whilst also identifying new directions for future research. Attacks via the speech interface as categorised under our taxonomy might be targeted at any voice-controlled system, including any voice-controlled digital assistant and any third-party applications accessible through it.

The taxonomy was developed according to established criteria for attack taxonomies, as described for example in Hansman and Hunt [50]. These criteria include the requirement that a taxonomy should be ’complete’ i.e., cover all possible attacks within its scope, and unambiguous i.e., it should be possible clearly to allocate every attack to one category within the scope of the taxonomy. In order to meet these criteria, a categorisation principle was chosen for the taxonomy of grouping attacks according to the nature of attacks via the

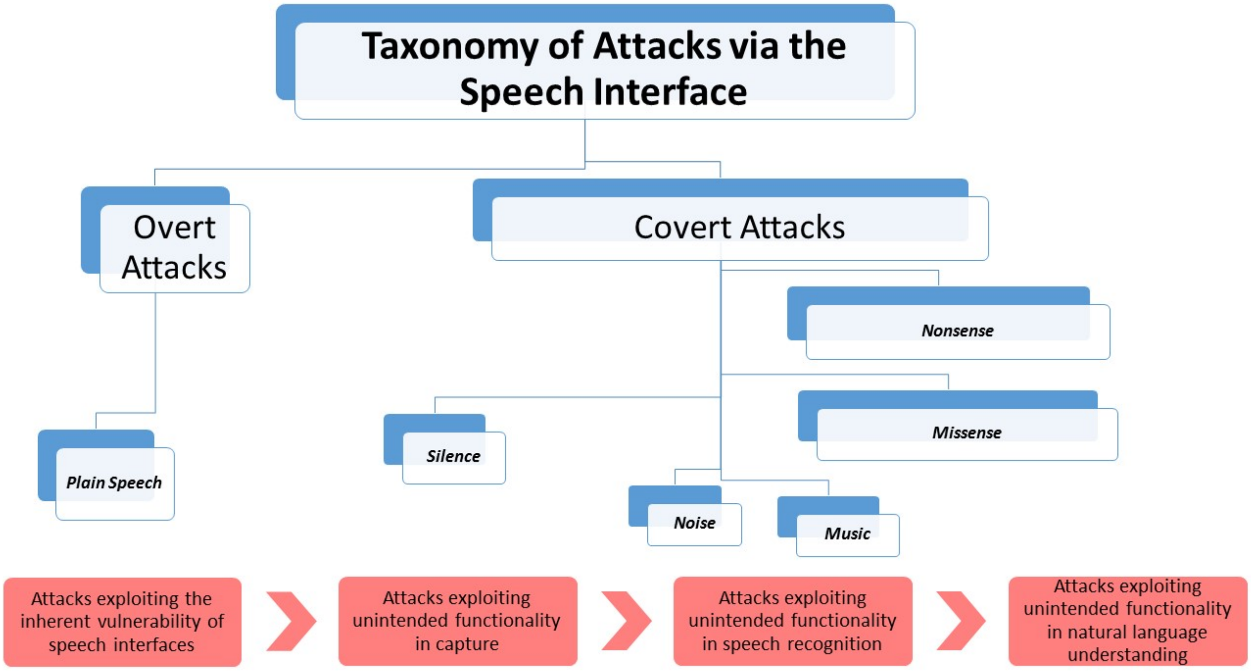
Taxonomy of Attacks via the Speech Interface
Overt Attacks
Covert Attacks
Plain Speech
Silence
Noise
Music
Missense
Nonsense
Attacks exploiting the inherent vulnerability of speech interfaces
Attacks exploiting unintended functionality in capture
Attacks exploiting unintended functionality in speech recognition
Attacks exploiting unintended functionality in natural language understanding

Figure 3. Taxonomy of Attacks via the Speech Interface

speech interface as they might be perceived by a human listener. Within the framework of this categorisation principle, six final categories of attacks via a speech interface were identified, namely attacks consisting of plain speech, silence, music, noise, nonsense, and 'missense', missense being unrelated speech which is misheard or misinterpreted by the target system. These final categories are ordered hierarchically in our taxonomy as detailed below. The principle of categorising attacks according to human perception ensures that the taxonomy is complete, as all attacks via a speech interface can be allocated to one of the six categories. The taxonomy is also unambiguous, in that it is not possible to allocate the same voice attack to more than one of the final categories. To the extent that speech processing by voice-controlled systems mimics human speech processing, the attack categories also reflect vulnerabilities in the different parts of the architecture of voice-controlled systems, as shown at the bottom of Figure 3.

In the taxonomy, the six final categories of voice attacks identified within the categorisation framework are primarily grouped into two categories: 'overt' attacks, which seek to gain unauthorised access to systems using the same voice commands as might be given by a legitimate user and are thus easily detectable by a human, and 'covert' attacks, which seek to gain access using speech commands which have been distorted in some way so as to escape detection by the victim. Another way of characterizing this division is as a distinction between attacks which make illicit use of the intended functionalities of a speech dialogue system, and attacks which exploit unintended functionalities.

Overt attacks exploit an inherent vulnerability in voice-controlled systems which arises from the difficulty of controlling access to a system via the 'speech space'. The plain-speech attacks investigated in prior work, such as that by Dhanjani et al. [31] discussed above, fall into the overt attack category. Covert attacks exploit gaps in the processes of capturing human speech or of translating the captured speech into computer executable actions in a voice-controlled system system. Malicious inputs in covert attacks may include input which consists in human terms of silence, as for example in the attacks demonstrated by Zhang et al. [35], noise, as for example in the attacks demonstrated by Carlini et al. [37], music, as for example in the attacks demonstrated by Yuan et al. [42], missense, as for example in the attacks demonstrated by Carlini and Wagner [41], and nonsense.

Nonsense attacks have yet to be demonstrated with respect to voice-controlled systems directly, although there has been some related work, such as in the attacks on a sentiment analysis system by Papernot et al. [43] by making nonsensical alterations to text. Similar attacks might be demonstrated in the context of voice-controlled digital assistants in future. Prior work on missense attacks in voice-controlled systems has to date been limited to attacks on speech recognition as incorporated in such systems. However, in related work, there have also been examples of missense attacks which target natural language understanding functionality, such the attacks on question answering by Jia and Liang [45] by making apparently inconsequential alterations to text. This suggests that, in the future, missense attacks on voice-controlled systems might target vulnerabilities in natural language understanding as well as in speech recognition. In a missense attack which targets natural language understanding functionality, words might be transcribed correctly by the target system, but their meaning would be misinterpreted. Such missense attacks might seek to exploit the shortcomings of current natural language understanding functionality in voice-controlled digital assistants in terms of being able to identify the correct meaning of words in context.

For missense attacks in the specific context of voice-controlled digital assistants, the need to circumvent the wake-up word activation presents a potential issue of linguistic plausibility. Unlike in the case of attacks hiding commands in silence, noise, or music, it is difficult to incorporate a device's wake-up word as part of an attack based on confusion of meaning. However, due to the known presence of false positives with respect to wake-up word recognition, attacking the activation function of a voice assistant with a missense attack is possible. This possibility was in fact demonstrated in an incident in which an Amazon Alexa device misinterpreted a word spoken in a private conversation as the wake-up word 'Alexa', and subsequently misinterpreted other words in the conversation as commands to send a message to a contact, resulting in a recording of a couple's private conversation in their home being sent to a colleague [51]. Whilst this transmission of private information occurred as a result of error rather than malicious intent, it highlights the potential for missense attacks on voice-controlled systems which include circumvention of the wake-up word.

V. Conclusion

This paper proposes a taxonomy of attacks via the speech interface which covers attacks investigated in prior work as well as attacks which may be possible in the future. The review of prior work in this paper indicates that the potential for attacks via a speech interface has yet to be comprehensively assessed. The scope of attacks via a speech interface can be expected to expand with the increasing sophistication of voice-controlled systems. Consequently, there is a need for further security-focussed research in the area of voice-controlled technology.

Future work should seek more extensively to demonstrate the potential for attacks in the various categories of the proposed taxonomy in the context of different technologies and use-case scenarios. Among the taxonomy categories, nonsense attacks and missense attacks targeting the natural language understanding functionality of voice-controlled systems represent new types of attacks which have yet to be demonstrated in practice, but may become possible in the future. Thus, such attacks should be a special focus of future work. The results of future work should ultimately be used as a basis for the development of more effective defence measures to improve the security of voice-controlled digital assistants and other voice-controlled systems.

Acknowledgment

This work was funded by a doctoral training grant from the Engineering and Physical Sciences Research Council (EPSRC).

References

[1] "Why Amazon's Alexa may soon become your new colleague," 2017, URL: https://www.inc.com/emily-canal/amazon-alexa-for-business.html [accessed: 2018-07-20].

[2] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components," IEEE Signal Processing Magazine, vol. 34, no. 1, 2017, pp. 67–81.

[3] D. Pogue, "At your command," Scientific American, vol. 315, no. 1, 2016, pp. 25–25.

[4] C. Franzese and M. Coyne, "The promise of voice: Connecting drug delivery through voice-activated technology," vol. 2017, 12 2017, pp. 34–37.

[5] "British navy warships 'to use Siri' as technology transforms warfare," 2017, URL: https://www.theguardian.com/uk-news/2017/sep/12/british-navy-warships-to-use-voice-controlled-system-like-siri [accessed: 2018-07-20].

[6] "The Voice-AI Revolution is a Conversational Interface of Everything," 2017, URL: https://medium.com [accessed: 2018-07-20].

[7] "A Murder Case Tests Alexa's Devotion to Your Privacy," 2017, URL: https://www.wired.com/2017/02/murder-case-tests-alexas-devotion-privacy [accessed: 2018-07-20].

[8] "Voice Hackers Will Soon Be Talking Their Way Into Your Technology," 2014, URL: https://www.forbes.com/sites/jasperhamill/2014/09/29/voice-hackers-will-soon-be-talking-their-way-into-your-technology/ [accessed: 2018-07-20].

[9] "Burger King triggers Google Home devices with TV ad," 2017, URL: https://nakedsecurity.sophos.com/2017/04/18/burger-king-triggers-ok-google-devices-with-tv-ad/ [accessed: 2018-07-20].

[10] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," Medical reference services quarterly, vol. 37, no. 1, 2018, pp. 81–88.

[11] W. Haack, M. Severance, M. Wallace, and J. Wohlwend, "Security analysis of the Amazon Echo," MIT, 2017.

[12] "Google uses Assistant to square up to Siri in AI arms race," 2017, URL: https://www.ft.com/content/f9423056-7efe-11e6-8e50-8ec15fb462f4 [accessed: 2018-07-20].

[13] J. R. Bellegarda and C. Monz, "State of the art in statistical methods for language and speech processing," Computer Speech & Language, vol. 35, 2016, pp. 163–184.

[14] P. Lison and R. Meena, "Spoken dialogue systems: the new frontier in human-computer interaction," XRDS: Crossroads, The ACM Magazine for Students, vol. 21, no. 1, 2014, pp. 46–51.

[15] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, vol. 1, 2005, p. 67.

[16] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," Communications of the ACM, vol. 57, no. 1, 2014, pp. 94–103.

[17] W. Xiong et al., "Achieving human parity in conversational speech recognition," arXiv preprint arXiv:1610.05256, 2016.

[18] P. Liang, "Learning executable semantic parsers for natural language understanding," Communications of the ACM, vol. 59, no. 9, 2016, pp. 68–76.

[19] M. McTear, Z. Callejas, and D. Griol, The conversational interface. Springer, 2016.

[20] R. Sarikaya et al., "An overview of end-to-end language understanding and dialog management for personal digital assistants," in IEEE Workshop on Spoken Language Technology, 2016, pp. 391–397.

[21] G. Mesnil et al., "Using recurrent neural networks for slot filling in spoken language understanding," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 3, 2015, pp. 530–539.

[22] A. Stolk, L. Verhagen, and I. Toni, "Conceptual alignment: how brains achieve mutual understanding," Trends in cognitive sciences, vol. 20, no. 3, 2016, pp. 180–191.

[23] H. Chung, J. Park, and S. Lee, "Digital forensic approaches for amazon alexa ecosystem," Digital Investigation, vol. 22, 2017, pp. S15–S25.

[24] "Alexa and Google Home Record What You Say, But What Happens To That Data?" 2016, URL: https://www.wired.com/2016/12/alexa-and-google-record-your-voice/ [accessed: 2018-07-20].

[25] H. Chung, M. Iorga, J. Voas, and S. Lee, "Alexa, can i trust you?" Computer, vol. 50, no. 9, 2017, pp. 100–104.

[26] A. H. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic, "Keyword spotting for google assistant using contextual speech recognition," in Proceedings of ASRU, 2017, pp. 272–278.

[27] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," Presented at WOOT, vol. 15, 2015, pp. 10–11.

[28] V. Këpuska and G. Bohouta, "Improving wake-up-word and general speech recognition systems," in Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence & Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2017 IEEE 15th Intl. IEEE, 2017, pp. 318–321.

[29] M. T. Islam, B. Islam, and S. Nirjon, "Soundsifter: Mitigating overhearing of continuous listening devices," in Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services. ACM, 2017, pp. 29–41.

[30] "How to use third-party Actions on Google Home," 2017, URL: https://www.cnet.com/uk/how-to/how-to-use-third-party-actions-on-google-home/ [accessed: 2018-07-20].

[31] N. Dhanjani, Abusing the Internet of Things: Blackouts, Freakouts, and Stakeouts. " O'Reilly Media, Inc.", 2015.

[32] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices. ACM, 2014, pp. 63–74.

[33] 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, 2015, pp. 1752–1755.

[34] P. J. Young, J. H. Jin, S. Woo, and D. H. Lee, "Badvoice: Soundless voice-control replay attack on modern smartphones," in Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on. IEEE, 2016, pp. 882–887.

[35] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," arXiv preprint arXiv:1708.09537, 2017.

[36] L. Song and P. Mittal, "Inaudible voice commands," arXiv preprint arXiv:1708.07238, 2017.

[37] N. Carlini et al., "Hidden voice commands," in 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, 2016.

[38] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," Stanford, 2017.

[39] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," arXiv preprint arXiv:1707.05373, 2017.

[40] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," arXiv preprint arXiv:1801.00554, 2018.

[41] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," arXiv preprint arXiv:1801.01944, 2018.

[42] X. Yuan et al., "Commandersong: A systematic approach for practical adversarial voice recognition," arXiv preprint arXiv:1801.08535, 2018.

[43] N. Papernot, P. McDaniel, A. Swami, and R. Harang, "Crafting adversarial input sequences for recurrent neural networks," in Military Communications Conference, MILCOM 2016-2016 IEEE. IEEE, 2016, pp. 49–54.

[44] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," arXiv preprint arXiv:1704.08006, 2017.

[45] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," arXiv preprint arXiv:1707.07328, 2017.

[46] I. Goodfellow, N. Papernot, and P. McDaniel, "cleverhans v0. 1: an adversarial machine learning library," arXiv preprint arXiv:1610.00768, 2016.

[47] P. McDaniel, N. Papernot, and Z. B. Celik, "Machine learning in adversarial settings," IEEE Security & Privacy, vol. 14, no. 3, 2016, pp. 68–72.

[48] D. Castelvecchi, "Can we open the black box of AI?" Nature News, vol. 538, no. 7623, 2016, p. 20.

[49] C. Szegedy et al., "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[50] S. Hansman and R. Hunt, "A taxonomy of network and computer attacks," Computers & Security, vol. 24, no. 1, 2005, pp. 31–43.

[51] "Amazon Alexa heard and sent private chat," 2018, URL: https://www.bbc.co.uk/news/technology-44248122 [accessed: 2018-07-20].