Metadata of the chapter that will be visualized in SpringerLink

Book Title	Information Systems Security and Privacy	
Series Title		
Chapter Title	The Security of the Speech Interface: A Modelling Framework and Proposals for New Defence Mechanisms	
Copyright Year	2020	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	Bispham
	Particle	
	Given Name	Mary K.
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	University of Oxford
	Address	Oxford, OX1 3QD, UK
	Email	mary.bispham@cs.ox.ac.uk
Author	Family Name	Agrafiotis
	Particle	
	Given Name	Ioannis
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	University of Oxford
	Address	Oxford, OX1 3QD, UK
	Email	ioannis.agrafiotis@cs.ox.ac.uk
Author	Family Name	Goldsmith
	Particle	
	Given Name	Michael
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	University of Oxford
	Address	Oxford, OX1 3QD, UK
	Email	michael.goldsmith@cs.ox.ac.uk
Abstract	This paper presents an attack and defence modelling framework for conceptualising the security of the speech interface. The modelling framework is based on the Observe-Orient-Decide-Act (OODA) loop model, which has been used to analyse adversarial interactions in a number of other areas. We map the different types of attacks that may be executed via the speech interface to the modelling framework, and present a critical analysis of the currently available defences for countering such attacks, with reference to	

the modelling framework. The paper then presents proposals for the development of new defence mechanisms that are grounded in the critical analysis of current defences. These proposals envisage a defence capability that would enable voice-controlled systems to detect potential attacks as part of their dialogue management functionality. In accordance with this high-level defence concept, the paper presents two specific proposals for defence mechanisms to be implemented as part of dialogue management functionality to counter attacks that exploit unintended functionality in speech recognition functionality and natural language understanding functionality. These defence mechanisms are based on the novel application of two existing technologies for security purposes. The specific proposals include the results of two feasibility tests that investigate the effectiveness of the proposed mechanisms in defending against the relevant type of attack.

Keywords

Cyber security - Speech interface - Human-computer interaction

The Security of the Speech Interface: A Modelling Framework and Proposals for New Defence Mechanisms

Mary K. Bispham^(✉), Ioannis Agraftotis, and Michael Goldsmith

Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK
{mary.bispham,ioannis.agrafiotis,michael.goldsmith}@cs.ox.ac.uk

[AQ1]

Abstract. This paper presents an attack and defence modelling framework for conceptualising the security of the speech interface. The modelling framework is based on the Observe-Orient-Decide-Act (OODA) loop model, which has been used to analyse adversarial interactions in a number of other areas. We map the different types of attacks that may be executed via the speech interface to the modelling framework, and present a critical analysis of the currently available defences for countering such attacks, with reference to the modelling framework. The paper then presents proposals for the development of new defence mechanisms that are grounded in the critical analysis of current defences. These proposals envisage a defence capability that would enable voice-controlled systems to detect potential attacks as part of their dialogue management functionality. In accordance with this high-level defence concept, the paper presents two specific proposals for defence mechanisms to be implemented as part of dialogue management functionality to counter attacks that exploit unintended functionality in speech recognition functionality and natural language understanding functionality. These defence mechanisms are based on the novel application of two existing technologies for security purposes. The specific proposals include the results of two feasibility tests that investigate the effectiveness of the proposed mechanisms in defending against the relevant type of attack.

Keywords: Cyber security · Speech interface · Human-computer interaction

1 Introduction

Voice control is becoming an increasingly mainstream modality of human-computer interaction, particularly with respect to the growing popularity of

Supported by a doctoral training grant from the Engineering and Physical Sciences Research Council (EPSRC).

© Springer Nature Switzerland AG 2020
P. Mori et al. (Eds.): ICISSP 2019, CCIS 1221, pp. 1–29, 2020.
https://doi.org/10.1007/978-3-030-49443-8_14

smart speakers such as Amazon Alexa.¹ Voice-controlled systems are being used to perform both virtual actions, such as diary management, and cyber-physical actions, such as controlling devices in a smart home. This new form of human-computer interaction has brought with it new security concerns with regard to attackers exploiting voice commands to perform nefarious actions. Prior research has demonstrated various types of attacks via the speech interface, including attacks in which malicious voice commands are hidden in some form of cover medium so as to make them imperceptible to the legitimate users of voice-controlled devices [14, 58]. Whilst there has been a significant amount of prior work demonstrating the various types of attacks that may be executed via a speech interface to gain control of a victim's system, there have been few attempts to conceptualise the security of the speech interface in a comprehensive framework. This paper provides such a framework, using the Observe-Orient-Decide-Act (OODA) loop model that has been used to model adversarial interactions in many contexts.

This modelling framework facilitates a critical analysis of the effectiveness of currently available defences to counter the various types of attacks that can be executed via a speech interface. This critical analysis concludes that attacks that are imperceptible as such by human listeners are particularly dangerous in their potential effects, because of various difficulties in defending against such attacks using currently available defences. Such attacks are made possible by the existence of unintended functionality at various stages of handling of speech input. In accordance with this conclusion of the analysis, the paper further presents proposals for the development of new defence mechanisms to protect voice-controlled systems against attacks that exploit gaps between human and machine perceptions of speech and natural language. These proposals involve the implementation of defence mechanisms at the dialogue management stage of handling of speech input, which would enable a voice-controlled system to block execution of an attack at the dialogue management stage and instead issue a verbal security alert to users via its speech synthesis functionality. This is in contrast to currently available defence mechanisms, which are applied at the voice capture, speech recognition or natural language understanding stages of speech input handling. In support of this high-level defence concept, the paper makes two specific proposals for defence mechanisms to counter attacks exploiting unintended functionality in speech recognition and natural language understanding respectively. Specifically, we propose the novel application of an existing speech recognition system and of machine translation for security purposes in voice-controlled systems.

The remainder of this paper is structured as follows. Section 2 presents some background on human-computer interaction by speech, as well as an overview of prior work on the security of the speech interface. Section 3 maps the various types of attacks via the speech interface described in Sect. 2 to the OODA loop

¹ A recent UK government survey, for example, reported that 8% of adults in the UK now own a smart speaker, see <https://gds.blog.gov.uk/2018/08/23/hey-gov-uk-what-are-you-doing-about-voice/>.

model, and reviews the defence measures currently available to counter such attacks, using the model as a framework. Section 4 presents our proposals for the development of new defence mechanisms, including a high-level concept for a defensive capability at the dialogue management stage of handling of speech input, as well as two specific proposals for defence mechanisms based on the novel application of existing technologies for security purposes. Section 5 concludes the paper and makes recommendations for future work.

This paper is an extended version of a previous paper in which our framework for modelling the security of the speech interface was first presented (see Bispham et al. [7]). The work presented here extends the work in the earlier paper with the proposals for the development of new defence mechanisms referred to above.

2 Background and Prior Work

2.1 Background on Human-Computer Interaction by Speech

The typical architecture of a voice-controlled system consists of a speech recognition component, a natural language understanding component, a dialogue management component, and a response generation component (see for example Lison and Meena [37]). An outline of the processing pipeline is shown in Fig. 1. Following capture of speech input by a microphone, the speech recognition component will transcribe a sequence of words from the captured speech signal, the natural language understanding component will extract from the sequence of words a representation of the user's intent, the dialogue management component will map the representation of user intent to an appropriate response action, and the response generation component will generate a verbal and/or non-verbal response to the user. Both the speech recognition and the natural language understanding components involve some form of machine learning, typically Deep Neural Networks (DNNs) combined with Hidden Markov Models (HMMs) for speech recognition, and Conditional Random Fields (CRFs) or Recurrent Neural Networks (RNNs) for natural language understanding (see for example McTear [41]).

The dialogue management component, as implemented in the current generation of voice-controlled digital assistants, typically maps the representation of user intent outputted by the natural language understanding component to an appropriate action based on a set of handcrafted rules, although there has been some research on the development of more sophisticated dialogue management capabilities based on reinforcement learning (see McTear). The response generation component executes the action determined by the dialogue management component, which might be a virtual action such as an update to a calendar entry or a cyber-physical action such as controlling a smart home device, and/or a verbal response to the user by speech synthesis.

In the current generation of voice-controlled digital assistants like Amazon Alexa and Google Assistant, speech recognition and natural language understanding are performed in the provider's cloud rather than on the user's local

device. These devices thus include an additional ‘wake-word’ functionality to trigger streaming of audio data from the user’s environment to the provider’s cloud (see for example Chung et al. [16]).

Fig. 1. Spoken dialogue systems architecture.

2.2 Prior Work on the Security of the Speech Interface

The speech interface represents a new type of attack surface for malicious actors seeking to gain unauthorised access to a system. As such it represents a new focus for security research, and various types of attacks via the speech interface have been demonstrated in prior work. Some of the attacks demonstrated in prior work use plain-speech voice commands, as might be spoken by the user themselves. As such attacks will be clearly audible by legitimate users of the target system, they rely on engineering a situation in which a user is not present with their device. An example of this type of attack is described by Dhanjani [18], who hypothesises an attack using plain-speech commands in an audio file which plays when a user is likely to be absent from their PC. Another example is demonstrated by Diao et al. [19], in the form an attack using plain-speech commands which are triggered by a malicious smartphone app during hours when a user can be expected to be asleep. By contrast to these examples, prior work has also demonstrated a number of attacks that are not audible by users of a voice-controlled system, even if they are present with their device. One of these attacks is the so-called ‘dolphin’ attack demonstrated by Zhang et al. [58]. The attack demonstrated by Zhang et al. shows that it is possible to hide malicious voice commands in audio signals that have a frequency above the range of human hearing. This attack exploits non-linearities in microphone technology that make it possible for the voice capture functionality of voice-controlled systems to be induced to accept non-audible input as a speech signal within human-audible range.

Another type of attack via the speech interface demonstrated in prior work is attacks using adversarial learning to exploit unintended functionality in the speech recognition component of the target system. Adversarial learning can be broadly defined as the process of identifying inputs to a target system that the system misclassifies in some way that is to an attacker’s advantage. In the context of voice-controlled systems, this might be done by crafting audio input that, although audible by legitimate users of a voice-controlled device, may not be recognised by them as a malicious voice command to their system. These attacks aim to mislead the target voice-controlled system to accept input that is out-of-scope of the system’s intended input space as a valid in-scope voice

command. One of the first examples of this type of attack demonstrated in prior work was presented by Carlini et al. [14]. They demonstrated that it was possible, by extracting from voice command recordings the core acoustic features used by speech recognition whilst removing other parts of the speech signal, to create audio input that was perceived by humans as white noise, but that was still recognised by the target system Google Now as a valid voice command. The attack demonstrated by Carlini et al. was a black-box attack requiring no inside knowledge of the target system, and was shown to be effective when played over the air to the Google Now assistant on a smartphone. Bispham et al. [8] demonstrate another type of attack in which malicious voice commands are masked in nonsensical word sounds that rhyme with words of a target command. The authors show that the nonsensical word sounds are recognised by Google Assistant as a valid command, whilst human listeners do not detect the target command in the nonsensical word sounds when hearing them out of context. The attack presented by Bispham et al. is also a black-box attack that is shown to be capable of being executed over-the-air.

A further attack targeting speech recognition is demonstrated by Carlini and Wagner [15], who show that it is possible, using a mathematical adversarial learning technique, to manipulate audio recordings of a spoken text so as to lead the recording to be transcribed by a speech transcription system as an entirely different text chosen by an attacker. Carlini and Wagner also demonstrate that it is possible using the same technique to hide target transcriptions in music recordings. Unlike the attack using white noise presented by Carlini et al. [14] and the attacks using nonsensical word sounds presented by Bispham et al. [8], the attacks demonstrated by Carlini and Wagner are demonstrated in relation to a separate speech transcription system rather than a voice-controlled system as such. Also unlike the other two attacks, these attacks are shown to be effective only as audio file input to the target system rather than as over-the-air input via a microphone, and are white-box attacks requiring inside knowledge of the target system, rather than black-box attacks. As such the attacks presented by Carlini and Wagner might not be seen as representing a real threat at present, given that an attacker is unlikely to have access to the inner workings of a commercial system, and that attacks via audio file input will not be possible in the case of smart speakers that are only accessible by sound. However, Carlini and Wagner claim that their attacks will be capable of execution over the air in future, and research from adversarial learning in image recognition suggests that it might be possible to convert this attack to a black-box attack in future. With respect to the latter, Papernot et al. [43] demonstrate that it is possible to craft adversarial input to an image recognition system using a substitute system that is constructed based on outputs from a target system without requiring details of its inner workings. Such an approach might also be taken in adversarial learning attacks on speech recognition.

In addition to the adversarial learning attacks on speech recognition described above, there has also been some work towards adversarial learning attacks on natural language understanding, albeit that most of this work has been out-

side the context of voice-controlled systems. Papernot et al. [44] 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. For example, substituting the word ‘I’ for the word ‘excellent’ in an otherwise negative review is shown in the paper to lead to it being classified as having positive sentiment by the target system. Papernot et al. point out that semantically coherent adversarial examples for attacks on natural language understanding are difficult to achieve using purely mathematical adversarial learning approaches. 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 by humans as unnatural. Liang et al. [36] 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-indetectable 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 is not fully automated, but requires human judgement in finding and making changes to parts of the original input identified as significant for text classification by the Fast Gradient Sign Method.

Whilst the attacks on natural language understanding described above were demonstrated outside the context of voice control, Bispham et al. [8] have demonstrated attacks on natural language understanding in voice-controlled systems, using linguistic adversarial learning methods. They present the results of experimental work showing that it is possible to mislead natural language understanding in third-party applications for Amazon Alexa (known as Skills) by replacing words in target commands or by embedding homophones of target command words in a different sense context so as to create apparently unrelated utterances that are accepted by the system as the target command. In an extended version of the original paper, published in this volume, the authors demonstrate further instances of the latter type of attack on Amazon Alexa Skills as well as on open-source natural language understanding technology RASA NLU (Bispham et al. [9]). This type of attack based on embedding homophones of target command words in a different sense context is termed a ‘word transplant’ attack by the authors.

Bispham et al. [10] have developed a taxonomy of potential attacks via the speech interface that is organised according to the nature of the attack in terms of human perception. The taxonomy divides such attacks into two high-level categories: overt attacks, which aim to take control of a target system using plain-speech voice commands; and covert attacks, in which malicious voice commands are concealed in a cover medium so as to make them imperceptible to human listeners.

Overt attacks are easily detectable by users if they are consciously present with their device, therefore the success of an overt attack relies on a user being distracted or leaving their device unattended. An example of an overt attack might be the activation of a smartphone by a voice command that is delivered via a malicious app whilst a user is away from their device. Covert attacks are by definition not detectable by users, and can therefore be executed even if the user is present with their device. Examples of covert attacks include the high-frequency attacks that hide voice commands in sound that is inaudible to humans demonstrated by Zhang et al. [58], the attacks that hide voice commands via an audio-mangling process that makes them appear to humans as meaningless noise demonstrated by [14], and the targeted use of nonsensical word sounds that trigger target commands in a victim's system despite these word sounds being perceived as meaningless by human listeners, as demonstrated by Bispham et al. [8]. Covert attacks are divided within the taxonomy into five sub-categories namely silence, noise, music, nonsense and 'missense' (missense being defined as the hiding of malicious voice commands in speech that appears to be unrelated to the attacker's intent).

Overt attacks exploit the inherent vulnerability of speech interfaces on account of the difficulty of controlling access to such interfaces. Covert attacks exploit unintended functionality in the handling of speech input by a voice-controlled system that allows it to accept input that is not a valid voice command. The 'silent' attacks demonstrated by Zhang et al. [58] exploit nonlinearities in analog-to-digital conversion of speech signals by a microphone, whereas the attacks demonstrated by Carlini et al. [14] are an example of attacks exploiting vulnerabilities in speech recognition. The attacks on Amazon Alexa Skills targeting natural language understanding demonstrated by Bispham et al. [8] are the first examples of attacks on natural language understanding in voice-controlled systems. Regarding the dialogue management and response generation components, as these functionalities are fully dependent on input from the preceding components in the current generation of voice-controlled systems, there are no attacks targeting these functionalities at present.

An attacker's goal in executing an attack via a speech interface will be to gain control of one of the three generic types of action that can be performed via a voice-controlled digital assistant or other speech-controlled system using a sound-based attack. These three types of action are data extraction, data input and execution of a cyber-physical action. Specific attacks on each type of action that might be possible based on the current capabilities of voice-controlled digital assistants include prompting disclosure of personal information such as calendar information as envisaged by Diao et al. [19], instigating a reputational attack by posting to social media in the victim's name as envisaged by Young et al. [56], and causing psychological or physical harm to the victim by controlling a device in their smart home environment as envisaged by Dhanjani [18].

Attacks via a speech interface require a channel through which the sound-based attack is delivered, and in the case of attacks involving theft of information, successful execution also requires a channel for data exfiltration. Sound-based

attacks might be delivered through various channels, including natural voice, radio or TV broadcasts, audio files that users are induced to open via a weblink or email attachment, as suggested by Dhanjani [18], or malicious smartphone apps, as suggested by Diao et al. [19]. A further possible attack delivery channel is via an intermediary device under the attacker's control. Some instances of compromise of internet-connected speakers have been reported.² Speakers that have been compromised in this way could be used as an attack delivery channel for sound-based attacks on a target voice-controlled digital assistant within the speakers' vicinity. Regarding data exfiltration channels, Diao et al. [19] envisage for example that a system could be prompted to call a phone number linking to an audio recording device, which would then be used to record personal information of the victim that the system might be prompted to disclose by further voice commands.

Attacks via the speech interface have the potential to expand in time by perpetuating over a number of dialogue turns, as well as in space by spreading to other speech-controlled devices. Alepis and Patskakakis [3] and Petracca et al. [46] both mention the possibility of attacks by voice 'spreading' from one device to another by hijacking of a device's speech synthesis functionality. An example of an attack via the speech interface spreading through both space and time was seen in an instance in which a Google Home device was prompted to provide data to its user in synthesised speech that was perceived by a nearby Amazon Echo device as a command. This prompted the Echo to provide data that was in turn perceived by the Google Home as a command, the consequence being to set in motion an 'endless loop' between the two devices.³ This instance represented an example of an 'attack' that spread both in space to another device as well as in time over a potentially endless number of dialogue turns. Whilst this particular instance represents merely a humorous anecdote, it is possible that more malicious actions might be performed using similar mechanisms.

3 Attack and Defence Modelling Framework

There are a number of techniques for attack and defence modelling in cyber security. One of the more well-known modelling techniques for cyber security applications is the cyber kill-chain [2], which is used to analyse the different stages of malware attacks. Other established attack modelling techniques in cyber security include attack graphs [52] and attack grammars [45]. Another type of attack

² See Wired, 27th December 2017, "Hackers can rickroll thousands of Sonos and Bose speakers over the internet", <https://www.wired.com/story/hackers-can-rickroll-sonos-bose-speakers-over-internet/> and Trend Micro report 2017, "The Sound of a Targeted Attack", <https://documents.trendmicro.com/assets/pdf/The-Sound-of-a-Targeted-Attack.pdf>.

³ See UPROXX, 12th January 2017 "You Can Make Amazon Echo and Google Home Talk to Each Other Forever", <http://uproxx.com/technology/amazon-echo-google-home-infinity-loop/> and cnet.com 15th February 2018, "Make Siri, Alexa and Google Assistant talk in an infinite loop", <https://www.cnet.com/how-to/make-siri-alexa-and-google-assistant-talk-in-an-infinite-loop/>.

model is the OODA loop. Originally developed for the military context [11], the OODA loop has been applied in many different areas, including cyber defence [31]. The OODA loop method represents the behaviour of agents in adversarial interactions as a continuous cycle through a four-stage loop in a shared environment, the four stages of the loop being observation (Observe), orientation (Orientation), decision (Decide) and action (Act). The four stages of the loop as presented by Klein [31] are shown in Fig. 2. Rule [50] explains that the Observe and Act stages of the OODA loop are the points at which it makes contact with the external world, whereas the Orient and Decide stages are internal processes. Rule further explains that an adversary's aim as modelled by the OODA loop is to interfere with decision-making within their opponent's loop by presenting them with "ambiguous, deceptive or novel" situations, whilst at the same time continuing to execute their own loop independently.

Fig. 2. The four stages of the OODA Loop (from Klein [31]).

For the purposes of this work, the OODA loop was considered to be the most suitable modelling technique. The reason for this was that the OODA model is capable of capturing the cyclical nature of interactions by speech. Therefore the OODA loop model is especially suitable for representing the ways in which the processes of human-computer interaction by speech may be hijacked by adversarial actions. Specifically, the capture of the speech signal by a microphone prior to speech recognition can be mapped to the Observe stage of the OODA loop; the combined functionality of the speech recognition and natural language understanding components can be mapped to the Orient stage; the dialogue management (DM) component can be mapped to the Decide stage; and the response generation and speech synthesis stages can be mapped to the Act stage. Figure 3 shows a mapping of non-malicious user-device interactions via speech to the OODA loop model.

Figure 4 shows a mapping to the OODA loop model of the different types of attacks via the speech interface as categorised in the taxonomy presented by Bispham et al. [10], in which an attacker replaces a legitimate user in interactions with the device. The position of each type of attack in the loop model corresponds to the specific vulnerability exploited by the attack, i.e. the point at which the attacker gains control of the target device's loop. Plain-speech (overt) attacks and silent attacks are positioned at the Observe stage, as these types of attack exploit the inherent vulnerability of the speech interface and unintended functionality in voice capture, respectively. All other types of attack

Fig. 3. The OODA Loop in User-Device Interactions (from Bispham et al. [7]).

(noise, music, nonsense and missense) are positioned at the Orient stage, as these types of attack exploit unintended functionality in speech recognition or natural language understanding. The attack model also shows an attack delivery channel for transmission of malicious input by sound, and a data exfiltration channel that is used if the aim of the attack is the extraction of data. The model further indicates the potential expansion of an attack in time over several dialogue turns, as well as the possible expansion in space to a second target. The attacker may be any agent capable of producing sound in an environment shared with a target. In the case of attacks involving extraction of data, the agent will also be capable of recording sound in the shared environment.

Fig. 4. The OODA Loop in Attacker-Target Interactions (from Bispham et al. [7]).

Figure 5 shows a mapping to the OODA loop model of currently available defence measures. The position of each defence measure in the loop corresponds to the type of system vulnerability that the defence measure aims to patch. Cyber

security defence measures are often categorised as either preventive or reactive [39]. Preventive defence measures, such as authentication and access control, prevent malicious payloads from being inputted to a system at all, whereas reactive defence measures, such as anomaly-based or signature-based defences, detect that a malicious payload has been inputted and trigger a response to counteract the attack [22]. In terms of defence measures for human-computer interaction by speech as represented by the OODA loop model, preventive defences are defences that are applied at the voice capture stage of speech input handling, represented by the Observe stage of the loop, whereas reactive defences are applied as part of speech recognition or natural language understanding, represented by the Orient stage of the loop. The preventive measures mapped to the Observe stage of the loop are user presence, access control, audio-technical measures, and voice authentication. Reactive measures mapped to the Orient stage are confidence thresholds, input validation, signature-based defences, and anomaly-based defences. As dialogue management and response generation are fully controlled by input from the preceding components in the current generation of voice-controlled systems, there is currently no scope for additional defences at the Decide stage of the loop.

Fig. 5. Defences against attacks via the speech interface in the OODA Loop model (from Bispham et al. [7]).

User Presence. Overt attacks via the speech interface using plain-speech voice commands are easily detectable by users if they are consciously present with their device. Whilst the ability to detect an overt attack may not prevent such attacks from being successful to some extent, as the attack may already be in the process of being executing as the user detects it, the immediate detection of an attack by a user clearly limits the potential effects of the attack, in that the attack is likely to be easily attributable, and the user will be able to prevent any further propagation of the attack. Therefore it is advisable for users to take preventative measures to ensure that overt attacks cannot be executed on their device whilst they are not present with it. Jackson and Orebaugh [27] recommend some basic preventative measures including unplugging a voice-controlled

device when leaving the home and not placing a voice-controlled device close to doors and windows to prevent voice commands being inputted to the device from outside a house. User prevention measures such as these apply only to overt attacks and do not represent a defence against covert attacks that are imperceptible to humans and may therefore be executed notwithstanding the conscious presence of the user.

Access Control. Some work has been done on the potential for using formal access control methods to secure interactions via a speech interface and other types of cyber-physical interactions. Agadacos et al. [1] use formal methods to develop a scheme for identifying unintended interactions that may be possible between devices in a smart home environment over ‘hidden’ physical channels, including voice. Petracca et al. [46] propose a system of access controls to secure audio channels to and from a smartphone. The paper proposes an extension to the Android operating system in smartphones, with the objective of enforcing security policies for communications over three audio channels, namely between the device’s speakers and its microphone, between the device’s speakers and external parties, and between external parties and the device’s microphone. The authors concede that their access control system is based on the assumption of a reliable means of authenticating the legitimate user of a device, which may not be a valid assumption. Gong and Poellabauer [24] argue that the ‘Audroid’ method developed by Petracca et al. is not effective against adversarial learning attacks.

Audio-Technical Defences. Some defence measures have been presented that are applied at the voice capture stage of the handling of speech input by a voice-controlled device, prior the speech recognition and natural language understanding stages, so as to prevent ‘silent’ attacks that exploit non-linearity in microphone technology. As mentioned above, such attacks mislead a voice-controlled digital assistant or other voice-controlled device to execute commands that are concealed in high-frequency signals that are outside the human audible range, an example being the attack demonstrated by Zhang et al. [58] mentioned above. Roy et al. [49] present a defence against inaudible attacks based on signal forensics that involves software rather than hardware changes to microphone technology. The applicability of such defence measures is limited to attacks that exploit vulnerabilities in the voice capture functionality of voice-controlled digital assistants; such measures are not effective against attacks that exploit vulnerabilities in the speech recognition or natural language processing functionalities.

Voice Authentication. Biometric voice authentication, also known as speaker recognition, is perhaps the most obvious defence measure that might be implemented to prevent attacks on systems that are accessible via a speech interface. Hasan [26] details how voice biometric authentication is performed using a standard set of acoustic features. In theory, voice biometrics represent a potential solution to all types of attack via the speech interface by ensuring that a speech-controlled device acts only on voice commands from an authorised user. In prac-

tice, however, voice biometrics remain vulnerable to spoofing attacks, as stated by Wu et al. [55]. In an overview of the state-of-the-art in speaker recognition, Hansen and Hasan state that unlike in the case of other types of biometrics such as fingerprints, voice is subject to a certain amount of variability within the same individual as well between individuals, implying that some degree of potential for false positives in voice biometric authentication may be inevitable [25]. The potential for false positives is exploited by attackers in voice spoofing attacks.

Confidence Thresholds. Voice-controlled systems generally implement some form of confidence threshold to prevent them from accepting input that cannot be matched to one of their actions with sufficient certainty by the speech recognition or natural language understanding functionalities [30]. Whilst confidence thresholds are implemented as an error prevention measure rather than as a defence measure, they may have some defence functionality in preventing covert attacks via the speech interface, by enabling the system to reject malicious input that is not sufficiently similar to the examples of legitimate input that were used in training the system. However, a confidence threshold is unlikely to be sufficient to prevent all attacks. This was seen for example with respect to the attacks targeting speech recognition in Google Assistant using nonsensical word sounds demonstrated by Bispham et al. [8].

Input Validation. Aside from confidence thresholds, another approach to error prevention for voice-controlled systems has been to restrict in some way the vocabulary that will be recognised by the system as valid input. Controlled Natural Language (CNL) has been used to prevent misunderstandings between machines and humans as to the intended meaning of natural language input. CNL is a general term for various restricted versions of natural language that have been constructed with a restricted vocabulary and syntax in order to enable every sentence in the language to be mapped unambiguously to a computer-executable representation of its meaning [33]. Restricted language models like these have been developed particularly for contexts where avoiding misunderstandings is a critical concern, such as human-robot interactions in military applications [17]. Although primarily an error prevention rather than a security measure, CNL enables natural language input to be validated in the same way as is often done for security purposes in non-speech interfaces [51]. Thus CNL can be seen a defence mechanism against attacks via the speech interface which is implemented as part of the natural language understanding functionality. However, CNL is not likely to be effective in preventing attacks targeting speech recognition. Kaljurand and Alumäe [29] discuss the use of CNL in speech interfaces for smartphones. They point to the additional challenges in using CNL in a speech-based application as opposed to a text-based application, noting the need to avoid homophones within the CNL that can be distinguished in written but not in spoken language. The approach proposed by Kaljurand and Alumäe potentially addresses issues of confusability between user utterances that are within the intended scope of a speech-controlled system. However, it may not be effective in preventing confusion with out-of-vocabulary sounds that are directed

to the system by a malicious actor. Thus CNL is unlikely to present a solution to preventing covert attacks that target the speech recognition functionality of a voice-controlled system. Enforcement of a CNL in the design of a speech interface might be effective in preventing some attacks that exploit ambiguities in natural language input. However, such an approach would clearly be contrary to the aim of most providers of voice-controlled systems to enable users to communicate with their devices in as flexible and natural a way as possible [40].

Signature-Based Defences. A potential defence against some types of attacks via the speech interface is detection of attacks based on known attack signatures using supervised machine learning. Carlini et al. [14], for example, propose a machine learning-based defence to their own audio-mangling attack on speech recognition in Google Now, in the form of a machine learning classifier that distinguishes audio-mangled sentences from genuine commands based on acoustic features. Signature-based defences using linguistic features might similarly be used to detect malicious input targeting natural language understanding. Carlini et al. demonstrate that their signature-based classifier is effective against the specific attacks presented in their paper with 99.8% detection rate of attacks. However, the authors themselves note that such defences do not represent a proof of security, and are vulnerable to an ‘arms race’ with attackers who are likely simply to craft more sophisticated attacks to evade such defences. Attackers have the upper hand in such arms races with respect to machine learning based systems, on account of the vast number of possible inputs to such systems, making it impossible for defenders to prepare systems for all possible input in training.⁴

Anomaly-Based Defences. One possibility for enabling voice-controlled systems to become resistant to previously unseen attacks via the speech interface could be defence measures based on some form of anomaly detection. Anomaly detection-based defences have been applied in other areas of cyber security, such as network defence (see Rieck and Laskov, [48] and Bhuyan et al. [6]). However, anomaly-based defence measures depend on reliable similarity and distance measures in terms of which malicious input can be distinguished as anomalous relative to legitimate input (see Weller et al. [53]). In the context of attacks via the speech interface, such quantifiably measurable indications of suspicious activity may be difficult to identify. Whilst a number of both phonetic and semantic distance measures have been developed (see Pucher et al. [47] and Gomaa and Fahmy [23]), none of these are fully reliable in terms of their ability to separate sounds and meanings that are perceived as different by human listeners. Kong et al. [32] present the results of an evaluative study that indicated significant differences between error rates in human perception of speech sounds and their transcription by different types of automatic speech recognition in terms of a phonetic distance

⁴ See Cleverhans blog, 15th February 2017, “Is attacking machine learning easier than defending it?”, <http://www.cleverhans.io/security/privacy/ml/2017/02/15/why-attacking-machine-learning-is-easier-than-defending-it.html>.

measure. Budanitsky and Hirst [13] compare different measures of semantic distance with implied human judgements of word meaning via a task that involved detection of synthetically generated malapropisms, finding that none of these measures was capable of alignment with human understanding of word meaning. Thus such distance and similarity measures do not provide a reliable basis for an anomaly detection-based defence against attacks that seek to exploit differences between human and machine perceptions of speech, and may also prevent the system from accepting legitimate input.

4 Proposals for New Defence Mechanisms

This section presents proposals for the development of new defence mechanisms against attacks via the speech interface, grounded in the attack and defence modelling framework presented in the previous section. As concluded in the analysis presented in the previous section, none of the currently available defence mechanisms provide a full solution to prevention of attacks via the speech interface.

The proposals presented here focus on covert attacks. Overt attacks using plain-speech commands are excluded from the scope of the proposals on the basis that the risk of such attacks can be minimised through simple user precautions, such as not leaving devices in listening mode unattended, as discussed in the previous section. Whilst user presence does not remove the risk of overt attacks completely, in that by the time an attack is detected by a user it may already be in the process of being executed, immediate detection by a user implies that an overt attack will be easily attributable, and that any propagation of the attack can be prevented. Thus overt attacks can be considered far less pernicious than covert attacks of which execution can be hidden from user perception.

The scope of the proposals for defence mechanisms is further limited to attacks targeting unintended functionality in speech recognition or natural language understanding. ‘Silent’ attacks that target unintended functionality at the voice capture stage of speech input handling are excluded on the basis that these are likely to be preventable by audio-technical defences, as discussed in the previous section. By contrast, in the case of covert attacks targeting speech recognition and natural language understanding, none of the currently available defence mechanisms is capable of providing a full solution.

Examples of the attacks falling within the scope of the proposals for defence mechanisms include the attacks targeting speech recognition in Google Now demonstrated by Carlini et al. [14], and the attacks targeting natural language understanding in Amazon Alexa Skills demonstrated by Bispham et al. [8].

Our proposals for the development of novel defence mechanisms against such attacks are presented in three subsections below. The first subsection presents a high-level defence concept envisaging the implementation of defence mechanisms at the dialogue management stage of handling of speech input, separate from the core speech recognition and natural language understanding functionalities. The next two subsections present specific proposals for countering attacks exploiting unintended functionality in speech recognition and natural language

understanding respectively, based on novel applications of existing technologies for security purposes. The proposal for a defence mechanism to counter attacks targeting vulnerabilities in speech recognition is presented in the second subsection. We propose the implementation of a knowledge-based speech recognition system in voice-controlled systems, in addition to the core machine learning-based speech recognition functionality, as a security measure. The proposal for a defence mechanism to counter attacks targeting vulnerabilities in natural language understanding functionality is presented in the third subsection. We propose a security measure based on cross-lingual comparison of user intent determination, whereby, in addition to the interpretation by the natural language understanding functionality of a transcription of speech input in its original language, interpretation would also be performed on a machine translation of the transcription, to increase the robustness of the natural language understanding functionality to adversarial learning attacks.

The proposals for new defence mechanisms follow principles of so-called speculative design for ensuring robustness of proposals for the development of future technologies by embedding speculative proposals in a current context. This is done firstly by grounding the proposals for new defence mechanisms in the modelling framework for assessing the security of the speech interface presented above, and secondly by linking the proposals to existing technologies and research, as detailed further below. The principles of speculative design are described in detail by Auger [5].

4.1 High-Level Defence Concept

Our high-level defence concept envisages a defensive capability for voice-controlled systems that would be able to detect potential attacks targeting speech recognition or natural language understanding as part of its dialogue management functionality, and produce security alerts as part of its response generation functionality, using its own speech synthesis capability. In terms of the modelling framework presented in the previous section, the proposed defence concept represents a defence at the ‘Decide’ stage of the OODA loop, with the security alerts being generated at the ‘Act’ stage of the loop. This is by contrast to currently available defence mechanisms implemented at the ‘Orient’ stage of a target system’s OODA loop as part of the speech recognition and natural language understanding functionalities.

As explained in the previous section, as dialogue management functionality in current commercial systems responds passively to input from the preceding components, the dialogue management component is not capable of protecting the systems against attacks targeting vulnerabilities in the preceding speech recognition or natural language understanding components. A security-aware dialogue management functionality would conversely be capable of detecting malicious input that has been misrecognised by the speech recognition or natural language understanding functionality as legitimate input. Such a dialogue management functionality would block execution of the command that such

input purports to contain, and instead prompt the response generation component to generate an alert to the legitimate user that the system has received input from its environment purporting to be a voice command of which the legitimate user may be unaware. Such security alerts would thus form part of the system’s repertoire for interacting with its users. This type of functionality has previously been considered with regard to conversational agents developed specifically for security purposes. Security-aware digital assistants capable of detecting and reporting suspicious events have for example been developed to support analysts in security operation centres, two of these being the Artemis bot developed by Endgame⁵ and the Havyn bot developed by IBM.⁶ In the case considered here, a voice assistant would report on potential attacks on its own functionality as part of its interaction with users in spoken natural language. Figure 6 shows our concept of implementing a detection capability at the dialogue management stage and an alert capability at the response generation stage in the context of the OODA loop-based attack and defence modelling framework.

Fig. 6. High-level defence concept.

In order to be able to detect a covert attack and trigger an alert to the legitimate user as envisaged above, a voice-controlled system would need to have a reliable ability to identify input that purports to contain a valid voice command, but that might be perceived by a human as out-of-scope. At an abstract level, a solution to detection of covert attacks on voice-controlled systems that exploit gaps between human and machine perception of speech and language is simply to develop speech recognition and natural language understanding processes that are as similar as possible to human speech and language processing.

⁵ See Endgame blog 20th January 2017, ‘Endgame Announces Artemis: ‘Siri For Security’ To Transform SOC Operations’, <https://www.endgame.com/news/press-releases/endgame-announces-artemis-siri-security-transform-soc-operations>.

⁶ See Medium blog, 13th February 2013, ‘Havyn: a cognitive assistant for cybersecurity’, <https://medium.com/cognitivebusiness/havyn-a-cognitive-assistant-for-cybersecurity-e6580898f49e>.

Such improvements would prevent mismatches from being present in the first place. This aim is of course shared to some extent with the more general goal of improving the performance of voice-controlled systems by reducing error rates. However, performance objectives for machine speech recognition and natural language understanding do not require as complete an alignment to human capabilities as security objectives do. From a performance perspective, speech recognition and natural language understanding in a voice-controlled system needs to match human abilities only to the extent that it is able to correctly classify inputs that are within the system's intended scope, relying on the assumption that non-malicious users will not direct input to the system that is not within its scope. From a security perspective, however, the technology needs also to be capable of rejecting maliciously crafted out-of-scope input. The covert attacks via the speech interface considered in the defence proposals made here all represent instances of malicious use of out-of-scope input to attack a voice-controlled system. Securing a voice-controlled system against such attacks requires the system to be capable of matching human capabilities not only in correctly recognising and interpreting in-scope input, but also in identifying out-of-scope input.

Defending voice-controlled systems against malicious out-of-scope input is difficult to achieve as part of machine-learning-based speech recognition and natural language understanding, because the space of out-of-scope input for a given system is likely to be too vast to represent comprehensively in a training dataset to the granularity required to distinguish it from the space of valid speech input to the system. Whilst this is a problem common to all machine learning-based systems, in the context of speech recognition and natural language understanding a particular issue arises on account of the challenges posed by the variability of word sounds and the ambiguity of word meanings in natural language. Due to the presence of variability in word sounds and meaning, some degree of overlap between legitimate and malicious input to a speech interface is likely to be inevitable, at least in as far as they are separable using current techniques. This implies that any measures capable of preventing the speech-controlled system from accepting all malicious input would simultaneously lead it to reject some legitimate inputs, thus damaging the usability of the system. Thus, rather than relying on improvements in speech recognition and natural language understanding technology as such to ensure security of a speech interface, it is necessary to consider additional separate defence mechanisms at the dialogue management stage. Specific proposals for such defence mechanisms to counter attacks on speech recognition and natural language understanding respectively are made in the following subsections.

4.2 Defences Against Attacks on Speech Recognition

Attacks that target speech recognition functionality in voice-controlled systems exploit mismatches between human and machine perceptions of speech sounds. Such attacks include the attacks in noise demonstrated by Carlini et al. [14], the attacks based on nonsensical word sounds demonstrated by Bispham et al. [8], and the attacks in unrelated speech and music demonstrated by Carlini and

Wagner [15]. An effective defence against such attacks would involve removing the potential for mismatches between human and machine speech recognition in the system with respect to out-of-scope input.

One possibility for reducing the potential for mismatches between human and machine speech recognition with respect to out-of-scope input might be to incorporate knowledge from advances in research on human speech recognition in voice-controlled systems. The aim of this would be to filter out input that is unlikely to have been produced naturally by a legitimate human user of a device, and is therefore likely to represent some form of covert attack on the system. The current state-of-the-art in speech recognition technology is dominated by machine learning-based approaches trained with large amounts of speech data. However, from a security perspective, speech recognition technology based on knowledge of human speech recognition might be expected to be more successful than machine learning-based approaches in closing down the space of out-of-scope input in which an adversary can inject malicious voice commands without being detected by a human listener.

One speech recognition system based on knowledge of human speech recognition has in fact been developed. This speech recognition system, named FlexSR, is not in commercial use but is described in a patent application [34]. The system has been developed in the first instance as a language learning tool to help non-native learners of various languages to improve their pronunciation in a target language. Rather than mapping acoustic features such as Mel-frequency Cepstral Coefficients to phonemes in a given language directly, the FlexSR speech recognition technology uses a set of 18 phonological features common to all languages that are linked to the characteristics of the human vocal apparatus (see Lahiri and Reetz [35], Arora et al. [4]). Whereas phonemes are a set of sound units that are specific to a particular language, phonological features are universal, in the sense that all sounds used across all languages can be characterised as some combination of such features. The FlexSR system initially maps acoustic features in a speech sample to a vector that represents the presence or absence of each of the phonological features. The combination of phonological features extracted from the acoustic features is then mapped to phonemes of a given language in a second step, in order to identify spoken words. The aim of the approach to speech recognition incorporated in FlexSR is to achieve high levels of accuracy in speech recognition whilst also accounting for the high levels of variability in the pronunciation of words in natural speech. FlexSR uses the TIMIT system for transcription of phonemes [21].

A speech recognition system based on new insights into human word recognition, such as FlexSR, might have potential security applications in being able to prevent covert attacks targeting speech recognition functionality. This would be achieved by ensuring that input not recognised by humans as a valid pronunciation of a target command is not accepted by a voice-controlled system, whilst simultaneously preserving flexibility and thus usability of the system by allowing for variation in different human pronunciations of the same words. A system such

as FlexSR is unlikely to be vulnerable to the same adversarial learning input as the core machine learning-based speech recognition in a voice-controlled system.

Rather than replacing machine learning-based speech recognition in voice-controlled systems, a system such as FlexSR would be applied as a security measure at the dialogue management stage, to filter out input that has been transcribed by the speech recognition component as a voice command, but that is in reality unlikely to have been produced naturally by the human vocal system, and thus might represent a covert attack. The reason for using a speech recognition system such as FlexSR in addition to machine learning-based speech recognition, rather than to replace it, would be that machine learning-based speech recognition may achieve higher levels of accuracy with regard to in-scope input. Thus the two different speech recognition systems would be implemented in tandem as complementary functionalities, with the machine learning-based system being responsible for handling in-scope input, and the knowledge-based system being responsible for excluding out-of-scope input. Rather than performing speech recognition from scratch, a knowledge-based system such as FlexSR would be used to confirm whether a given audio input is likely to represent a genuine attempt to vocalise a voice command that has been transcribed by the core speech recognition system. A voice command identified by the core speech recognition functionality would be executed only if the knowledge-based system recognises the input as a legitimate human-comprehensible input. Such an approach would minimize the potential for covert attacks on speech recognition, by ensuring a closer alignment between machine and human perception of meaningful speech sounds, without affecting the performance of the system in terms of being able to handle variability in human pronunciations of the same words.

In order to investigate the potential defensive capability of FlexSR in a specific context, FlexSR's response to adversarial commands in noise, as demonstrated by Carlini et al. [14], was tested.⁷ Specifically, FlexSR's response was tested to five adversarial commands in noise that had been demonstrated by Carlini et al. in black-box attacks on Google Now, as made available by the researchers.⁸ The tests used an implementation of FlexSR as a language learning tool on a smartphone. In this implementation, FlexSR provides feedback on pronunciation by advising on how the expression of phonological features in a learner's speech should be changed in order to achieve correct pronunciation. If a learner's vocalisation of a word differs from the correct pronunciation in a way or to a degree that makes the target word unrecognisable, FlexSR may instead identify the insertion, deletion or substitution of a phoneme in the user's speech. Users are required to enter the word or phrase that they are attempting to pronounce, and then speak this word or phrase into the phone's microphone. FlexSR then compares phonological features extracted from the user's speech to the phonological features associated with correct articulation of phonemes in the

⁷ The authors are grateful to the University of Oxford's Faculty of Linguistics, Philology and Phonetics for providing access to the FlexSR system for the purposes of this work.

⁸ See <http://www.hiddenvoicecommands.com/black-box>.

target word or phrase in order to provide feedback. The example from Table 1 of correction of pronunciation of the word ‘Tweet’ as “Feature voice should increase for T” indicates that the phonological feature ‘voice’ in the phoneme transcribed as ‘T’ should be articulated more clearly to achieve correct pronunciation.

Details of FlexSR’s response to the adversarial commands in noise demonstrated by Carlini et al., as well as to a legitimate version of the corresponding target command in live human speech, are shown in Table 1. The table shows FlexSR’s judgement as to whether the adversarial and legitimate inputs represented a correct pronunciation of the target command, and, if not, how many features were identified as mispronounced, as well which specific features were identified as mispronounced in what respect. The most significant finding from these tests was that FlexSR did not recognise any of the adversarial commands in noise as correct pronunciations of the corresponding target command. In one case, FlexSR refused to accept the noise command as speech input altogether. This in itself was not sufficient to separate the adversarial commands from the legitimate target commands as spoken by a human, as FlexSR also identified three of the target commands in natural speech as incorrectly pronounced. However, with the exception of the adversarial command for the ‘Hey Google’ wake phrase, FlexSR did identify a much larger number of mispronounced features for the adversarial commands in noise than for target commands in natural speech for which incorrect pronunciation was identified. This suggests the number of features identified as mispronounced by FlexSR could be used as criteria for distinguishing legitimate ‘natural’ input from malicious ‘unnatural’ input in a bespoke implementation of FlexSR for security purposes. Any bespoke implementation of FlexSR for security purposes would need to address the issue of false positives that is evident in the feasibility tests using the implementation of FlexSR as a language learning tool.

The use of two different speech recognition systems with equivalent but different functionality as a security mechanism can be characterised as a cyber mimic defence. Cyber mimic defence involves using redundant alternative processing units of different but equivalent functionality to increase the robustness of a system to adversarial input. This approach has previously been applied in network defence to detect zero-day attacks (see for example Liu et al. [38]). The use of a second speech recognition system to increase the robustness of a voice-controlled system to attack has parallels to this approach.

4.3 Defences Against Attacks on Natural Language Understanding

Attacks that target natural language understanding functionality in voice-controlled systems exploit inadequacies in current methods for representing meaning in such systems. Examples of attacks of this type are the attacks on Amazon Alexa Skills demonstrated by Bispham et al. [8], in which it was shown to be possible to mislead a dummy Amazon Alexa Skill to accept an unrelated utterance as a target command. Current technologies for natural language understanding clearly represent only very crude approximations of the ‘true’ processes of natural language understanding in the human brain. Defences against attacks

Table 1. FlexSR noise attacks tests - Results.

Target command	human (non-adversarial) / noise (adversarial)	FlexSR correctly pronounced yes/no	FlexSR no. of features mispronounced	Details of FlexSR features mispronounced
What is my current location	human	NO	2	1) Feature voice should decrease for T in what; 2) Feature stop should increase for T in current
	noise	NO	5	1) Feature stop should increase for T; 2) Feature str should increase for z in is; 3) Feature nas should increase for N in current; 4) Feature cor should increase for T in current; 5) Feature nas should increase for N in location
Tweet goodbye	human	NO	1	Feature voice should increase for D
	noise	NO	4	1) Feature stop should increase for T; 2) Feature stop should increase for T; 3) Feature voice should increase for D; 4) Feature labial should increase for B
OK Google	human	NO	1	Feature RTR should increase for AX
	noise	NO	1	Feature cor should increase for EY
Turn on airplane mode	human	YES	0	none
	noise	NO	n.a.	<i>Not recognised by FlexSR as valid speech input</i>
Call 911	human	YES	0	none
	noise	NO	6	1) Feature nas should increase for N in nine; 2) Feature nas should increase for N in nine; 3) Feature cor should increase for N in one; 4) Feature rtr should increase for AH; 5) Feature nas should increase for N; 6) Insertion of NG at the end

targeting natural language understanding will need to close the gaps between human and machine understanding of the meaning of utterances that leave the system vulnerable to malicious exploitation.

In theory, similar to defences against attacks on speech recognition based on knowledge of human speech recognition as proposed above, defences against attacks on natural language understanding could be developed based on knowledge of human construction of meaning from spoken utterances, so as to mitigate the threat of attacks that exploit differences in human and machine understanding of natural language. In practice, however, much about the human processes for the construction of meaning remains unknown. At a high level, research from

neurolinguistics suggests that the human construction of meaning from sentences emerges from a complex interplay of word meanings and syntax, as shown for example by Fedorenko et al. [20] and Johnson and Goldberg [28]. Such studies confirm the compositional nature of meaning construction in the human brain in general, but do not shed light on this at a level of granularity that might elucidate the process of meaning construction in specific contexts. It is possible that future advances in neuroscience and linguistics may identify features of meaning representation in the human brain that capture the essence of the specific distinctions of meaning made in human natural language understanding, and that are capable of replication in machine natural language understanding. The discovery of such features would enable the natural language understanding of voice-controlled systems to become more closely aligned with human understanding both with respect to in-scope and out-of-scope input, thus minimising the potential for malicious exploitation of gaps between human understanding of spoken language and its imitation in artificial systems. However, this possibility remains futuristic and nebulous at present.

A different approach to defending against attacks targeting natural language understanding is suggested by the work of Navigli and Ponzetto [42] on an approach to word sense disambiguation using multilingual semantic networks. The basis of the approach to word sense disambiguation proposed by Navigli and Ponzetto is that the set of word senses associated with a given word is unlikely to remain constant across different languages. Different senses of the same word in one language are likely to be translated as different words in another language, and thus the translation of a word as used in a particular context can be used to determine the correct word sense. This idea might be adapted to detect covert attacks on natural language understanding in voice-controlled systems by translating utterances inputted to a voice-controlled system and comparing the interpretation of user intent from the utterance in different languages. In the case of a non-malicious command, the interpretation of user intent is likely to remain constant across languages. However, in the case of malicious input aiming to mislead natural language understanding, for example by crafted use of homophones, the interpretation of user intent is unlikely to remain constant in different languages. A defence mechanism based on cross-lingual comparison of natural language understanding outputs could be implemented at the dialogue management stage of handling of speech input by a voice-controlled system. In the event that the intent extracted from a user utterance changes in a different language, execution of the intent would be blocked.

The potential of such an approach to defend against attacks on natural language understanding in voice-controlled systems can be trivially demonstrated in the context of the ‘word transplant’ attacks on natural language understanding in Alexa Skills and RASA NLU using homophones of target command words demonstrated by Bispham et al. [9], using the readily available machine translation technology Google Translate.⁹ Google Translate uses RNNs for sequence-to-sequence mapping of input in one language to output in another language (see

⁹ See <https://translate.google.co.uk/>.

Wu et al. [54]). Table 2 shows the successful adversarial commands used in the word transplant attacks on natural language understanding in Alexa Skills and RASA NLU demonstrated by Bispham et al. and their translation by Google Translate into German. It is evident that, with just one exception, the number of content words shared between target command and adversarial command drops significantly in translation, and thus that the interpretation of user intent from adversarial commands is unlikely to remain constant across languages.

The effectiveness of this approach as a potential defence against word substitution attacks on natural language understanding, as also demonstrated by Bispham et al. [8] in relation to voice-controlled systems as well by other researchers in related areas, is difficult to test in the absence of a real multilingual voice-controlled system to use in testing. However, it is reasonable to speculate that this approach might at least minimise the effectiveness of such attacks, as it would be less likely for a word substitution attack to be effective across different languages than in just one language. The defence approach proposed here might also be used to defend against any attacks using non-grammatical and/or meaningless combinations of real words as a cover medium for covert attacks, i.e. nonsense attacks targeting natural language understanding in voice-controlled systems. Again this would be based on the supposition that a nonsensical string of words would be less likely to mislead natural language understanding in two languages than just one.

Similar to the use of two different speech recognition systems as a defence against attacks on speech recognition, performing natural language understanding with respect to two or more languages rather than just one in order to

Table 2. Feasibility test for defence against attacks on NLU based on cross-lingual comparison.

Target Command (original language)	Adversarial Command (original language)	no. of words shared between target and adversarial command (original language)	Target Command (translation)	Adversarial Command (translation)	no. of words retained between target and adversarial command (translation)
tell me the current balance	I kept my balance in the current	2	Sag mir das aktuelle Guthaben	Ich habe mein Gleichgewicht im Strom gehalten	0
show me all my transactions	the transactions were for show	2	Zeig mir alle meine Transaktionen	Die Transaktionen waren für die Show	1
pay a bill for electricity	bill of an anchor	1	Strom bezahlen	Rechnung eines Ankers	0
think my card is stolen	your card is an ace	1	Ich glaube meine Karte ist gestohlen	Ihre Karte ist ein Ass	1

increase the robustness of this process against malicious input has some commonality with the cyber mimic defence approach of using equivalent but different functionality as a security mechanism as described for example by Liu et al. [38].

5 Conclusion and Future Work

In this paper, we firstly use a modelling framework based on the OODA loop model to analyse the effectiveness of currently available defences against attacks via the speech interface. The analysis concludes that current defence measures are not adequate to prevent all types of attacks via the speech interface, particularly with respect to attacks that exploit gaps between human and machine perceptions of spoken language. In accordance with this conclusion, the paper further makes proposals for the development of new defence mechanisms against human-imperceptible attacks which exploit unintended functionality in the speech recognition and natural language understanding components of voice-controlled systems. These defence mechanisms would be implemented at the dialogue management stage of speech input handling. To counter human-imperceptible attacks targeting speech recognition, a defence mechanism is proposed based on the use of an alternative speech recognition system as a security measure, whereby the outputs of the core speech recognition functionality and the alternative speech recognition system would be compared at the dialogue management stage. A voice command would be executed only if the alternative speech recognition system confirms that it is legitimate human-generated speech. To counter human-imperceptible attacks targeting unintended functionality in natural language understanding, a defence mechanism is proposed based on cross-lingual comparison using machine translation, whereby the interpretation of an utterance by the natural language understanding functionality in the system's original language would be compared to its interpretation in translation to another language at dialogue management stage. A voice command would be executed only if the same user intent was identified in the translation of an utterance as in its original language. On detection of a potential attack targeting speech recognition or natural language understanding, the dialogue management component of the voice-controlled system would block execution of the command identified by the core speech recognition and natural language understanding components, and instead issue a verbal alert to the legitimate user of the device via its response generation and speech synthesis functionalities. Future work should focus on the further development and testing of the proposals for defence mechanisms outlined in this paper.

References

1. Agadakos, I., et al.: Jumping the air gap: modeling cyber-physical attack paths in the internet-of-things. In: *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy*, pp. 37–48 (2017)
2. Al-Mohannadi, H., Mirza, Q., Namanya, A., Awan, I., Cullen, A., Disso, J.: Cyber-attack modeling analysis techniques: an overview. In: *IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 69–76 (2016)
3. Alepis, E., Patsakis, C.: Monkey says, monkey does: security and privacy on voice assistants. *IEEE Access* **5**, 17841–17851 (2017)
4. Arora, V., Lahiri, A., Reetz, H.: Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *J. Acoust. Soc. Am.* **143**(1), 98–108 (2018)
5. Auger, J.: Speculative design: crafting the speculation. *Dig. Creativity* **24**(1), 11–35 (2013)
6. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: Network anomaly detection: methods, systems and tools. *IEEE Commun. Surv. Tutor.* **16**(1), 303–336 (2014)
7. Bispham, M.K., Agrafiotis, I., Goldsmith, M.: Attack and defence modelling for attacks via the speech interface. In: *Proceedings of International Conference on Information Systems Security and Privacy, ICISSP 2019*, pp. 519–527 (2019)
8. Bispham, M.K., Agrafiotis, I., Goldsmith, M.: Nonsense attacks on Google assistant and missense attacks on Amazon Alexa. In: *Proceedings of International Conference on Information Systems Security and Privacy, ICISSP 2019*, pp. 75–87 (2019)
9. Bispham, M.K., Janse van Rensburg, A., Agrafiotis, I., Goldsmith, M.: Black-box attacks via the speech interface using linguistically crafted input. In: Mori, P., et al. (eds.) *ICISSP 2019, CCIS 1221*, pp. xx–yy, revised and extended paper. Springer, Cham (2020)
10. Bispham, M.K., Agrafiotis, I., Goldsmith, M.: A taxonomy of attacks via the speech interface. In: *Proceedings of CYBER 2018: The Third International Conference on Cyber-Technologies and Cyber-Systems*, pp. 7–14 (2018)
11. Boyd, J.R.: The essence of winning and losing. *Unpublished Lecture Notes* **12**(23), 123–125 (1996)
12. Brehmer, B.: The dynamic OODA loop: Amalgamating Boyd’s OODA loop and the dynamic decision loop (2005)
13. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Resources*, vol. 2, p. 2 (2001)
14. Carlini, N., et al.: Hidden voice commands. In: *25th USENIX Security Symposium (USENIX Security 2016)*, Austin, TX (2016)
15. Carlini, N., Wagner, D.: Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint [arXiv:1801.01944](https://arxiv.org/abs/1801.01944)* (2018)
16. Chung, H., Park, J., Lee, S.: Digital forensic approaches for Amazon Alexa ecosystem. *Dig. Invest.* **22**, 15–25 (2017)
17. Ciesielski, A., Yeh, B., Gordge, K., Basescu, M., Tunstel, E.: Vocal human-robot interaction inspired by Battle Management Language. In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3379–3384 (2017)
18. Dhanjani, N.: *Abusing the Internet of Things: Blackouts, Freakouts, and Stakeouts*. O’Reilly Media Inc., Sebastopol (2015)

19. Diao, W., Liu, X., Zhou, Z., Zhang, K.: 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, pp. 63–74. ACM (2014)
20. Fedorenko, E., et al.: Neural correlate of the construction of sentence meaning. *Proc. Natl. Acad. Sci.* **113**(41), 6256–6262 (2016)
21. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: DARPA TIMIT acoustic-phonetic continuous speech corpus cd-rom. NASA STI/Recon technical report (1993)
22. Giraldo, J., Sarkar, E., Cardenas, A.A., Maniatakos, M., Kantarcioglu, M.: Security and privacy in cyber-physical systems: a survey of surveys. *IEEE Des. Test* **34**(4), 7–17 (2017)
23. Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. *Int. J. Comput. Appl.* **68**(13), 13–18 (2013)
24. Gong, Y., Poellabauer, C.: An overview of vulnerabilities of voice controlled systems. arXiv preprint [arXiv:1803.09156](https://arxiv.org/abs/1803.09156) (2018)
25. Hansen, J.H., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Process. Mag.* **32**(6), 74–99 (2015)
26. Hasan, M.R., Jamil, M., Rahman, M., et al.: Speaker identification using MEL frequency cepstral coefficients. *Variations* **1**(4) (2004)
27. Jackson, C., Orebaugh, A.: A study of security and privacy issues associated with the Amazon Echo. *Int. J. Internet Things Cyber-Assur.* **1**(1), 91–100 (2018)
28. Johnson, M.A., Goldberg, A.E.: Evidence for automatic accessing of constructional meaning: Jabberwocky sentences prime associated verbs. *Lang. Cognit. Process.* **28**(10), 1439–1452 (2013)
29. Kaljurand, K., Alumäe, T.: Controlled natural language in speech recognition based user interfaces. In: International Workshop on Controlled Natural Language, pp. 79–94 (2012)
30. Khan, O.Z., Sarikaya, R.: Making personal digital assistants aware of what they do not know. In: INTERSPEECH, pp. 1161–1165 (2016)
31. Klein, G., Tolle, J., Martini, P.: From detection to reaction-a holistic approach to cyber defense. In: Defense Science Research Conference and Expo (DSR) 2011, pp. 1–4. IEEE (2011)
32. Kong, X., Choi, J.-Y., Shattuck-Hufnagel, S.: Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5810–5814. IEEE (2017)
33. Kuhn, T.: A survey and classification of controlled natural languages. *Comput. Linguist.* **40**(1), 121–170 (2014)
34. Lahiri, A., Reetz, H., Roberts, P.: Method and apparatus for automatic speech recognition. US Patent App. 15/105,552 (2016)
35. Lahiri, A., Reetz, H.: Distinctive features: phonological underspecification in representation and processing. *J. Phonet.* **38**(1), 44–59 (2010)
36. Liang, B., Li, H., Su, M., Bian, P., Li, X., Shi, W.: Deep Text Classification Can be Fooled. arXiv preprint [arXiv:1704.08006](https://arxiv.org/abs/1704.08006) (2017)
37. Lison, P., Meena, R.: Spoken dialogue systems: the new frontier in human-computer interaction. *XRDS: Crossroads ACM Mag. Stud.* **21**(1), 46–51 (2014)
38. Liu, W., Chen, F., Hu, H., Cheng, G., Huo, S., Liang, H.: A novel framework for zero-day attacks detection and response with cyberspace mimic defense architecture. In: Proceedings of 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 50–53 (2017)

39. Loukas, G., Gan, D., Vuong, T.: A taxonomy of cyber attack and defence mechanisms for emergency management networks. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 534–539. IEEE (2013)
40. McShane, M., Blissett, K., Nirenburg, I.: Treating unexpected input in incremental semantic analysis. In: Proceedings of The Fifth Annual Conference on Advances in Cognitive Systems, Cognitive Systems Foundation, Palo Alto, CA (2017)
41. McTear, M., Callejas, Z., Griol, D.: The Conversational Interface. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-32967-3>
42. Navigli, R., Ponzetto, S.P.: Joining forces pays off: multilingual joint word sense disambiguation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1399–1410 (2012)
43. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint [arXiv:1602.02697](https://arxiv.org/abs/1602.02697) (2016)
44. Papernot, N., McDaniel, P., Swami, A., Harang, R.: Crafting adversarial input sequences for recurrent neural networks. In: Military Communications Conference, MILCOM 2016–2016 IEEE, pp. 49–54 (2016)
45. Patten, T., Call, C., Mitchell, D., Taylor, J., Lasser, S.: Defining the malice space with natural language processing techniques. In: Cybersecurity Symposium (CYBERSEC), pp. 44–50. IEEE (2016)
46. Petracca, G., Sun, Y., Jaeger, T., Atamli, A.: Audroid: preventing attacks on audio channels in mobile devices. In: Proceedings of the 31st Annual Computer Security Applications Conference, pp. 181–190. ACM (2015)
47. Pucher, M., Türk, A., Ajmera, J., Fecher, N.: Phonetic distance measures for speech recognition vocabulary and grammar optimization. In: 3rd Congress of the Alps Adria Acoustics Association, pp. 2–5 (2007)
48. Rieck, K., Laskov, P.: Detecting unknown network attacks using language models. In: Büschkes, R., Laskov, P. (eds.) DIMVA 2006. LNCS, vol. 4064, pp. 74–90. Springer, Heidelberg (2006). https://doi.org/10.1007/11790754_5
49. Roy, N., Shen, S., Hassanieh, H., Choudhury, R.R.: Inaudible voice commands: the long-range attack and defense. In: 15th USENIX Symposium on Networked Systems Design and Implementation NSDI 2018), pp. 547–560. USENIX Association (2018)
50. Rule, J.N.: A Symbiotic Relationship: The OODA Loop, Intuition, and Strategic Thought. US Army War College (2013)
51. Schneider, M.A., Wendland, M.-F., Hoffmann, A.: A negative input space complexity metric as selection criterion for fuzz testing. In: El-Fakih, K., Barlas, G., Yevtushenko, N. (eds.) ICTSS 2015. LNCS, vol. 9447, pp. 257–262. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25945-1_17
52. Janse van Rensburg, A., Nurse, J.R., Goldsmith, M.: Attacker-parametrised attack graphs. In: 10th International Conference on Emerging Security Information, Systems and Technologies (2016)
53. Weller-Fahy, D.J., Borghetti, B.J., Sodemann, A.A.: A survey of distance and similarity measures used within network intrusion anomaly detection. IEEE Commun. Surv. Tutor. **17**(1), 70–91 (2015)
54. Wu, Y., et al.: Google’s neural machine translation system: bridging the gap between human and machine translation. arXiv preprint [arXiv:1609.08144](https://arxiv.org/abs/1609.08144) (2016)

55. Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H.: Spoofing and countermeasures for speaker verification: a survey. *Speech Commun.* **66**, 130–153 (2015)
56. Young, P.J., Jin, J.H., Woo, S., Lee, D.H.: BadVoice: soundless voice-control replay attack on modern smartphones. In: 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 882–887. IEEE (2016)
57. Young, S., Gašić, M., Thomson, B., Williams, J.D.: POMDP-based statistical spoken dialog systems: a review. *Proc. IEEE* **101**(5), 1160–1179 (2013)
58. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W.: DolphinAttack: inaudible voice commands. arXiv preprint [arXiv:1708.09537](https://arxiv.org/abs/1708.09537) (2017)

Author Queries

Chapter 14

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Please note that for Reference [9] the page numbers (pp. xx-yy) will be updated after they are finalised.	