The Speech Interface as an Attack Surface: An Overview

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 and related 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 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. The attack surface presented by the speech interface comprises not only the voice-controlled device itself, but the entire process of human-computer interaction including the human user. In accordance with this, our taxonomy categorises attacks via the speech interface according to human perceptions of the attacks, whilst also aligning the categories of the taxonomy to vulnerabilities in various parts of the architecture of voice-controlled systems. This paper is an extended version of a previous paper in which our taxonomy of attacks via the speech interface was first presented.**

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. This paper presents a comprehensive overview of the types of attacks that might be targeted at voice-controlled systems, and categorises these attacks in a high-level taxonomy. This paper is an extended version of a previous paper in which our taxonomy was first presented (Bispham et al. [1]).

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 [2]. Sarikaya [3] 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 that can be used to control smart home devices, control of systems via a speech interface has extended beyond purely virtual environments to include also cyber-physical systems. Pogue [4] 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 [5] [6]. There is some speculation that communication with computers via natural language represents the next major development in computing technology [7].

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 [8]. 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 [9]. The reality of this possibility was recently illustrated by a TV advertisement that 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 [10]. 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 [11]).

This paper provides a review of the research that 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 that 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. [12] 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 of some indirectly relevant work in related areas of research. This section also includes some speculation on the potential for attacks via the speech interface that are not possible on current commercial systems, but may become possible in future based on current trends in research on speech dialogue systems. Section IV proposes a new high-level taxonomy of attacks via the speech interface, including attacks that have been demonstrated in prior work as well as attacks that may be possible in the future. Section V concludes the paper and contains some suggestions for future research.

II. Background on Voice-Controlled Systems

Speech interfaces that 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 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 [13].

Input to a speech dialogue system is provided by a microphone that captures speech sounds and converts these from analog to digital form. Bellegarda and Monz [14] describe the task of the speech recognition component as the task of extracting from a set of acoustic features the words that 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 that generated the speech and then extracts from the words a semantic representation of the intent that 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 [15]). In current systems, the speech recognition and natural language understanding components are the components most likely to be targeted in an attack via the speech interface. As explained further below, in current systems the dialogue management and subsequent components are fully controlled by input from the speech recognition and natural language understanding components, and can therefore not be targeted directly.

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 [16]). 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. [17] 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. [18]).

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 that 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 [19] gives as an example of semantic parsing the instance where a request to cancel a meeting is mapped to a logical form that 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 [20]). Current speech dialogue systems typically use semantic representations known as semantic frames (see Sarikaya et al. [21]). Semantic frames provide a structure for representing the meaning of utterances that requires firstly identification of the general domain or concept that 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 [3] states 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. [22], 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. [23] give the examples of

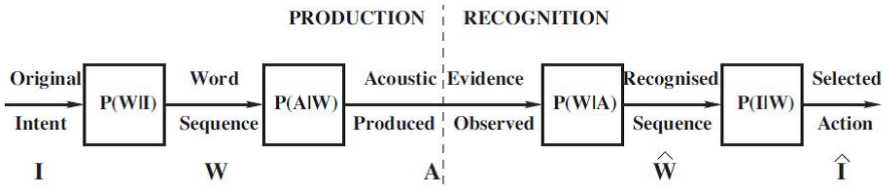

Figure 1. An example of integrated speech and language processing: personal assistance seen as information transmission across a noisy channel [14]

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.

Dialogue management is the task of determining the most appropriate action that should be taken in response to a user's request. The dialogue management component then instructs the response generation and (in the case that a verbal action is required) speech synthesis components of system to take the necessary action. Sarikaya [3] states that the dialogue management task in personal digital assistants is far more challenging than in older speech recognition systems. Older speech recognition systems were commonly limited to one general purpose, such as providing travel information. Digital assistants, by contrast, are designed to perform a large number of tasks, including providing information on many different topics, connecting with web applications to fulfil a variety of user requests, and controlling devices in the Internet of Things. Sarikaya describes the structure of a dialogue manager in a digital assistant as consisting of a dialogue state tracker, which updates the 'state' of the dialogue based on the representation of user intent generated by the natural language understanding module, and a dialogue policy that controls the execution of tasks in response to the user request.

The dialogue management component in current speech dialogue systems is on the whole still rule-based, i.e., it maps user intent to dialogue states and dialogue states to actions based on hand-crafted rules, as stated by McTear [20]. The dialogue management capabilities in current systems are thus fully dependent on input from the speech recognition and natural language understanding components and do not therefore represent a separate point of attack. Rule-based dialogue management systems have the advantage of limiting the potential for error and unintended functionality in the dialogue management process (see McTear [24]). However, such systems are also likely to be lacking in flexibility and limited in scope. There has been some research on the eventual replacement of current rule-based systems by more sophisticated dialogue management systems based on reinforcement learning, which would enable voice assistants to learn directly from their interactions with users. Young et al. [25] propose ideas for dialogue management based on Partially Observable Markov Decision Processes (POMDPs), which model a dialogue as a Markov process with transition probabilities between states, for which a probability distribution over all possible states is continuously maintained. This approach seeks to represent the uncertainty inherent in the fact that a user's intent is not directly observable, but rather inferred probabilistically from their utterance. Systems based on POMDPs combine Bayesian inference for belief state tracking to determine the most likely interpretation of a user's utterances with reinforcement learning for optimisation of the dialogue policy, whereby a reward function is used to train the system as to the most appropriate action to take in response to a user utterance based on user feedback.

Modern voice-controlled digital assistants implement the generic components of speech dialogue systems in the context of a cloud-based service that 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 provider's cloud. Chung et al. [26] 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 of a 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 that is continuously recorded and deleted [27]. Wake-up word activation can be triggered by false positives. Chung et al. [28], for example, refer anecdotally to accidental activation of the Alexa assistant by a sentence containing the phrase 'a Lexus' (see also Michaely et al. [29]), and Vaidya et al. [30] 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 [31] 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 that has frequencies overlapping with those of human speech (see Islam et al. [32]). The vulnerability of wake-up word recognition to false positives was 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

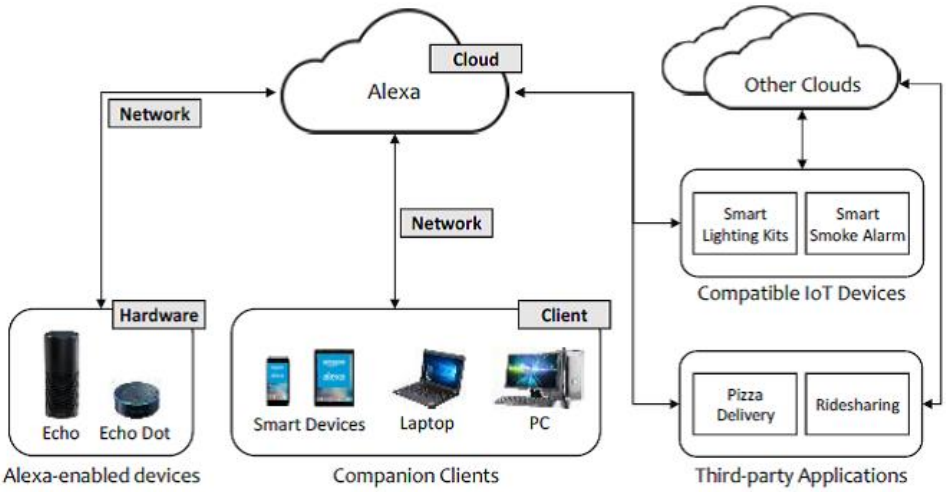

Figure 2. Amazon Alexa Ecosystem [26]

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 [33].

The current generation of voice-controlled digital assistants have also introduced platforms for the development of third-party voice applications that 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) [34]. Such apps can be used for example to enable users to access information services or to purchase products.

III. Attacks via the Speech Interface in Prior and Related 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 on the current generation of voice-controlled digital assistants is presented, and summarised in Table I. Our review further includes some speculation on attacks that are not possible in relation to the current generation of voice-controlled systems, but that may become possible in the future based on current research trends. The review is concerned with sound-based attacks only, whilst recognising that attacks by sound are only a subset of the potential attacks that 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. Our review of attacks in prior and related work is organised according the mechanism of attack that they relate to. These mechanisms are plain speech, inaudible sound injection, adversarial learning, and active attack.

A. Plain Speech

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 [35], who describes a security vulnerability identified in Windows Vista that 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 that 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. [36] investigate possibilities for gaining unauthorised access to a smartphone via a malicious Android app that 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 that connects to a recording device, and then to disclose information, such as the victim's calendar schedule, by synthesised speech that 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.

B. Inaudible Sound Injection

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 [37]. 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 attacks using this mechanism 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. [38] also describe a 'silent' attack on smartphones via the voice command interface that enables an attacker to perform actions such as calling fee-paying phone numbers, posting to Facebook 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 attach a Raspberry Pi-based tool that is recognised by the phone as headphones with a microphone. Zhang et al. [39] and Song and Mittal [40] 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.

C. Adversarial Learning

There has also been some prior work towards using adversarial machine learning in attacks on voice-controlled digital assistants. Adversarial learning can be broadly defined as a process of identifying unexpected input that a machine learning-based system classifies 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 that an attacker can exploit to their advantage. 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. [41]). 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. [42] explain that processes of adversarial machine learning rely on identifying 'adversarial regions' in a classification category that 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 [43]). 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. [44]), but has recently also been applied to speech recognition. One example is the work presented by Vaidya et al. [30], who used audio mangling to distort commands issued to the 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. [45], who also proved the possibility of prompting Google Now to execute mangled commands that 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. [47], 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. [48] 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. [49] 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 [50] 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 degree of similarity to the original phrase, by making perturbations to the original recording that did not affect the original phrase as heard by humans. Schöenherr et al. demonstrate a similar type of attack on open-source speech recognition system Kaldi [51]. 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., Carlini and Wagner and Schöenherr et al. would be perceived by victims as ordinary speech and would therefore by more difficult to detect. Schöenherr et al. refer to this type of attack as "psychoacoustic hiding". To date, such work has

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

Paper	Attack Mechanism	Target Component	Human Perception of Attack
Dhanjani [35]	plain speech	speech interface in PC (Windows Vista)	standard voice command
Diao et al. [36]	plain speech	speech interface in voice-controlled digital assistant (Google Voice Search)	standard voice command
Kasmi and Esteves [37]	inaudible sound injection	voice capture in voice-controlled digital assistant (Google Now, Siri)	silence
Young et al. [38]	inaudible sound injection	voice capture in voice-controlled digital assistant (Siri)	silence
Zhang et al. [39]	inaudible sound injection	voice capture in voice-controlled digital assistant (Apple Siri, Amazon Alexa, Microsoft Cortana and others)	silence
Song and Mittal [40]	inaudible sound injection	voice capture in voice-controlled digital assistant (Google Now, Amazon Alexa)	silence
Vaidya et al. [30]	adversarial learning	speech recognition in voice-controlled digital assistant (Google Now)	white noise
Carlini et al. [45]	adversarial learning	speech recognition in voice-controlled digital assistant (Google Now) / speech recognition (CMU Sphinx)	white noise
Yuan et al. [46]	adversarial learning	speech recognition in speech transcription system (Kaldi)	music
Iter et al. [47]	adversarial learning	speech recognition in speech transcription system (WaveNet)	unrelated language
Cisse et al. [48]	adversarial learning	speech recognition in voice-controlled digital assistant (Google Voice)	unrelated language
Alzantot et al. [49]	adversarial learning	speech recognition in speech transcription system (TensorFlow)	unrelated language
Carlini and Wagner [50]	adversarial learning	speech recognition in speech transcription system (DeepSpeech) speech recognition in speech transcription system (DeepSpeech)	music unrelated language
Schöenherr et al. [51]	adversarial learning	speech recognition in speech transcription system (Kaldi)	unrelated language
Papernot et al. [52]	adversarial learning	natural language understanding in sentiment analysis system	nonsensical language
Liang et al. [53]	adversarial learning	natural language understanding in text classification system	unrelated language
Jia and Liang [54]	adversarial learning	natural language understanding in question answering system	unrelated language
Alzantot et al. [55]	adversarial learning	natural language understanding in sentiment analysis and textual entailment systems	unrelated language
Kuleshov et al. [56]	adversarial learning	natural language understanding in spam filtering, fake news detection and sentiment analysis systems	unrelated language
Li et al. [57]	adversarial learning	natural language understanding in sentiment analysis and toxic content detection systems	unrelated language
Bispham et al. [58]	adversarial learning	speech recognition in Google Assistant natural language understanding in Amazon Alexa Skills	nonsensical language unrelated language

been limited to speech-to-text transcription, i.e., it has not yet demonstrated mistranscription of voice commands capable of executing an action.

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. [46] 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. Another type of adversarial learning attack on speech recognition is presented by Bispham et al. [58], who present the results of work demonstrating a black-box attack in which voice commands to a target system are hidden in nonsensical word sounds that are perceived as meaningless by humans. One further, currently hypothetical, type of adversarial learning attack on speech recognition arises from the development of voice-controlled systems that are capable of interacting with users in more than one language (see for example Lopez-Moreno et al. [59]). It could be possible for attackers to identify instances where input in one language is misclassified by a system as a different input in another language. Depending on the language capabilities of the human listener, an adversarial learning attack prompting mistranscription of a utterance in one language as a different utterance in another language would be perceived by the human listener either as unrelated speech, or else as nonsensical or unintelligible speech.

Adversarial learning has also recently been applied to some areas of natural language understanding. This work has been performed mainly outside the context of voice-controlled systems, although there has been some preliminary work on attacks targeting natural language understanding in voice-controlled digital assistants, as discussed below. 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. [52]), text classification (see Liang et al. [53]), and question answering (see Jia and Liang [54]). Papernot et al. [52] use the forward derivative method, a white-box adversarial learning method, to identify word substitutions that 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. Whereas the altered sentence will appear unnatural to a human, the target system is not capable

of identifying the nonsensical nature of the adversarial input.

Papernot et al. state that the lack of naturalness of the adversarial examples in their attacks on natural language understanding will need to be addressed in future work. By contrast to Papernot et al., Liang et al. [53] 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 that 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 [54] 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 that 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.

Kuleshov et al. [56] use a word replacement approach in an adversarial learning attack targeting spam filtering, fake news detection and sentiment analysis. Their attack selects acceptable replacement words according to a semantic similarity measure based on 'thought vectors' in the form of averages of individual word vectors, and a syntactic similarity measure based on a language model, with the stated aim of 'formalising' the process of generating adversarial examples in natural language classification. The attacks demonstrated by Kuleshov et al. are white-box attacks, in that they rely on knowledge of objective function calculations in order to optimise the attack. Li et al. [57] demonstrate an attack on sentiment analysis and toxic content detection systems under both white-box and black-box conditions, using different types of perturbation of text including deliberate misspellings as well as word replacement. They note that character-level perturbations have a higher success rate in generating adversarial examples than word-level perturbations. Whilst all of the attacks on natural language understanding described above are demonstrated outside the context of voice-controlled systems, Bispham et al. [58] present a proof-of-concept study for attacks targeting natural language understanding in a voice-controlled digital assistant, using third-party Skills for Amazon Alexa as a specific example of a target system. The attack concept developed in the proof-of-concept study involves word replacement in a target command, as well as the transplant of content words from a target command to another meaning context where they are used in a different sense. These processes are shown to generate adversarial utterances that trigger target actions in a dummy Alexa Skill, whilst appearing to humans to have an unrelated meaning. The examples of attacks on natural language understanding described here are indicative of the fact that natural language understanding technology currently represents only a crude approximation of human language understanding that is easily destabilised.

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 for adversarial learning attacks on natural language understanding, in that unlike in the case of adversarial learning attacks targeting speech recognition, it is difficult to incorporate a device's wake-up word as part of an attack based on confusion of meaning. However, given the known presence of false positives with respect to wake-up word recognition, this type of attack should not be dismissed as impossible.

D. Active Attack

All of the attacks described in the prior and related work summarised above are 'passive' attacks, in the sense that they seek to exploit vulnerabilities that are already present in a target system. There is also the possibility of 'active' attacks that seek to undermine the functionality of the system itself. Miller et al. [60] refer to these attack types as 'foiling' and 'tampering', respectively. An example of active attack on a natural language interface was seen in an attempt by Microsoft to launch a social media chatbot named Tay. Tay was intended to learn human-like language use from interactions with humans on social media platform Twitter. Within a short time of launching the chatbot had to be closed down on account of having been flooded by some users with offensive language and views, which it then proceeded to imitate (see Følstad and Brandtstæd [61]). In the context of cloud-based voice assistants, active attacks might involve manipulating the response behaviour of the system for malicious ends. Rather than passively exploiting weaknesses in the speech recognition and natural language understanding functionalities of a voice-controlled system, such attacks would seek actively to undermine the system's ability to respond appropriately to spoken input by manipulating the dialogue management functionality. The potential for active attacks on voice-controlled digital assistants arises from the aim of providers of such systems to enable cloud-based assistants to continually 'improve' in interactions with their users. The capacity of voice-controlled digital assistants to learn from feedback from user conversations can be expected to increase with the introduction of commercially available voice assistants based on reinforcement learning. This capacity for learning might be abused by attackers aiming to confuse the system using various means, such as inconsistent verbal inputs over time, incongruous feedback in dialogue turns, or inappropriate corrections of a target system's responses. Attackers might for example launch a denial of service-type attack by mass disconfirmation of legitimate commands. Such attacks remain hypothetical at time of writing, as the current generation of voice-controlled digital assistants still use rule-based rather than reinforcement learning-based dialogue management technology, as explained above. However, this type of attack may become significant in future.

A different type of active attack affecting human interaction with voice-controlled systems in future might arise from the voice-controlled systems themselves, via the evolution of machine-generated languages that diverge from human language use. Whilst mismatches between human and machine understanding of natural language have generally been viewed as failure on the part of machines to attain human levels of language understanding, it is also possible to view such mismatches as a failure on the part of humans to grasp

the way in which meaning is represented by a machine. This was illustrated by an instance in which two bots were observed to develop a language for communication between themselves that was unintelligible to humans. This occurred as an unintended consequence of research by Lewis et al. [62], the aim of which was to train two bots to negotiate with one another in natural language using reinforcement learning. In the course of the learning process, the bots began to deviate from natural English in their language use, instead using apparently nonsensical strings of words in their communication with each other. This deviation was presumed to have effected more efficient communication between the two bots in achieving an optimal outcome in their negotiations. The development of bots capable of autonomously evading human language understanding may represent an increasingly significant future security threat, given the potential for loss of control over the behaviour of such systems by their human users. A malicious actor might be able to trigger a machine-machine reinforcement learning process in a target system with the specific aim of prompting it to behave in a way that was unintended by its human developers.

IV. Taxonomy of Attacks via the Speech Interface

Reflecting on the review of prior work and related work in Section III, we propose a high-level taxonomy of categories of attacks via the speech interface. This taxonomy is presented in Figure 3. The principle behind the taxonomy is to identify the various categories of non-speech and speech sounds that humans are capable of perceiving, and to group attacks via the speech interface according to these categories, rather than according to the attack mechanism used by an attack or by the specific technical vulnerability that it exploits. The last column of Table I shows the perceptual category that might be allocated to the attacks described above by humans. By applying this categorisation principle, our taxonomy is capable of encompassing attack types that have been shown to be possible in relation to the current generation of voice-controlled systems, as well as attacks that may become possible in future as the state-of-the-art in voice control advances. Thus our 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, and might be delivered via any speaker-enabled device capable of producing sound in the target system's environment.

In the taxonomy, attacks via the speech interface 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 that 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 that make illicit use of the intended functionalities of a speech dialogue system, and attacks that exploit unintended functionalities. Overt attacks use plain speech to exploit an inherent vulnerability in voice-controlled systems that arises from the difficulty of controlling access to a system via the 'speech space'. Covert attacks exploit gaps in the processes of capturing human speech or of translating the captured speech input into computer executable actions in a voice-controlled system. Covert attacks include attacks using inaudible sound injection, adversarial learning, and active attack, as discussed above.

Within the two primary categories of overt and covert attacks, attacks are grouped hierarchically into six final sub-categories based on human perceptual categories, as shown in Figure 3 and explained further below. Malicious inputs in overt attacks consist by definition of ordinary speech. Thus a single sub-category of 'plain-speech' attacks was identified for overt attacks. The attacks demonstrated in prior work using standard voice commands, such as those demonstrated by Dhanjani et al. [35] discussed above, fall into this sub-category. Malicious inputs in covert attacks may include input that consists in human terms of silence, as for example in the attacks demonstrated by Zhang et al. [39], noise, as for example in the attacks demonstrated by Carlini et al. [45], music, as for example in the attacks demonstrated by Yuan et al. [46], nonsense, as for example in the attacks on Google Assistant hiding malicious commands in nonsensical word sounds demonstrated by Bispham et al. [58], and unrelated speech, as for example in the attacks demonstrated by Carlini and Wagner [50]. Based on these examples of attacks in prior work, and in accordance with the categorisation principle chosen for the taxonomy of grouping attacks according to the nature of attacks as they might be perceived by a human listener, five sub-categories of covert attacks via a speech interface were identified, namely attacks consisting of silence, music, noise, 'nonsense', and 'missense'. Nonsense as a malicious input in covert attacks is defined as input that is made up of words or sounds that are in legitimate use in the relevant language, but that combines them in such a way that they do not convey any meaning in terms of human understanding. Missense is defined as unrelated speech that is misheard or misinterpreted by the target system as a target command.

Our taxonomy accords with established criteria for attack taxonomies, as described for example in Hansman and Hunt [63]. 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. 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 sub-categories. The taxonomy is also unambiguous, in that it is not possible to allocate the same voice attack to more than one of the six final sub-categories.

At the bottom of Figure 3, the attack categories based on human perceptual distinctions as identified in the taxonomy are aligned to the technical vulnerabilities in the architecture of the current generation of voice-controlled systems that might be targeted by each type of attack. The taxonomy of attacks categorised according to human perception as aligned to technical vulnerabilities at various points of the handling of speech input by voice-controlled systems represents the entire attack surface presented by a speech interface. To the extent that speech processing by voice-controlled systems mimics human speech processing, the attack categories in the taxonomy based on human perception correspond to vulnerabilities in the parts

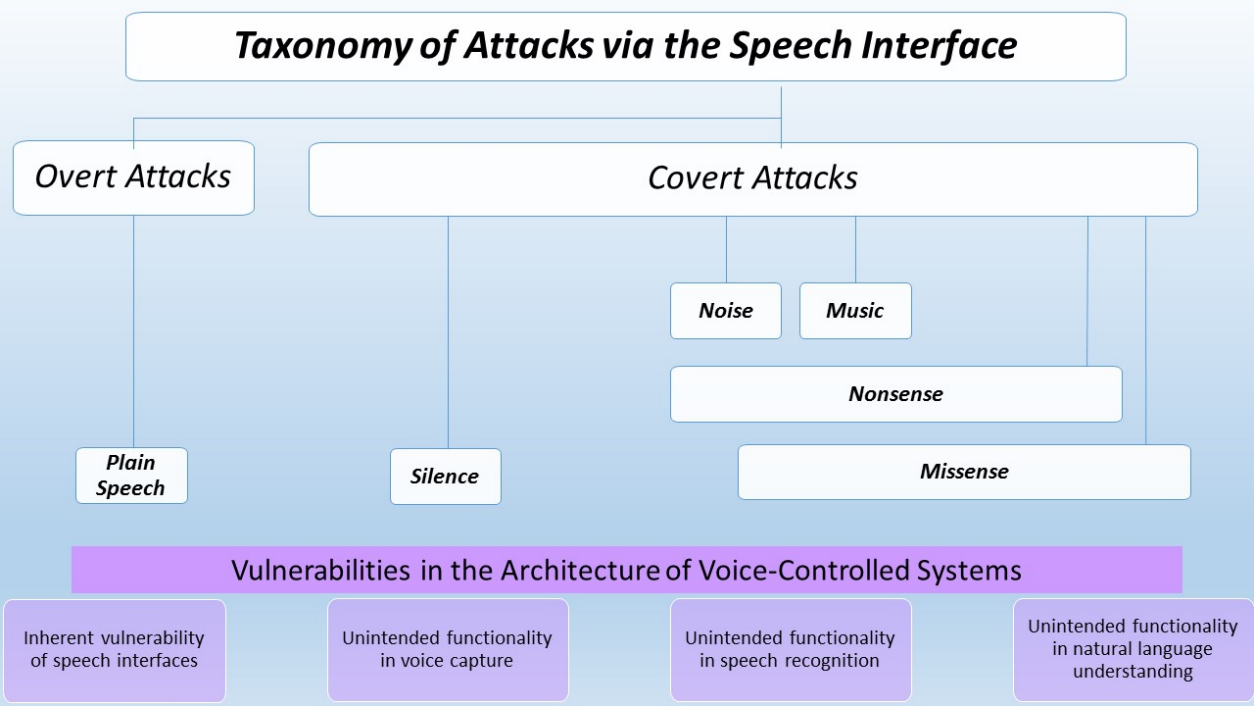

Figure 3. Taxonomy of Attacks via the Speech Interface aligned to Vulnerabilities in the Architecture of Voice-Controlled Systems

of the architecture of voice-controlled systems that represent equivalent human processes, although this correspondence is not exact. The alignment presented in Figure 3 covers the technical vulnerabilities that are present in the current generation of voice-controlled digital assistants, namely the vulnerability arising from the inherent difficulty of controlling access to a system by sound, vulnerabilities in the voice capture process, vulnerabilities in speech recognition, and vulnerabilities in natural language understanding. Whilst the categories of attack based on human perception can be expected to remain stable over time, their alignment to vulnerabilities in the architecture of voice-controlled systems might be expected to shift in future to include new vulnerabilities as the state-of-the-art in voice-controlled systems progresses. Thus, for example, missense attacks might be aligned in future not only to vulnerabilities in the speech recognition and natural language understanding components of voice-controlled systems, but also to vulnerabilities in the dialogue management component, such as the vulnerability presented by the potential for mistraining in the context of dialogue management functionality based on reinforcement learning, as well as the vulnerability presented by the potential for the evolution in reinforcement learning-based systems of bot-generated language that is incomprehensible to humans, as discussed above.

As reflected in the alignment in Figure 3, attacks in plain-speech exploit the inherent vulnerability of speech interfaces on account of the difficulty of controlling access to a system by sound. Attacks in silence attacks exploit vulnerabilities in the voice capture process, as is shown by the alignment of silent attacks to the voice capture component of the architecture in Figure 3. Attacks that use music and noise as malicious input exploit unintended functionality in speech recognition, as is shown by the alignment of these attack categories to the speech recognition component of the architecture. As further reflected in the alignment in Figure 3, nonsense attacks on current voice-controlled systems might be targeted either at the speech recognition or the natural language understanding components of a target system. The attacks in which malicious voice commands were hidden in nonsensical word sounds demonstrated by Bispham et al. [58] can be categorised as nonsense attacks targeting the speech recognition level of handling of speech input in a voice-controlled system. As regards attacks targeting the natural language understanding level, nonsense attacks have yet to be demonstrated with respect to voice-controlled systems directly, although there has been some related work that could be described as nonsense attacks on natural language understanding, such as in the attacks on a sentiment analysis system by Papernot et al. [52] by making nonsensical alterations to text discussed in Section III. Similar attacks might be demonstrated in the context of voice-controlled digital assistants in future.

Similar to nonsense attacks, missense attacks might also be targeted at either the speech recognition or natural language understanding component of current voice-controlled systems, as is also shown in the alignment in Figure 3. Missense attacks targeting speech recognition rely on mistranscription of adversarial input by a target system as a target command. In a missense attack that targets natural language understanding functionality, on the other hand, words might be transcribed correctly by the target system, but their meaning would be misinterpreted. This type of missense attacks would 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. Prior work on missense attacks in voice-controlled systems has to date been focussed primarily on attacks on

speech recognition as incorporated in such systems, as for example in the work of Carlini and Wagner [50]. The attacks described in the proof-of-concept study presented by Bispham et al. [58] for attacks that trigger a target command in an Amazon Alexa Skill using unrelated utterances would fall into the category of missense attacks targeting natural language understanding. There has also been more extensive work demonstrating missense attacks that target natural language understanding functionality in related research areas, such the attacks on question answering by Jia and Liang [54] by making apparently inconsequential alterations to text, or in the work by Kuleshov et al. [56] using word replacement to mislead spam filtering, toxic content detection and sentiment analysis systems, as described in Section III.

As discussed above, attacks on future systems may also include attacks targeting speech recognition in multilingual systems, prompting a target system to mistranscribe input in one language as different input in another. Such attacks would be classed as either nonsense or missense attacks, based on whether or not the cover language used by an attacker was understood by a human listener. As also discussed above, future attacks might further include attacks in which a target system's ability to respond appropriately to spoken input is actively undermined by mistraining of a dialogue management component based on reinforcement learning, as well as attacks that are based on facilitating the evolution of human-incomprehensible languages in autonomous bot-to-bot interactions in reinforcement learning-based systems. The former type of attack would represent a missense attack, with the adversarial input being perceived by human listeners as unrelated language, whereas the latter type of attack would represent a nonsense attack, with the adversarial input being perceived by human listeners as nonsensical language.

V. Conclusion and Future Work

This paper proposes a taxonomy of attacks via the speech interface that covers attacks investigated in prior and related work, as well as attacks that may be possible in the future. The review of prior and related 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 and missense attacks targeting the natural language understanding functionality of voice-controlled systems represent types of attacks that have yet to be explored fully in practice. Thus, such attacks should be a special focus of future work. Looking further into the future, attacks based on language confusion in multilingual systems, as well as attacks based on mistraining of dialogue management or facilitation of bot-generated languages in reinforcement learning-based systems, may become a reality requiring the attention of security researchers.

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. As a first step in this direction, Bispham et al. [64] present some attack and defence modelling work in which the attack categories in the taxonomy presented here are mapped to currently available defences against attacks via the speech interface, enabling an assessment of the effectiveness of current defences against the various types of attack.

Acknowledgment

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

References

[1] M. K. Bispham, I. Agrafiotis, and M. Goldsmith, "A taxonomy of attacks via the speech interface," Proceedings of Third International Conference on Cyber-Technologies and Cyber-Systems, 2018.

[2] "Why Amazon's Alexa may soon become your new colleague," 2017, URL: https://www.inc.com/emily-canal/amazon-alexa-for-business.html [accessed: 2019-05-05].

[3] 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.

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

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

[6] "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: 2019-05-05].

[7] "The Voice-AI Revolution is a Conversational Interface of Everything," 2017, URL: https://medium.com [accessed: 2019-05-05].

[8] "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].

[9] "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: 2019-05-05].

[10] "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: 2019-05-05].

[11] 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.

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

[13] "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: 2019-05-05].

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] M. McTear, "Conversational modelling for chatbots: Current approaches and future directions," Technical report, Ulster University, Ireland, Tech. Rep., 2018.

[25] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," Proceedings of the IEEE, vol. 101, no. 5, 2013, pp. 1160–1179.

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

[27] "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: 2019-05-05].

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] "Amazon Alexa heard and sent private chat," 2018, URL: https://www.bbc.co.uk/news/technology-44248122 [accessed: 2019-05-05].

[34] "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: 2019-05-05].

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

[36] 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.

[37] 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.

[38] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[51] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," arXiv preprint arXiv:1808.05665, 2018.

[52] 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.

[53] 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.

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

[55] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," arXiv preprint arXiv:1804.07998, 2018.

[56] V. Kuleshov, S. Thakoor, T. Lau, and S. Ermon, "Adversarial examples for natural language classification problems," 2018.

[57] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," arXiv preprint arXiv:1812.05271, 2018.

[58] M. K. Bispham, I. Agrafiotis, and M. Goldsmith, "Nonsense attacks on Google Assistant and missense attacks on Amazon Alexa," Proceedings of International Conference on Information Systems Security and Privacy, 2019.

[59] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 5337–5341.

[60] D. J. Miller, X. Hu, Z. Qiu, and G. Kesidis, "Adversarial learning: a critical review and active learning study," arXiv preprint arXiv:1705.09823, 2017.

[61] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of HCI," interactions, vol. 24, no. 4, 2017, pp. 38–42.

[62] M. Lewis, D. Yarats, Y. N. Dauphin, D. Parikh, and D. Batra, "Deal or no deal? End-to-end learning for negotiation dialogues," arXiv preprint arXiv:1706.05125, 2017.

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

[64] M. K. Bispham, I. Agrafiotis, and M. Goldsmith, "Attack and defence modelling for attacks via the speech interface," Proceedings of International Conference on Information Systems Security and Privacy, 2019.