I. J. Computer Network and Information Security, **2022, x, xx-xx**
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

Securing Voice Communications Using Audio Steganography

Anthony Phipps
London Metropolitan University/Cyber Research Centre, London, N7 8DB, UK
E-mail: arp0264@my.londonmet.ac.uk, tsaphip1@londonmet.ac.uk

Karim Ouazzane and Vassil Vassilev
London Metropolitan University/Cyber Research Centre, London, N7 8DB, UK
E-mail: k.ouazzane@londonmet.ac.uk, v.vassilev@londonmet.ac.uk

Received: January 2022, Year; Accepted: Date Month, Year; Published: Date Month, Year

Abstract: Although authentication of users of digital voice-based systems has been addressed by much research and many commercially available products, there are very few that perform well in terms of both usability and security in the audio domain. In addition, the use of voice biometrics has been shown to have limitations and relatively poor performance when compared to other authentication methods. We propose using audio steganography as a method of placing authentication key material into sound, such that an authentication factor can be achieved within an audio channel to supplement other methods, thus providing a multi factor authentication opportunity that retains the usability associated with voice channels. In this research we outline the challenges and threats to audio and voice-based systems in the form of an original threat model focusing on audio and voice-based systems, we outline a novel architectural model that utilises audio steganography to mitigate the threats in various authentication scenarios and finally, we conduct experimentation into hiding authentication materials into an audible sound. The experimentation focused on creating and testing a new steganographic technique which is robust to noise, resilient to steganalysis and has sufficient capacity to hold cryptographic material such as a 2048 bit RSA key in a short audio music clip of just a few seconds achieving a signal to noise ratio of over 70 dB in some scenarios. The method developed was seen to be very robust using digital transmission which has applications beyond this research. With acoustic transmission, despite the progress demonstrated in this research some challenges remain to ensure the approach achieves its full potential in noisy real-world applications and therefore the future research direction required is outlined and discussed.

Index Terms: **Cyber Security, Audio Security, Steganography, User Experience, Accessibility**

1. Introduction

Imagine a world where you could use any one of your trusted personal devices that has a speaker to help strongly authenticate you by playing a simple and pleasant sound? No more logging into twenty different applications on your TV with a fiddly remote control. No longer the worry that using a banking application via a smart speaker could be carried out by other household members without you being present with your device. In this research we have explored the concept that user selected sounds with hidden keys embedded in them could be used to improve the security and usability of voice-controlled devices.

To set the context, it must be recognised that the way we interact with digital systems in our daily lives is changing. From checking our smartphones for news and social media, asking our smart speakers to play music or turn the heating on through to even driving our cars, the days when using a computer only meant sitting in front of a keyboard, mouse and screen are over. Touch screens and voice interfaces have become the most prevalent way we command many of today's consumer electronics. In particular, voice interfaces and the smart assistants that sit behind them have become ubiquitous. These are incorporated into phones, cars, smart speakers, and everyday household appliances. According to research, smart speaker usage has increased since the COVID-19 pandemic with popular uses including playing music, requesting a weather forecast and internet search [1]. In 2018, 50% of all adults had used voice for internet search and were over a billion voice searches per month [2]. Useability challenges impact a substantial proportion of the world's population for whom the use of existing digital technology is a struggle. Cognitive capabilities, physical difficulty in handling and manipulating devices, visual impairment, physical pain, social exclusion, or financial challenges impact the use and adoption of all digital technology. Speech interfaces have increased accessibility for many with cognitive and physical challenges. The largest technology platforms and providers such as Apple, Amazon and Google are all racing to deliver artificial intelligence driven, conversational technology that will change every sector of society and fundamentally change our relationship with technology [3]. The use of voice computing in the consumer and domestic context is very common for tasks such as playing music, home automation and internet search. By comparison, voice-based applications via smart

speakers and IOT devices have had relatively low acceptance in high security and corporate or business contexts. For non-voice based high security and corporate scenarios, multi-factor authentication is widely used. Multi-factor authentication systems in high security environments such as online banking often have favourable perception with regards to security but at the expense of usability [4]. In addition, there is a changing regulatory landscape in the banking and payments world with strong, multi factor authentication becoming an increasing mandatory requirement for many activities such as conducting financial transactions or retail payments [5].

With sensitive voice applications, the additional step of multi factor authentication or step up authentication is not yet common practice and providing an extra authentication has the potential to undermine usability if reliant on existing authentication methods that require user interaction. Online security concepts such as authentication and identification can also prove to be a barrier to vulnerable persons, and this can result in their exclusion from certain digital services and a reduction in the amount of time they spend online [6].

According to research carried out by the Office for National Statistics (ONS) in the UK, nearly 10% of the workforce in UK lacks digital skills and at least 5% of the population have disabilities which limit their ability to use digital technologies [7]. During the same period the voice assistant market has grown with hundreds of millions of devices equipped with voice assistant software. Amazon's Alexa and Google Assistant alone are now available on more than 500 million devices.

In this article, the authors propose an approach and outline initial experimental results of using audio steganography as an authentication factor for voice enabled devices. The approach utilises the audio capabilities of voice enabled devices to transfer authentication material embedded into sound to enable a hands-free authentication experience. In the first section the current approaches to authentication and related work are outlined. The second section focusses on the threats to the current approaches and a flexible architectural model to address the threats. The third section details the development of both the software and physical experimental set up to conduct initial testing, the results and analysis.

The concluding section of the article deals with a potential architectural model that could be used and what future work should focus on to make the audio authentication method more robust to environmental factors, and to explore further novel uses of this technique.

2. Related Work and Recent Developments in Authenticating Voice Channels

Authentication is the process of identifying or proving the identity of a user or process. Typically, authentication is categorised into one of three factors: something you know, something you have and something you are. An essential key part of ensuring that authentication is effective and reliable is to have a robust enrolment procedure that binds the user and the user identity within the system they wish to use. The current approaches to authentication for conversational computing and voice assistants varies depending upon the platform and the particular application. We will look at each of these in turn.

Account linking & device binding

Pre-authenticating the users device or linking a pre-existing user account to the voice enabled device is by far the most common approach in the scenarios for the interactions with consumer devices such as smart speakers, televisions, and mobile phones. Typically, the user will be required to download a mobile app and create an account that involves a process that the user has to go through to provide identification via a confirmation email or from a federated identity provider such as a social media account [8]. By asserting the identity of the user upon enrolment, the service provider is reliant upon the device being physically secured and only used by trusted parties. In these scenarios, the user journeys tend to be unauthenticated at time of use, relying on techniques such as physical security, pre-registration, and use via previously authenticated devices that are uniquely identifiable. It must be noted that using "wake words" or phrases to enable the devices into a mode ready to accept voice commands is not authentication, as no identity is asserted. This approach does not lend itself to applications where high confidence is required in the authentication process, multiuser scenarios such as an office or shared home, automated telephony systems such as Interactive Voice Response (IVR) systems, or where there is a regulatory requirement to strongly authenticate and assert the identity of a user [5].

Voice biometrics

Biometric authentication can be said to be authentication by a personal characteristic. In authentication terms, this is the concept of "something you are" and is subdivided into things you do (behavioural biometrics) and things that are (physiological). Biometric techniques can be either static (like a fingerprint, iris) or dynamic (behavioural, voice, heart ECG). Within the audio domain, the main method employed for biometric authentication presently is voice biometrics. Voice biometric techniques can be split into two categories: speaker verification and speaker identification. Speaker-verification authenticates a person is who she or he claims to be. It works by holding a database of reference voiceprints

and comparing the speakers captured voice print, with one that had been captured during a previous enrolment procedure. The challenges facing this method include considerations of background noise, health issues that might impact the speakers voice, and the quality of the audio or telephony equipment involved in both the enrolment and subsequent authentications. Also, as the stored voiceprint and one captured at a later authentication will be different due to noise, vocal imperfections and inflections, and the equipment used, the result of the authentication process is usually a matching or confidence score rather than a binary yes or no [9]. The performance of voice biometrics and speaker verification systems typically require the user to respond to a prompt with memorable but unique information such as either a set passphrase or password which maybe more difficult to conceal in shared spaces than a typed password. As a result, the issues with voice biometrics mean they perform poorly when compared with other biometric techniques despite high levels of usability and accessibility [10] [11] [12]. Current voice biometric based systems are more likely to incorrectly grant access than any other biometric methods and are much more likely to deny access to legitimate users of fingerprint, retina and iris scan biometric systems. When comparing biometric techniques in a recent study, the perceptions of security by users showed voice was perceived in the bottom third of methods in terms of security [13]. The accuracy of a biometric system is dependent on two measurable characteristics. The first is the rate at which the genuine users fail to authenticate, the false rejection rate – FRR (1).

$$\text{FRR} = \frac{\text{Number of False Rejections}}{\text{Number of Authentication Attempts}} \tag{1}$$

The second accuracy measure is the rate at which unauthorised users are authenticated when they should not be, the false acceptance rate FAR (2).

$$FAR = \frac{Number\ of\ False\ Acceptances}{Number\ of\ Authentication\ Attempts} \tag{2}$$

The point at which these two are equal is called the Crossover Error Rate (CER) and is generally accepted as the overall measure of the accuracy of the system. This can be seen in Fig 1. Error Rates in Biometric Authentication Systems. The CER can be optimised by adjusting the sensitivity of various attributes of an authentication system however these are typically a trade-off, e.g. tuning the system for lower false acceptances (FAR) can and often does lead to higher false rejects (FRR). It is the target for a well-designed system to have cross over error rate (CER) at as low or as near to zero as possible within the bounds of usability and other performance concerns (Fig. 1).

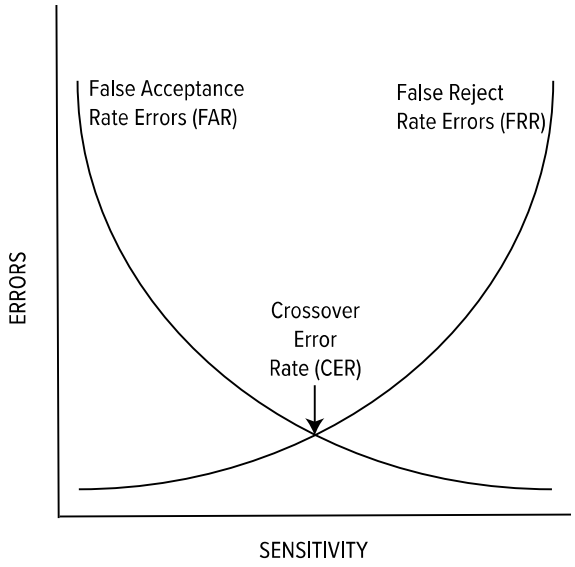

Fig 1. Error Rates in Biometric Authentication Systems

To compare the performance of various biometric techniques, the table below shows the strengths, weaknesses, performance, and security of various solutions below, presented by Timişoara (2016), Rui et al (2019) and Sabhanayagam et al (2018) [10] [11] [12] is shown in Table 1.

Table 1. Comparison of Popular Biometric Methods Showing Accuracy and Performance Drawing on Studies by Timişoara (2016), Rui et al (2019) and Sabhanayagam et al (2018)

Biometric Method	Acceptability	Accuracy	Permanence	Usability	Risk of Spoofing
Iris Scan	Medium	High	High	Medium	Low
Retina Scan	Low	High	High	Low	Low
Facial Recognition	High	High	Medium	High	Medium
Fingerprint	Medium	High	High	High	Medium
Voice	High	Low	Low	High	High

It can therefore be seen that the issues with the current voice biometric authentication solutions are problematic. Whilst they offer high levels of usability and accessibility, voice authentication can be subject to noise, health impacts affecting the subjects voice, poor quality micro phones and replay or spoofing attacks. As a result, voice biometrics are more likely to incorrectly grant access that any other method and are much more likely to deny access to legitimate users of fingerprint, retina and iris scan biometric systems. This can also be seen by comparison data taken from a recent study in Table 2. Biometric Performance Evaluation (Sabhanayagam et al. 2018).

Table 2. Biometric Performance Evaluation (Sabhanayagam et al. 2018)

Performance Criteria	Iris Scan	Retina Scan	Facial Recognition	Fingerprint	Voice
FAR	0.94%	0.91%	1%	2%	2%
FRR	0.99%	0.04%	10%	2%	10%
CER	0.01%	0.08%	-	2%	6%

Recent Developments

The challenges to the widely used approaches are therefore poor performance of voice biometrics; lack of authentication at point of use; poor perception of security when the user has to speak a password or memorable information out loud; vulnerability to replay attacks; addressing regulatory requirements for multi factor authentication and stepping outside of the voice channel to complete an authentication such as by using a screen or typing passwords which undermines the usability case for voice enablement.

Whilst each of the various biometric techniques continue to improve as the hardware and software improve, several proposals to overcome the current challenges have been published that augment existing biometric systems with either multi-factor authentication or by the addition of a challenge response. None of these have seen widespread adoption due to various challenges described as follows. Some of the techniques proposed rely on the proximity of Bluetooth enabled devices such as the users smartphone. In a scheme developed by Jansen (2005) et al. a multimode authentication framework was developed that embedded a challenge response protocol and ensured only users with a valid token were authenticated if they and their devices were in an authorised location [14]. The limitation of this approach is the reliance on the user being at a pre-authorised location. Whilst this isn't much of a restriction for smart speakers and TV's it is for mobile based voice assistants and applications. The approach taken by Jansen (2005) was expanded by, Hocking (2013) who proposed the concept of a cooperative and distributed approach to user authentication on mobile devices called *Authentication Aura* supplemented by continuous authentication within a users personal network leveraging facial biometrics and voice recognition [15]. By leveraging the presence of multiple devices within the vicinity of an authentication, greater trust can be achieved however this is at the expense of usability when the user has to conduct an additional biometric authentication. With the increased prevalence of screen and keyboard-less Internet of Things (IoT) devices, Gu and Lui (2016) developed a method for IoT devices to agree keys with each other called the group audio-based authentication scheme for IoT devices (GAB-IoT). The method however requires that the user has a smart phone which has a pre-enrolled security association. The IoT devices are then required to be co-located in an area where they can all receive an audio signal to the device with the security key [16]. In a scheme set out by Burch, Angelo and Spring, audio is the chosen media for a proximity based beacon [17]. In the scheme, mutual authentication between devices is achieved by exchanging audio signals between a desktop and the users mobile device via an audio authentication server. This approach relies on acoustic communication in both directions which doubles the opportunity for interference from noise and there is no reference to hiding the sound or making part of the overall user experience. In a radically different approach taken by Feng et al. (2017), a system is proposed called Vauth. The Vauth system aims to provide ongoing authentication of the speaker by ensuring the voice originates from the speakers throat by using a matching algorithm to match the inputs from the microphone and an accelerometer however this approach is only practical for phones and had limited accuracy in its current arrangement [18].

In another study into using voice assistants for financial applications, research carried out by the authors (Vassilev et al. 2020) created a cloud based two factor authentication scheme for banking applications such as making payments and balance enquiries. In the scheme, the voice assistants are pre-registered to the authorised user and the user is issued a token capable of both receiving and emanating a multitude of audio and radio based signals. The scheme utilised the cloud service provider API of the voice assistant to furnish the requests to the bank via an Open Banking API. A practical implementation of the system utilising radio beacons validated the approach as feasible in a real-world context of a user possessing a smart speaker wishing to conduct secure transactions via a voice assistant [6].

Typically, the authentication steps researched in this review improve the security of the overall authentication process for a system but potentially at the detriment of the usability and acceptability. As a result, most of the solutions currently used for authenticating voice assistants and voice activated technology do not use multi factor authentication and have limitations in terms of privacy, usability, and security [19]. Therefore this research is motivated by addressing the current and emerging threats, and the usability and security performance issues that have been outlined in this section.

3. Threat Model

Threat modelling provides a way of deriving security requirements in the development of a system or architectural model. By detailing the elements of the system and the relevant threats facing each, appropriate mitigations and design decisions can be made. This requires an understanding of the Tools, Techniques and Procedures (TTP's) that might be used to attack the system. In this section, an outline of the current and emerging threats to voice-based systems are detailed.

Current and Emerging Threats

High profile vulnerabilities reported in the media, and the work by security researchers has drawn much attention to the many constraints of voice only interactions with smart speakers and phones, and the lack of command confirmation, effective voice authentication or any additional authentication factors [20] [21]. Examples of these are as follows:

An investigation into the use of voice biometrics by a reporter in the UK revealed an issue where his telephone banking account service was accessed by his non-identical twin brother. The reporter enabled voice access to his bank account by enrolling in a telephone banking service. With the reporters consent, his brother made several attempts and eventually succeeded in accessing the account by mimicking his brothers voice. A critical flaw in the system was that it allowed for repeat attempts to access the account without alerting the customer or the bank [19].

Freely available high-quality tools are now available that allow for voices to be cloned and for text to speech (TTS) services to be leveraged to automate attacks on voice biometric systems. Improvements in machine learning and natural language processing have greatly increased the quality of the output of synthetic speech systems. The amount of data required to train synthetic voices has decreased with only modest amount of audio or video content needed to train realistic voice clones. Demonstrations of how such attacks can be carried out using modest resources have been shown at hacking conferences and published by Seymor et al. [22]. One of the attacks demonstrated utilises an online voice simulation from a company called Lyrebird and was trained by providing 30+ voice samples and then can be commanded to convert any text to speech using the newly created "voice avatar" [23]. In another method demonstrated by Seymore et al. which was more robust to factors such as noise and interference was utilising open source tools Tacotron [24] [25] and Wavenet [26] and utilising voice samples freely available to the public. It was concluded that given samples of sufficient quality, basic voice authentication is relatively easy to subvert.

Covert methods of attack are emerging that allow the malicious actors to operate and gain covert access to voice-controlled systems and assistants utilising inaudible ultrasonic sounds [21]. Another covert method which is incomprehensible to the human owners of such systems, utilises non-sensical word sequences which are interpreted as commands [27] [28]. Tools to exploit misinterpretation of maliciously crafted commands have been developed to exploit errors in interpreting commands. Exploiting adversarial linguistic models, tools such as "LipFuzzer" have demonstrated how mis interpretation of commands can lead to dangerous semantic inconsistencies and exploitable security weaknesses [29].

Research has also demonstrated that it is possible to use laser light to remotely inject inaudible and invisible malicious commands into voice control enabled devices such as smart speakers, tablets, and phones across large distances and even through glass windows and from adjacent buildings [30].

Another emergent source of concern is fake voice applications and skills. In the same way fake and malicious computer programmes and mobile apps have been created, there is a threat from malicious voice-based apps and skills. Typically, skills are developed by third parties and a threat actor can create a skill that impersonates another skill by having an invocation name that is similar or the same as a legitimate application or it can be a skill that appears to be legitimate but has hidden and unwanted functionality. Such techniques rely on the voice assistant core service having insufficient checks and assurance over 3rd party applications registered for use. It has been demonstrated that it is possible to register fake and malicious voice applications and skills by Zhang et al. [31]. In what has been termed as a Voice Squatting Attack (VSA), these fake and malicious skills can be invoked by ensuring the fake skill has an invocation command that sounds like a legitimate one. Other adversarial techniques used by malicious skills and apps include fake authentication (requesting sensitive data from the user) and fake termination by leaving the skill open for listening or further malicious operations. These are summarised in Table 3. Classes of Attack.

Table 3. Classes of Attack

Class of Attack	Examples
Fake Skills & Voice Squatting	The attacker crafts a fake skill that is used to capture sensitive information through interaction or covert recording. Invocation of the skill can exploit phonetic similarity to legitimate skills.
Voice Impersonation & Deep Fake	Attacker mimics the legitimate speaker for verification/authentication using either a human or a trained voice model (deep fake) in order to be able to respond to verification challenges.
Covert & Side Channel Attacks	Attacker uses laser light, ultrasound, infrasound or fragments of sound embedded into other sounds to initiate covert commands.
Voice Synthesis	Speech synthesis and/or text-to speech adapted to the characteristics of the target or brute force attack. This category can include adversarial synthetic voice models designed to exploit weaknesses in voice interaction models.
Replay Attack	Attacker replays a covertly recorded voice sample of a legitimate user and replays in back to the system to authenticate.
Technical Exploitation	A conventional "hack" where the attacker utilises malicious code and/or exploits a coding flaw or other vulnerability in the implementation of the system

Creating the Threat Model

In order to create a threat model for voice-based systems, it is helpful to consider a general model of the architectural arrangements of such systems. The World Wide Web Consortium (W3C) Voice Interaction Community have created a standardised model which is a generalised architecture for intelligent personal assistants, the most prevalent voice based application. It contains the components and tasks required to deliver a voice based intelligent assistant. Whilst the model is architecture focused rather than specifically looking at threats to security, it provides a useful reference from which to start to build a threat model from [32]. The model consists of three layers and is helpful to introduce the inner components

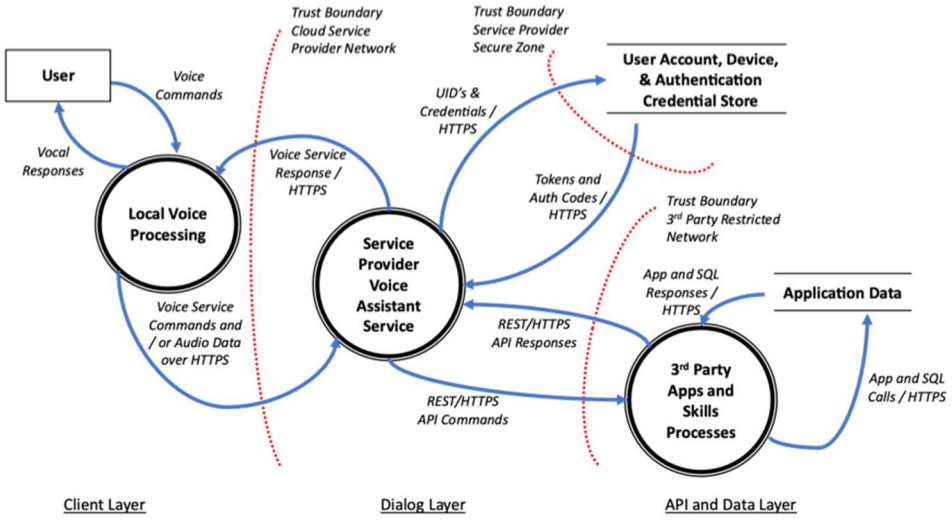

Figure 2. Simplified General Threat Model for Voice Enabled Assistants and Systems

used within a typical architecture. For this research an adapted version of this model was considered and simplified at the start of the creation of the threat model with trust boundaries added to each layer as shown in Figure 2 shown above.
The client layer in the modified model has a number of key characteristics. The Intelligent Personal Assistant Client is the voice assistant client that utilises the client hardware to open the microphone for voice input upon the utterance of a wake word or push of a dedicated hardware button, listen for dialogue, and pass this on to the dialogue layer in addition, a loudspeaker to issue responses. A number of other services are often client layer such as holding a unique device identifier, provision of location information, local voice processing (both listening and for voice response and synthesis), volume controls, status lights and indicators and in some cases a visual display.
The dialogue layer consists of a set of services that mediates communications from the client layer and the applications and skill provided by the core service provider or a third-party app or skill. In order to interact with the services a voice interaction model is required. Typically, a service layer interprets dialogues, manages speech recognition, converts text to speech and manages dialogue. The dialogues are referenced against a registry and then interpreted as intents. The voice interaction model set by the provider allows for the speech to be understood in a way that allows the devices to awaken, launch, or invoke specific actions and accept voice-based data into applications. In many deployments, this is also known as the skill interaction model.
The core service provider layer handles the services needed to provide the service. The functionality provided by the core service provider in the model includes functionality such as user accounts, authentication information, device registration, natural language understanding and a selection service enabling voice applications to access core service provider skills and data or 3rd party provided data and applications. In addition, it is the natural language understanding provided by the core service provider that can convert the utterances from the user into meaningful text for processing.
The 3rd party API and data layer comes into play if the service selected is a 3rd party app, a third party "skill" or if the device is executing a command or service that requires 3rd party data. Third party apps and skills utilise web services, data and other functionality that is hosted separately from the core service provider. In the model the third-party skills or apps have to be registered with and accessed via the core service provider, and are called upon when the user interaction requires this. This could for example be when the user requests the voice assistant to open and play music from a 3rd party streaming service or requests local weather provided by a 3rd party.

Table 5. Attack Tools Techniques and Procedures

Tactic / Technique	Description	Mitre ATT&CK Reference
Audio Capture (Enterprise)	In this attack, the threat actor gains access to the targets microphone or video camera on their device to capture audio. The audio is reviewed to listen for sensitive information such as conversation or sounds (keystrokes). This TTP can leverage malware to stream data or write to a file for later collection.	https://attack.mitre.org/techniques/T1123/
Capture Audio (Mobile)	In this attack, the threat actor gathers audio from the target's mobile device microphones. This can be via malicious applications, inappropriate permissions, malware and physical access.	https://attack.mitre.org/techniques/T1429/
Obfuscated Files or Information: Steganography	This technique is used by attackers to hide information in images, audio or video. Steganography is then used for malware, data hiding and data exfiltration.	https://attack.mitre.org/techniques/T1027/003/
Collection	The collection tactic, is where the attacker is collecting data of interest to their goal (in this case digital or analogue audio information). To do this the attacker either leverages the targets computer or mobile device, or an application that handles audio data such as streaming apps, voice and video calls.	https://attack.mitre.org/tactics/TA0009/
Credential Access	The tactic of credential access is where the attacker steals account names and passwords. Traditional techniques supporting the tactic include key logging, brute force guessing, network interception. In the context of this research, this would include listening for any audio that has usernames or passwords embedded in it.	https://attack.mitre.org/tactics/TA0006/

The next phase of developing the threat model was to further enumerate the researched threats, build attack trees as graphical representations of the components or elements required for a successful attack. Additionally the Mitre ATT&CK framework was used to gather additional information about the threats, specifically the applicable Techniques Tools and Procedures (TTP's) as shown above in Table 5 [33]. Once the target of an attack has been located and reconnaissance carried out, often the main objective of the threat agent or adversary is to gain the capability to gain account access by stealing credentials such as user ID's, tokens, and passwords. The TTP's typically deployed therefore include the interception of keys and passwords in transmission; and the capture, replay, brute force guessing, offline cracking, and other attacks on key and password storage as shown in Table 5. Once credentials are obtained, the threat actors can choose how and when to attack. This can then facilitate onward stages of access to other systems and escalation of privileges in order to achieve the threat actors aims. In addition to the audio specific threats, the threat model developed proposes to mitigate the same threats any other authentication scheme would face.

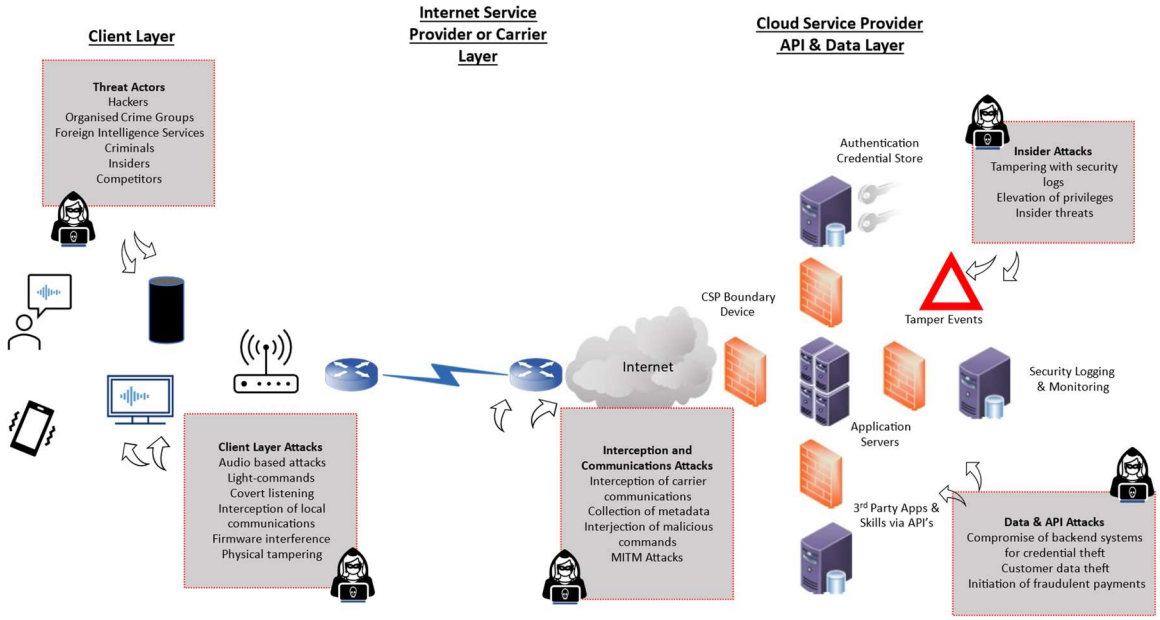

Figure 3. Architectural View of the System Threat Model

The next phase of the research was to consider how threats identified in threat modelling might be addressed in a way that minimises user authentication friction whilst maintaining or even enhancing security. Creating an initial simple model of how an authentication system could incorporate audio steganography into voice assistant and IOT authentication scenarios helped derive the focus for experimentation as shown in Figure 3.

With audio, anything transmitted through the air can be subject to local trivial interception, capture and replay. In the proposed model, the main purpose of the audio steganography is to facilitate transmission whilst preserving the user and audio user experience. It must therefore be assumed that this information can and will be intercepted, captured, retained, and replayed by a threat actor.

The proposed model as shown in Figure 4. Architectural Model – Audio Steganography Key Arrangements, assumes the use case of an enrolled smart speaker or tv. In the model the user initiates a voice command which is received by the voice assistant device. Once the originator of the request and its validity has been confirmed, the voice assistant responds with two things alongside the command response: a session key and the users pre-enrolled audio public key. The audio key can be decoded by the steganography decoding and checked against the locally stored private key. In addition, the voice assistant service provider generates a unique session token which is sent to the users pre-enrolled device. Using the private key in the local key storage, the users device then generates a session key with the token, the location, device ID and user ID. Once the session key is received the voice assistant service provider can execute the command and any onward actions.

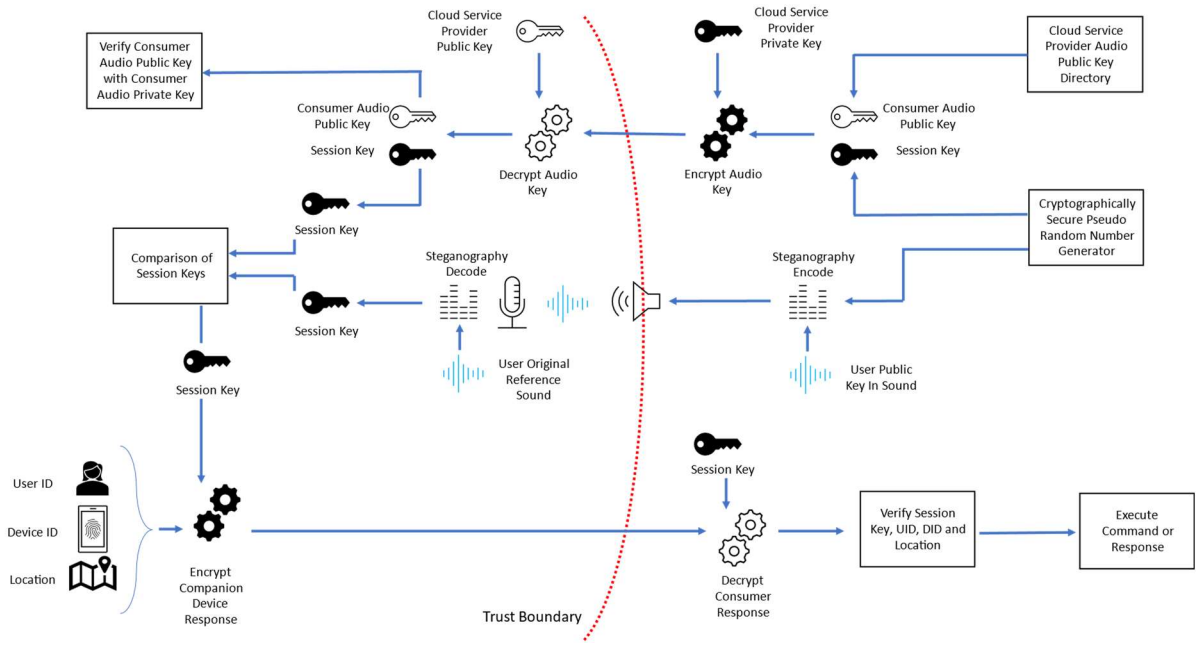

Figure 4. Architectural Model – Audio Steganography Key Arrangements

4. Experimental Setup

An experiment was devised to test, analyse, and evaluate the use of steganography to hide authentication material such as tokens or cryptographic keys into short audio sound clips. The experiment set out to demonstrate the ability to encode a data within a sound, transmit that data embedded in a sound to another receiving machine and then decode or recover the data from the transmitted sound in the contexts that might be encountered as part of an authentication scheme such as via digital transmission and via acoustic broadcast. The experiment was focused on the ability to hide information within an audio signal. Referencing a simple steganography model as per Figure 5. Steganography Model, the carrier or cover message in this instance is an audio sound and the message the data we wish to hide. Steganography is used to hide information rather than secure it. With this in mind, part of the carrier sound can be used as the key for encoding and decoding purposes as shown in Figure 6. This as part of developing an architectural model but for the experiment, a method was developed that could be used with adjustable parameters to compensate for background noise and other environmental factors [34].

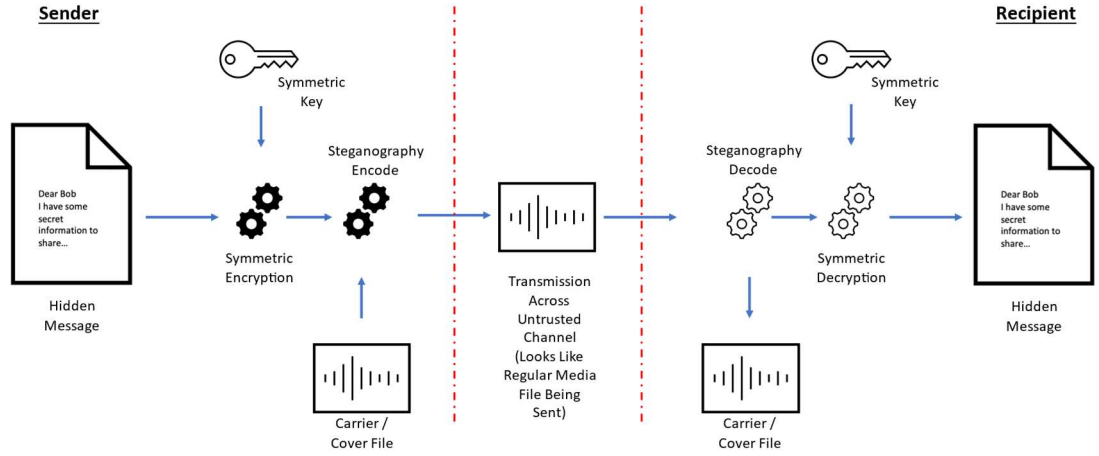

Figure 5. Steganography Model

I. J. Computer Network and Information Security, 2020, 5, 56-73
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

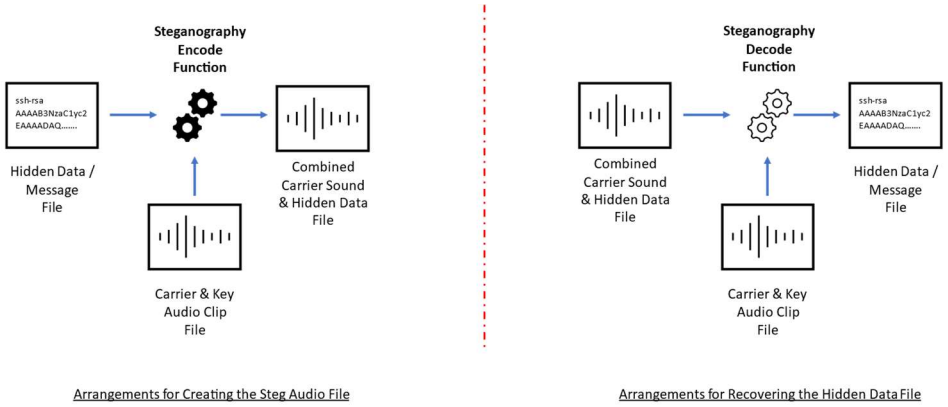

Figure 6. Array Based Steganography Program Inputs and Outputs

Figure 7. Studio Layout

5. Steganographic Technique for Experimentation

At a basic level the approach taken was to convert the information to be hidden into an array with each element representing the ASCII value of the character, to divide each element by a sensitivity variable to reduce its magnitude and then add the resultant array to the original sound x_o [n] to encode the hidden information into the carrier sound. The original or carrier signal in the digital domain is a series of discrete values from the start of the sound at $n = 0$ to the last sample at $n = n_{max}$ and can be shown in (3) and simplified as in (4):

$$x_o\,[n] = x[n] + x[n-1] + x[n-2] + x[n-3] \ldots . \quad 0 < n < n_{max} \tag{3}$$

$$x_o = \sum_{n=0}^{n_{max}} x[n\text{-}n_{max}] \tag{4}$$

We can represent the encoded signal by considering the inputs and deriving the output as a function of *a* [n] which is the numeric ASCII Code values of the range 33 to 126 (all the valid characters that could be used in a message or key) and *s* which is a sensitivity variable selected by the user, and *l* which is the length of hidden information or key to be hidden. We then create a series of values called the modifier array x_m to combine with the original signal based on the ASCII values of the key material to be embedded and a scaling factor known as the sensitivity variable shown as *s* (5).

$$x_m[n] = \frac{a[n]}{s} \qquad 0 < n < l \tag{5}$$

A simple method of combining modifier array x_m and the original signal x_o, is to add each element to encode the message for the samples up to the length l of the hidden information as in (6):

$$x_t[n] = x_o[n] + x_m[n] \quad 0 < n < l \tag{6}$$

We can represent the captured signal x_c as the discrete values between zero and the end of the sound from the extracted from a .wav file as in (7)

$$x_c = \{x[n]\} \quad 0 < n < n_{max} \tag{7}$$

When the encoded sound is transmitted via digital means, both the original signal $x_o[n]$ and $x_c[n]$ have a common datum and length making comparison and signal extraction (decoding) a straightforward processing and provided the transmission media and any file storage has integrity, the accuracy was expected to be exceptionally high. The encoded signal can therefore be represented as (8) and decoding simply the reverse up to the maximum length of the hidden data given as n_{kl} :

$$x_t = \sum_{n=0}^{n_{max}} x_o[n\text{-}n_{max}] + \sum_{nk=0}^{n_{kl}} x_m[n\text{-}n_{kl}] \tag{8}$$

In order to make the challenge of decoding the embedded information more robust to errors, the encoding process was modified. By encoding each of the hidden data points over multiple samples, it was envisaged that the technique can be made more robust to lost samples, misalignment and noise. The ASCII characters of the information to be hidden are scaled with the sensitivity variable and added to the original sound as before but instead of one character per sample, the scaled character is added to a number of samples g, to form a group. The greater the size of the group, the more robust the technique however the less information can be stored within a sound clip of a given timespan. When decoding the sound, the comparison is made as before however a rolling average of difference is taken for each of the groups where the sample group size is g (9):

$$xa[n] = \sum_{n=0}^{n=g-1} \frac{1}{g} x_c[n] + \sum_{n=g}^{n=2g-1} \frac{1}{g} x_c[n] + \sum_{n=2g}^{n=3g-1} \frac{1}{g} x_c[n] + \sum_{n=3g}^{n=3g-1} \frac{1}{g} x_c[n] \ \ldots \tag{9}$$

For the experiment we considered values of g based upon the length of sound clip we wished to use, the amount of data we want to hide and the quality of the sound (resolution, bit depth and sampling frequency). Considering the use case for hiding a 2048 bit cryptographic key, realistic values of g were derived for experimentation purposes in Table 6:

Table 6. Audio Sample Characteristics

Property	Value
Sample rate	44.1 kHz
Samples per second	44,100
Sample resolution	16 bit
Bits per sample	2^{16} or 65,536
Normalised to:	-32,768 to 32767
Bits per second	1,411,220 bps

If we encode one character per sample and take the example of a 2048 bit RSA key (381 ASCII characters) we can see that it takes just under a 100th of a second to transmit the whole key (10)(11):

$$t = \frac{l * g}{p} \tag{10}$$

$$t = \frac{1*381}{44{,}100} = 0.00864 \text{ seconds} \tag{11}$$

If we use the groups average approach with a g value of 100 we can see it takes just under a second to transmit the key (12).

$$t = \frac{100*381}{44{,}100} = 0.864 \text{ seconds} \tag{12}$$

If we use a *g* value of 1000 we can see it takes almost 9 seconds to transmit the same amount of information (13):

$$t = \frac{1000*381}{44{,}100} = 8.64 \text{ seconds} \tag{13}$$

If the hidden material is to be repeated for resilience and error correction purposes, a relatively short period is desirable. Therefore, for the experiment using short sound clips the working range of up to 1 second for the hidden material was used which translates to a value of *g* of 100 samples or less.

The next step was to generate simulated noise. The lab took place under the controlled conditions of a recording studio. For cryptographic schemes however, there is a need to preserve the fidelity of key material exactly. Even a single error can render the key unusable and therefore the scheme will fail. Using the grouped averages mitigates this, but to what extent needed to be tested. Therefore, for the experiment, a noise signal was created using the pseudo random number generator in code. The signal was a normally distributed set of values around a scaling factor that could be controlled, thus simulating different noise levels. For repeatability, the random number generator was seeded to provide the same values when automating tests within a loop. The simulated noise signal was therefore created as a vector and then added to the steganographic encoded signal as a series of discrete values where x_s is the noise, x_o the original signal, and x_t is the total sum of the original signal and the noise (14) and simplified as (15):

$$x_t[n] = x_o[n] + x_s[n] \quad 0 < n < n_{max} \tag{14}$$

$$x_t = \sum_{n=0}^{n_{max}} x_o[n\text{-}n_{max}] + x_s[n\text{-}n_{max}] \tag{15}$$

For the experiment, the percentage of hidden message accurately recovered was measured along with the signal to noise ratio. The signal to noise ratio was calculated in dB as follows (16):

$$SNR_{dB} = 10 \log_{10} \left(\frac{x_o}{x_s}\right) \tag{16}$$

To measure the percentage of message recovered, each of the elements were compared in turn, and then a percentage calculated where *m* is the number of matched elements between x_d and x_c (17)(18):

$$\text{Match } \% = 100 \left(\frac{m}{l}\right) \tag{17}$$

$$\text{Match } \% = 100 \frac{mg}{l} \tag{18}$$

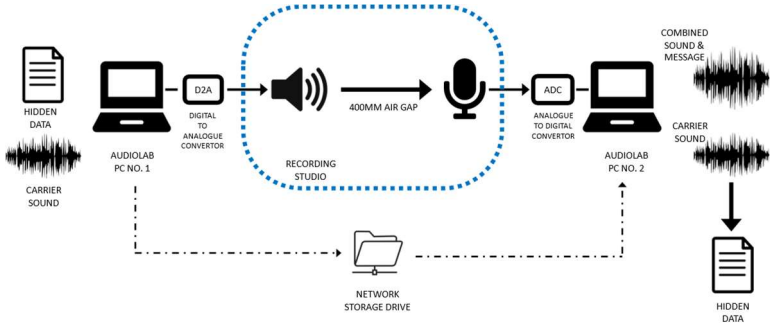

Figure 8. Audio Steganography Experimental Layout

I. J. Computer Network and Information Security, 2020, 5, 56-73
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

6. Results and Discussion

The experimentation was carried out in a number of phases. The acoustic experiments were carried out in an acoustically controlled environment. The first phase consisted of the transmission of voice and music sound clips containing a steganographic embedded text message of 189 bits, sent over digital media and then via acoustic transmission. The experiment was arranged as per Figures 7 and 8, and was repeated at different encoding depths by varying the sensitivity variable. Each of the encoded sound files created were played back on the sending machine (Audio Lab Machine 1) several times to ensure multiple opportunities were available to capture the encoded sound correctly on Audio Lab Machine 2, which was then used to record at least three loops of the encoded sound and the resulting capture saved for later analysis.
The Table 7. below, shows the results for each sensitivity variable setting, the outcome of the decoding process, the decoded output itself, and a subjective assessment of whether the encoding process impacted the audio signal or was clearly visible in a simple time vs. amplitude plot. In the case of both the voice and the music carrier samples, the decoding of the hidden text ceased at sensitivity variable settings of above 10,000 noting the larger the number, the lower the encoding depth. It was easier to audibly detect interference in the voice sample as expected as there are less frequencies

Table 7. Music & Voice Sample Tests Results Table

Music Sample Tests			
Sensitivity Variable	**100% Decoded?**	**Visual Interference?**	**Audible Interference?**
100	Y	H	Y
1000	Y	L	N
5000	Y	L	N
10000	Y	L	N
50000	N	L	N
75000	N	L	N
100000	N	L	N
500000	N	L	N
Voice Sample Tests			
Sensitivity Variable	**100% Decoded?**	**Visual Interference?**	**Audible Interference?**
100	Y	H	Y
1000	Y	M	Y
5000	Y	M	N
10000	Y	L	N
50000	N	L	N
75000	N	L	N
100000	N	L	N
500000	N	L	N

present and more quiet periods in the clip. The interference manifested itself as a short click at the deeper encoding depths and as such it was harder to hear any interference in the music samples. It was also noted that the samples that failed to decode (at sensitives above 10,000) had a high incidence of similar characters. At sensitivity variable settings of 5000 and 10,000 the messages were recovered reliably whilst not exhibiting any outward signs of interference or hidden information. These can be considered the optimum settings.
The experiment was expanded to test the transmission of cryptographic key material, an RSA key of 2048 bits. The tests were completed using encoding of 1 character per carrier sound sample. The purpose of the steganography is to hide or embed the key in a pleasant sound, rather than a secure it. To establish the optimum working range of the test was repeated with sensitivity variable values between 100 and 100,000 in 100 increments. The decoded key was then checked, character by character and the percentage of key match to the original recorded. To use the recovered message as a key, 100% accuracy has to be achieved, and the working range can clearly be seen in the plot in Figure 9. which shows the percentage of recovered key material in the audio sample test. This test established the working range of between 100 and 16,500 for the sensitivity variable in this scheme. Further tests were carried out to establish the impact of encoding each hidden character in groups of samples rather than single samples and then looking at the impact of this on the resilience and robustness of the signal recovery to noise interference. With this method, each of the hidden data points are encoded over multiple samples with the aim of making the steganography technique more robust to lost samples, misalignment and noise.

I. J. Computer Network and Information Security, 2020, 5, 56-73
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

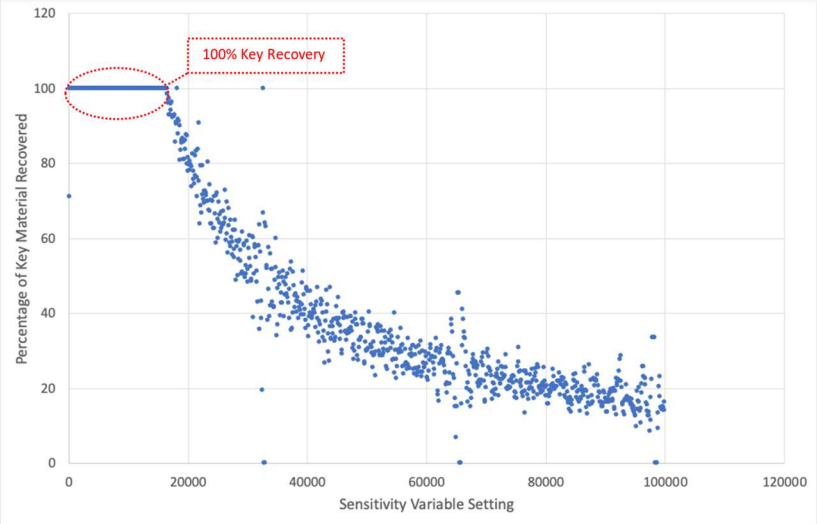

Figure 9. Percentage of Recovered Key Material in Audio Sample Test

Tests were carried out using the grouped average technique, with group sizes of 10 to 200 samples per encoded character, in increments of 10 and there was no impact on the key recovery. The group size of 100 was selected as this was calculated as the optimum for the length of sound clip under test. A side effect is that the encoding can be seen slightly more clearly when plotting the amplitude of each sample using the grouped average method because the encoding spans multiple samples.

Interference and Noise

An objective of audio steganography is to make the alteration to the carrier sound so minute that it is undetectable to the observer whether by the human auditory system or by digital monitoring. In order to understand the impact on the signal quality of encoding a message within the audio signal, calculations were made comparing the original and encoded signal. All audio steganography techniques impact the Signal to Noise Ratio (SNR) and this is significant as it helps understand the distortion the hidden data is introducing and has a direct impact on the amount of data that can be hidden [35]. To quantify this, the hidden message was considered as adding noise to an original audio clip. Testing in the noise controlled

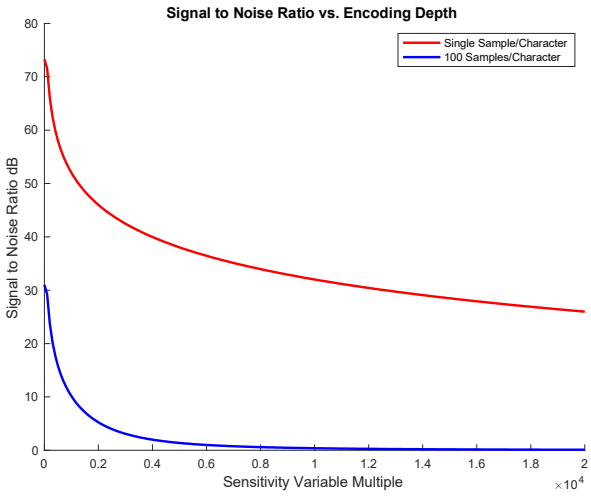

Figure 10. SNR vs Encoding Depth Using Single and Grouped Average Steganography

environment of a recording studio, the data was collected from tests comparing the Signal to Noise Ratio (SNR) between the original and encoded signal at different sensitivity variable settings and comparing the grouped average approach and single sample per character encoding. The higher the SNR, the better performance, as this signifies there is more original signal than unwanted data or noise. It can be seen in results in Figure 10. which compares the SNR to the encoding depth first using single character per sample and then the grouped average steganography, that both grouped and single sample

methods have better SNR's at lower sensitivity variable settings. The next test conducted was to investigate the impact of noise on the message recovery. For the test, a progressively stronger noise signal, created from a normally distributed set of random values was added to the signal to be captured. The percentage of message recovered was used as a measure of effectiveness as anything less than a 100% recovery would be unusable for cryptographic schemes unless additional error correction protocols were used. The test was repeated at different embedding levels by adjusting the sensitivity variable. Each run of the test the noise signal was increased in stages until the signal could no longer be recovered.
It can be seen by the results outlined in Figure 11. that at the more robust sensitivity variable levels in the range of up to

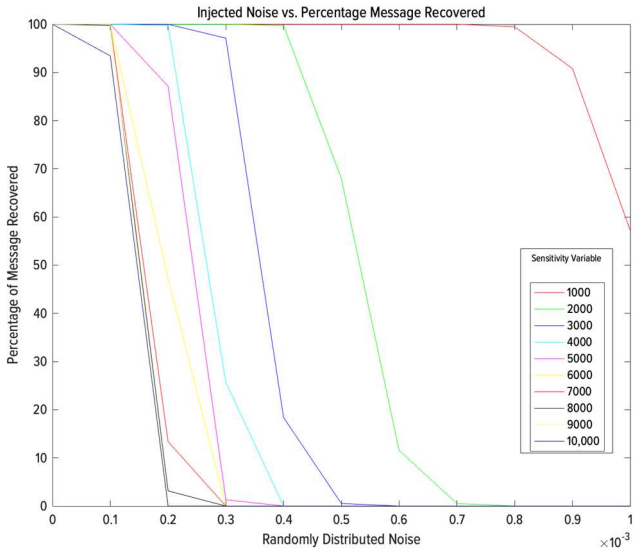

Figure 11. Impact of Noise on Message Recovery

2000, the message recovery is fully effective. At the less detectable but highly sensitivity variable settings towards 10,000, the message recovery accuracy drops below 100% quickly as more noise is added e.g. as little as 0.0001 times the amplitude of the original signal. This was expected and there is a trade-off between detectability in the form of interference in the audio sample and the message recoverability.

A remaining challenge to address is the alignment of the original signal to the captured signal. This is because if the signal is captured acoustically over the air, there is no common datum or origin between the original to the captured signal that contains the embedded information. A simple technique is to shift the signal along sample by sample until the best match is found. To achieve this, one can represent the captured signal as the discrete values between zero and n_{max} which is the end of the extracted sound as follows in (19):

$$x_c = \{x[n]\} \quad 0 < n < n_{max} \tag{19}$$

We can then compare x_c to the original signal x_o by cycling through values of n until the difference between the two across the series is at its minimum (20):

$$x_o\,[n] \overset{\Delta}{\rightarrow} x_c\,[n - n_k] \tag{20}$$

Digitally transmitted sounds did not require this however using the max min feature in MATLAB the signals were scaled, although this also required manual experimentation.

I. J. Computer Network and Information Security, 2020, 5, 56-73
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

Steganalysis

As well as comparison of the original sound by plotting the signal amplitude vs. time, two other simple techniques can be used to detect the presence of the impact of hiding information. Histogram analysis is an approach taken to steganalysis to detect the presence of hidden messages in image files [36]. It is also a way of quantifying the difference between the original message and a file containing a hidden message. The technique works if the encoding method is in the time domain. In order to understand the impact of encoding a message into the audio file, histograms were created for both the original and encoded signals using first the 100 sample grouped encoding and the single character sample encoding technique. Both outcomes were as expected. For both tests the difference between the signals was relatively small however, when using the 100 sample grouped average technique the histogram plot shows minimal difference between the encoded and original signals as per Figure12. The difference was more marked when using a single sample per

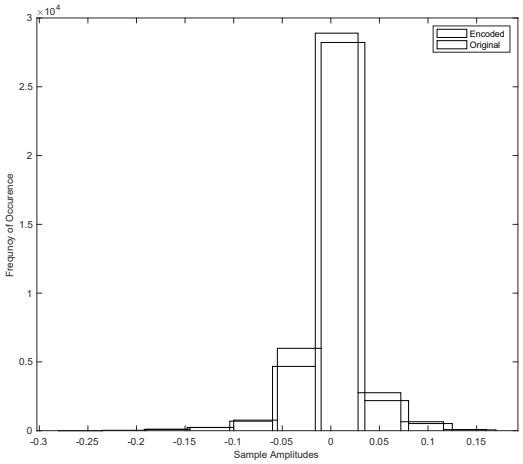

Figure 12. 10 Bin Histogram Comparison of Original and 100 Sample Grouped Encoded Signal

character when encoding as expected. The differences are of course only possible to display if the person conducting the analysis has access to the original audio signal without the encoding.

Another method to investigate the presence of hidden data within a sound is to look in the frequency domain. Spectrograms are a visual representation of the spectrum of frequencies present in the sound sample as it varies over time and plotting the spectrograms for each test allowed for comparison between the original sound with and without the encoded output sound files that contain the hidden data to see if there were visibly any difference between the two as shown in Figure 13.

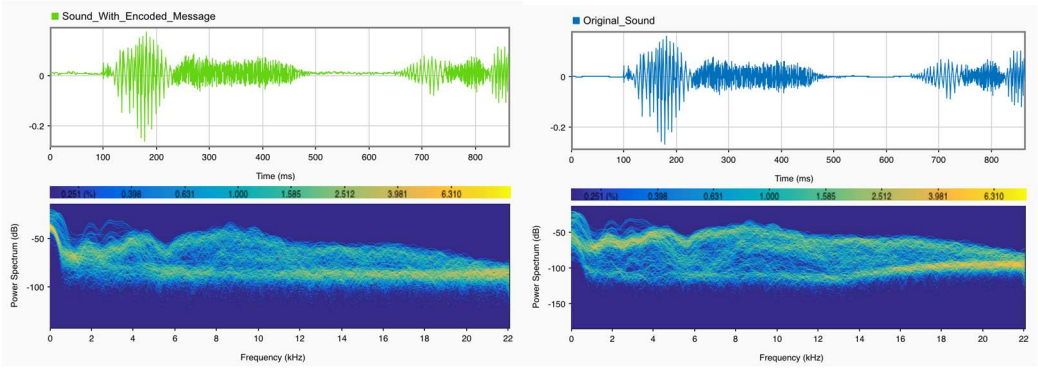

Figure 13. Comparison Spectrographs of Original vs Encoded Signal

As the steganography method developed was based upon minutely modifying the amplitude of parts of the sound rather than changing the sound itself, there was very little visible difference between the input and output frequencies present.

7. Conclusion and Future Scope

The ever-increasing ubiquity of voice driven computing and IOT devices is providing new levels of convenience and accessibility but is also driving new threats to our digital identities and assets. Current architectures have weaknesses that have led to comparatively low uptake of voice computing for high security applications or high-risk transactions. It is therefore likely that interest in securing these services will grow considerably.

In this research, the threats to voice and audio-based systems such as the lack of authentication at time of use, poor performing voice biometrics, deep fake and impersonation attacks and a number of covert side channel attacks were outlined in a new threat model. Additionally, the use of audio steganography to transmit key material was proposed as a novel way of mitigating these threats in an architectural model. The flexible architectural model that addresses the threats captured in the threat model can be applied to smart speaker, IoT and TV type applications for future refinement and development. Finally outlined in this research was the development and testing of an audio steganography technique that can be used to hide the key material or any data within sound that does not add any additional data size, is simple to implement and can be tuned to mitigate small amounts of interference, errors and noise. The technique which effectively amplifies groups of samples of a given reference sound proportionate to a scaling factor and the ASCII value of the data to be hidden, was able to be tuned such that recoveries were possible using a range of between 100th and a 16,500th of the size of the original sample and withstanding randomised noise signals up to 0.008 of the original amplitude with zero hidden payload data loss. Higher noise levels impact the data recovery with progressively less hidden data recoverable as noise levels increased. Based on these findings and as authenticating users using audio steganography to transmit key material is central to this approach, selecting and refining a high performing technique should be the focus of future development. Experimentation has shown this to be a promising but challenging method. The steganography technique developed for the experiment was robust in the digital domain and can be used to hide sufficient information for all manner of purposes including transmission of crypto key material. Significant challenges remain when transmitting the key material acoustically due to the issues of noise, scaling and alignment. These challenges however can be addressed in future research leveraging signal processing techniques deployed in other domains as discussed in the analysis section such as repetition, use of error correction algorithms, dynamic time warping or time shifting signals, and further tuning. In addition, the development of the architectural model to transmit shorter session keys or PIN type numbers would mean other less efficient steganography techniques could also be deployed.

References

[1] G. Kesten, “15 mind-blowing stats about voice assistants,” Adobe Inc., 21 September 2020. [Online]. Available: https://blog.adobe.com/en/publish/2020/09/21/mind-blowing-stats-voice-assistants.html#gs.cu22jz. [Accessed 10 August 2021].

[2] A. Marchick, “Voice Search Trends,” Alpine AI, April 2018. [Online]. Available: https://alpine.ai/voice-search-trends/. [Accessed 4th May 2018].

[3] J. Vlahos, Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think, USA: Houghton Mifflin Harcourt USA , 2019.

[4] N. Gunson, D. Marshall, H. Morton and M. Jack, “User perceptions of security and usability of singlefactor and two-factor authentication in automated telephone banking,” *Computers & Security, vol 30, no. 4, pp. 208-220,* vol. vol 30, no. no. 4, pp. pp. 208-220, 2011.

[5] European Banking Authority, “Opinion of the European Banking Authority on the elements of strong customer authentication under PSD2,” European Banking Authority, 2019 June 21. [Online]. Available: https://eba.europa.eu/sites/default/documents/files/documents/10180/2622242/4bf4e536-69a5-44a5-a685-de42e292ef78/EBA%20Opinion%20on%20SCA%20elements%20under%20PSD2%20.pdf. [Accessed 29 February 2020].

[6] V. Vassilev, A. Phipps, M. Lane, K. Mohamed and A. Naciscionis, “Two-Factor Authentication for Voice Assistance in Digital Banking Using Public Cloud Services,” in *Confluence 2020 10th International Conference on Cloud Computing, Data Science and Engineering*, Noida , 2020.

[7] UK Office for National Statistics, “Office for National Statistics,” ONS, 18 February 2020. [Online]. Available: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork. [Accessed 29 February 2020].

[8] M. Saltzman, “How to Set Up A Smart Speaker: Step-by-step tips for getting started with Amazon Echo, Apple HomePod or Google Home assistants,” AARP, 11 October 2019. [Online]. Available: https://www.aarp.org/home-family/personal-technology/info-2019/smart-speakers-set-up-instructions.html. [Accessed 29 May 2021].

[9] J. A. Markowitz, “Voice Biometrics,” *Communications of The ACM,* vol. 43, no. No.9, pp. pp 66-73, 2007.

[10] C. Otti, “Comparison of Biometric Identification Methods,” in *11th IEEE International Symposium on Applied Computational Intelligence and Informatics* , Timişoara, 2016.

[11] Z. Rui and Z. Yan, “A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification,” *IEEE Access,* vol. vol. 7, pp. pp. 5994-6009, 2019.

[12] T. Sabhanayagam, V. Prasanna Venkatesan and K. Senthamaraikannan, “A Comprehensive Survey on Various Biometric Systems,” *International Journal of Applied Engineering Research,* vol. 13, no. 5 (2018), pp. pp. 2276-2297, 2018.

[13] O. Buckley and J. R. C. Nurse, "The Language of Biometrics: Analysing Public Perceptions," *Journal of Information Security and Applications,* vol. 47, pp. 112-119, August 2019.

[14] W. Jansen, S. Gravila and V. Korolev, "NIST Computer Security Resource Centre," National Institute of Standards and Technology (NIST), 2005. [Online]. Available: https://csrc.nist.gov/publications/detail/nistir/7200/final. [Accessed 3rd December 2018].

[15] C. Hocking, S. Funnel, N. Clarke and P. Reynolds, "Co-operative user identity verification using an Authentication Aura," *Computers and Security,* vol. 39, pp. 486-502, 2013.

[16] Z.-l. Gu and Y. Liu, "Scalable Group Audio-Based Authentication Scheme for IoT Devices," in *12th International Conference on Computational Intelligence and Security*, 2016.

[17] L. Burch, M. Angelo and B. Masoud, "Proximity Based Authentication". United States of America Patent US 9,722,984 B2, 1st August 2017.

[18] H. Feng, K. Fawaz and K. G. Shin, "Continuous Authentication for Voice Assistants," in *MobiCom '17 Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, Utah, 2017.

[19] D. Simmons, "BBC fools HSBC voice recognition security system," BBC News - Technology , May 2017. [Online]. Available: https://www.bbc.co.uk/news/technology-39965545. [Accessed 30th August 2018].

[20] 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*, 2014.

[21] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang and W. Xu, "DolphinAtack: Inaudible Voice Commands," in *ACM Conference on Computer and Communications Security (CCS)*, Dallas, 2017.

[22] J. Seymour and A. Aqil, "Your Voice is My Passport," in *Black Hat USA 2018 Website Whitepapers*, Las Vegas, 2018.

[23] Lyrebird, ""We create the most realistic artificial voices in the world"," 2018. [Online]. Available: https://lyrebird.ai. [Accessed 3rd March 2019].

[24] Y. Wang, "Audio samples from "Tacotron: Towards End-to-End Speech Synthesis"," [Online]. Available: https://google.github.io/tacotron/publications/tacotron/index.html. [Accessed 3rd March 2019].

[25] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech*, Stockholm, Sweden, 2017.

[26] A. v. d. Oord, S. Dieleman and H. Zen, "WaveNet: A Generative Model for Raw Audio," Deepmind, 8th September 2016. [Online]. Available: https://deepmind.com/blog/wavenet-generative-model-raw-audio/. [Accessed 3rd March 2019].

[27] M. K. Bispham, I. Agrafiotis and M. Goldsmith, "Nonsense Attacks on Google Assistant," 6th August 2018. [Online]. Available: https://www.cs.ox.ac.uk/people/mary.bispham/. [Accessed December 2018].

[28] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in *IEEE Symposium on Security and Privacy Workshops*, San Francisco, California, USA, 2018.

[29] Y. Zhang, L. Xu, A. Mendoza, G. Yang, P. Chinprutthiwong and G. Gu, "Life after Speech Recognition: Fuzzing Semantic Misinterpretation for Voice Assistant Applications," in *Network and Distributed Systems Security (NDSS) Symposium* , San Diego, CA, USA, 2019.

[30] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin and K. Fu, "Lightcommands: Laser-Based Audio Injection on Voice-Controllable Systems," Defense Advanced Research Projects Agency (DARPA) , 4th November 2019. [Online]. Available: https://lightcommands.com/20191104-Light-Commands.pdf. [Accessed 29 February 2020].

[31] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian and F. Qian, "Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems," in *Proceedings of 2019 IEEE Symposium on Security and Privacy (SP)* , San Francisco, 2019.

[32] D. D. Dirk Schnelle-Walka, "W3C Github Repository - Intelligent Personal Assistant Architecture," World Wide Web Consortium, 24 March 2020. [Online]. Available: https://w3c.github.io/voiceinteraction/voice%20interaction%20drafts/paArchitecture.htm. [Accessed 16 October 2020].

[33] MITRE Corporation, "MITRE ATT&CK," MITRE, 27th January 2021. [Online]. Available: https://attack.mitre.org/. [Accessed 13th February 2021].

[34] A. Phipps, K. Ouazzane and V. Vassilev, "Enhancing Cyber Security Using Audio Techniques: A Public Key Infrastructure for Sound," in *IEEE 19TH INTERNATIONAL CONFERENCE ON TRUST, SECURITY AND PRIVACY IN COMPUTING AND COMMUNICATIONS (TRUSTCOM 2020)*, Guangzhou, CHINA, 2020.

[35] H. Dutta, R. K. Das, S. Nandi and S. R. M. Prasanna, "An Overview of Digital Audio Steganography," *IETE Technical Review,* vol. 37, no. 6, pp. 632-650, 2020.

[36] B. Choudhury, R. Das and A. Baruah, "A Novel Steganalysis Method Based on Histogram Analysis: Lecture Notes in Electrical Engineering, vol 315," in *Advanced Computer and Communication Engineering Technology*, Switzerland, Springer International, 2015, pp. pp 779-789.

I. J. Computer Network and Information Security, **2020, 5, 56-73**
Published Online October xxxx in MECS (http://www.mecs-press.org/)
DOI:

Anthony Phipps is a member of the cyber research centre at London Metropolitan University, conducting research into audio based cyber security applications. His interests include steganography, the cyber security aspects of digital accessibility, and voice computing. He started his career as an engineer in electrical and electronic engineering. For over two decades he has specialised in Information and Physical Security in the energy and financial services sectors. He obtained his first degree in Electrical and Electronic Engineering from the University of Greenwich in 1997 and a masters degree from University of Westminster in Information Technology Security in 2002. He is currently a Lead Engineer for Lloyds Banking Group, managing a team of security design engineers.

Professor Karim Ouazzane is a Senior Professor of Computing and Knowledge Exchange, Head of Research and Enterprise, Chair of the European Cyber Security Council (Brussels), and founder of the Cyber Security Research Centre (CSRC), London Metropolitan University, London, UK. His research interests include artificial intelligence (AI) applications, Cyber security, bimodal speech recognition for wireless devices, big data, computer vision, hard and soft computing methods, flow control and metering, optical instrumentation and lasers. He has carried out research in collaboration with industry through a number of research schemes such as The Engineering and Physical Sciences Research Council (EPSRC), KTP, EU Tempus, LDA (London Development Agency), KC (Knowledge Connect) and more. He has also published over 150 papers, three chapters in books, is the author of three patents and has successfully supervised 25 PhDs.

Vassil Vassilev is a Professor in AI and Intelligent Systems and Head of the Cyber Security Research Centre at London Metropolitan University. Prof. Vassilev graduated with an MSc in Electrical Engineering from the Technical University in Sofia and obtained a PhD in Artificial Intelligence from the Bulgarian Academy of Sciences. He was active during the nineties in both academia and industry in his home country of Bulgaria. He founded the Department of Informatics at the New Bulgarian University and led several EU funded projects dedicated to the transformation of Higher Education in Bulgaria. After a couple of years in the offshore software development business, Dr. Vassilev joined the academic community in the UK with positions at the University of Northumbria, the Open University, and London Metropolitan University.

