Enhancing Cyber Security Using Audio Techniques: A Public Key Infrastructure for Sound

Anthony Phipps
Cyber Security Research Centre
London Metropolitan University
London, UK
arp0264@my.londonmet.ac.uk

Karim Ouazzane
Cyber Security Research Centre
London Metropolitan University
London, UK
k.ouazzane@londonmet.ac.uk

Vassil Vassilev
Cyber Security Research Centre
London Metropolitan University
London, UK
v.vassilev@londonmet.ac.uk

***Abstract* —This paper details the research into using audio signal processing methods to provide authentication and identification services for the purpose of enhancing cyber security in voice applications. Audio is a growing domain for cyber security technology. It is envisaged that over the next decade, the primary interface for issuing commands to consumer internet-enabled devices will be voice. Increasingly, devices such as desktop computers, smart speakers, cars, TV's, phones an Internet of Things (IOT) devices all have built in voice assistants and voice activated features. This research outlines an approach to securely identify and authenticate users of audio and voice operated systems that utilises existing cryptography methods and audio steganography in a method comparable to a PKI for sound, whilst retaining the usability associated with audio and voice driven systems.**

Keywords— Authentication, Steganography, Two-factor Authentication, Cyber Security, Audio Security

I. Introduction

Audio is a growing domain for cyber security technology. It is envisaged that over the next decade, the primary interface for issuing commands to consumer internet-enabled devices will be voice. Increasingly, devices such as desktop computers, smart speakers, cars, TV's, phones and Internet of Things (IOT) devices all have built in voice assistants and voice activated features. Already in the context of digital services, 50% of all adults have used voice for internet search and there are over a billion voice searches per month. [1] Powerful drivers such as accessibility, increased accuracy, device design (screen less and keyboard less devices), convenience and speed of communication will drive the trend for increased use of audio and voice as a channel to interact with information technology. [2] Additionally, verbal communication is much faster than the typical typing speed of the average person and recent advances in machine learning for voice recognition and biometrics have improved accuracy of this technology however, significant challenges remain to enable this technology in high security and high reliability environments. Despite the relatively inexpensive and non-intrusive nature of voice and audio methods of authentication, they are still relatively low performance in noisy environments. [3] At the outset, the purpose of this research was to investigate the new methods of identification and authentication of users accessing IT systems based on audio processing in a model that can contain factors such as something you have, something you know, something you are, and contextual information such as user ID, device, location, background sounds, health information, emotional state, combined with cryptographic information.

II. Motivation and Rationale

With voice increasingly becoming the interface of choice for users of information systems, security techniques must evolve. Many of today's authentication and identification solutions for voice channels (such as voice biometrics) have serious limitations in terms of security and usability. [4] New forms of attack are emerging that allow malicious actors to gain covert access top voice-controlled systems and assistants which are inaudible or incomprehensible to the human owners of such systems. [5]

Figure II-1 Increasing Application for Voice Control in Cars

As can be seen in *Figure II-1 Increasing Application for Voice Control in Cars* voice control can be as trivial as asking for music to play to more safety related information such as "are my brake pads still ok?" Research has drawn attention to serious limitations of voice only interactions with smart speakers and phones and the lack of command confirmation, voice authentication and any additional authentication factors. [6] [7] In addition, recent research has also shown it is possible to use 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. [8] Imagine this scenario: A remote laser triggering your car to come off autopilot whilst driving on the motorway. Even with the addition of a vocal confirmation prompt, the attacker could easily anticipate and add this.

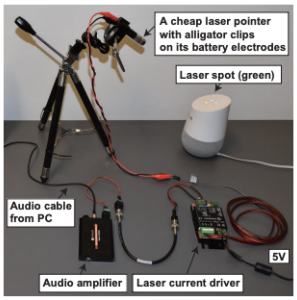


Figure II-2 Setup for low-cost laser attack on Google Home (Sugawara et al. 2019)

With cybercrime on the rise and increased regulatory requirements, there is demand for an uplift in the techniques and technologies available for users and consumers to interact with online services using voice and audio technology. Multi-factor voice driven authentication systems in high security environments such as online banking often have favourable perception with regards to security but at the expense of usability. [9] [10] In addition, there is a significant proportion of the world's population who struggle to use existing digital technology due to physical restrictions, cognitive abilities or social and financial limitations. This could include physical difficulty in handling and manipulating devices, visual impairment, physical pain, social exclusion or financial challenges. According to the UK Office for National Statistics (ONS), 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. [11] At the same time, Amazon sold hundreds of millions of devices equipped with their voice assistant software Alexa, and Google Assistant is now available on more than 500 million devices. Online security concepts such as authentication and identification can prove 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. [12]

What is therefore required is a model that can provide identification and authentication using audio techniques to compliment voice, that can eliminate usability challenges whilst improving overall security and accessibility of digital services.

III. Background Concepts

The research draws upon a number of key concepts which are briefly introduced:

A. Authentication

Authentication is the process of identifying or proving the identity of a user or process. When referring to the use of computer systems, authentication is the check that users are who they claim to be. Typically, authentication is categorised into one of three factors: something you have, something you know, something you are. A key part of ensuring that authentication is secure and consistent is to have a robust enrolment process that binds the user and the user identity.

B. Biometric Authentication

Biometric authentication can be said to be authentication by characteristic. Broadly speaking this is the concept of "something you are" and can be subdivided into things you do (behavioural) and things you are (physiological). Biometric techniques can be either static (like a fingerprint, iris) or dynamic (behavioural, voice, heart ECG) etc. Within the audio domain, the main method employed for biometric authentication presently is voice biometrics. The main two voice biometric techniques in use are speaker verification and speaker identification. [13]

C. Speaker Verification

Speaker-verification authenticates that 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. Key challenges of this method include considerations of background noise, health impacts on the speaker's voice, and the quality of the audio or telephony equipment involved in both the enrolment and authentication. Also, as the stored voiceprint and one captured at a later authentication will be different, the result of the authentication process is usually a matching score rather than a binary yes or no. [13]

There are various types of speaker verification:

Table III-1 Speaker Verification in Voice Systems (Markowitz 2000)

Verification Method	User Action
Text-dependant verification	User is prompted to enter username and speak a password
Text-dependent verification with speech recognition	User is prompted to say a specific phrase, account number or PIN
Text-prompted verification	User enters a PIN or password then respond to prompts to repeat words or numbers
Text-independent verification	Users voice is verified covertly

D. Speaker Identification

Speaker identification refers to the identification of an unknown speaker. This technique does not require the user to respond to specific commands or prompts.

E. Steganography

The word steganography is made up of the Greek words steganos, which means "covered," and graphia, which means "writing." [14] Unlike cryptography, the objective of steganography is to hide the existence of the message. Steganography is therefore another term for covert communication. It works by hiding messages in inconspicuous objects that are then sent to the intended recipient. [15] The most important requirement of any steganographic system is that it should be impossible for an eavesdropper to distinguish between ordinary objects and objects that contain secret data. In the digital realm, image steganography is the most common and audio steganography is the second most popular type of digital steganography. [16]

F. Audio Steganography

Audio steganography is a method of hiding information within sounds or within the soundtrack to a video clip. Its applications vary from creating covert communications channels, subliminal manipulation, digital watermarking and digital rights management for the copyright protection of digital media. A key challenge is that the Human Auditory System (HAS) is more sensitive in comparison to the Human Visual System (HVS) so techniques used must preserve a high degree of fidelity over the original audio signal. The HAS operates over an exceptional range greater than one billion to one and a range of frequencies of over a thousand to one. [27] When using digital media, the secret information that is hidden within an audio file is not limited to audio information, rather any digital files or information can be hidden. This technique is the main technique used in the experimental phase of this research.

G. Steganography Environments and Domains

There are various techniques that can be applied to hiding data. The first thing to consider is the transmission media and environments the audio signal will have to travel through between the encoding stage (where the hidden data or signal

is added) and the decoding stage (where the hidden information is then recovered and separated from the carrier). As described by Bender et al. [17] there are many different transmission environments to consider and four classes are shown in Figure III-1 Steganography Transmission Environments (Bender et al. 1996).

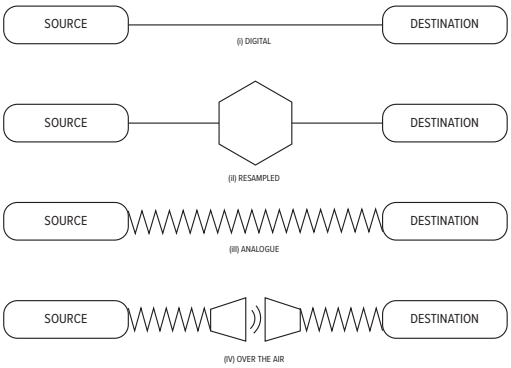


Figure III-1 Steganography Transmission Environments (Bender et al. 1996)

Consideration of the transmission environments and the relative challenges, advantages and disadvantages. In the first example (i), the environment is digital end to end and thus a sound file is copied from machine to machine without any modification. This method has the least constraints on the hiding techniques that can be used in the digital domain but has limited application and the sampling is exactly the same for both encoder and decoder. In the second example (ii), the signal is resampled. A key consideration here is if the sampling rate is higher or lower when resampled. This environment will preserve the signals magnitude and phase for the most part, but changes the temporal characteristics of the signal. In the third example (iii) the signal is transmitted in an analogue fashion. Absolute signal magnitude, sample quantisation, and temporal sampling rate are not preserved however phase is preserved. In the final transmission method (iv), the signal is played in an analogue state over an air gap and recaptured with a microphone. This is the most challenging transmission method as the signal will be subjected to drift, echoes leading to non-linear modifications, phase changes, different frequency components and general interference. The transmission method has a significant impact on the type and amount of data that can be hidden in an audio signal using steganographic techniques.

H. Voice and Audio System Attack Techniques

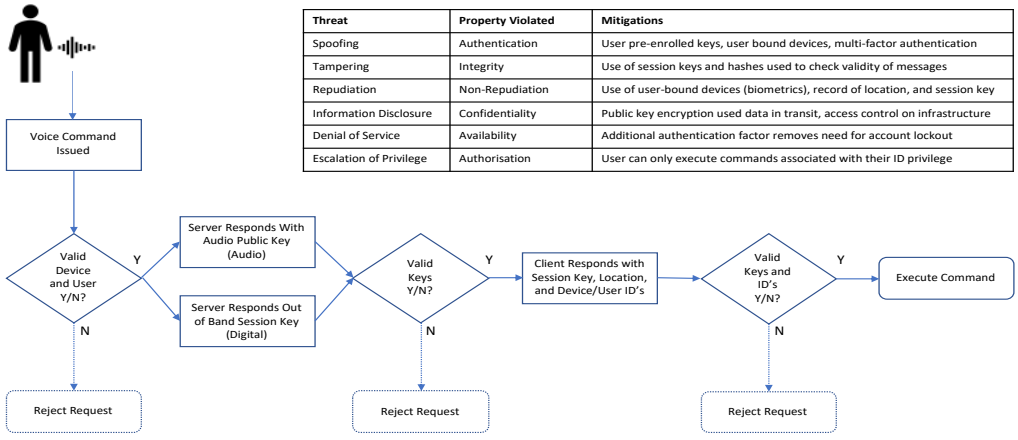
Voice and audio systems of authentication and identification can be attacked like any other cyber system. As attack techniques evolve, new counter-measures will be needed. Classes of attack include replay attacks, voice synthesis as well as attacks in the implementation of speech and audio interfaces like any other IT system or application. At a high level these can be classified thus:

Table III-2 Classes of Audio Attack

Attack Vector	Technique
Replay	Attacker injects covertly recorded voice sample
Voice Synthesis	Speech synthesis and/or text-to speech adapted to the characteristics of the target or brute force attack
Impersonation & Deep Fake	Attacker mimics the target for verification/authentication

IV. Developing a Conceptual Model

In assessing the security of voice and sound driven user journeys, the authentication process either relies on voice recognition, a binding of the device (authenticating the device rather than the user), or the verification of text as previously discussed. Given the issues described in the motivation and rationale section, the model development has centred on offering an additional authentication factor combined with minimal user friction. The model development is still underway but the security provided by the model developed thus far is the application of both existing public key cryptography techniques and new steganography and authentication paths. The novelty in this model using audio and an out of band path to complete the overall authentication. Work has commenced on understanding the threats using both the STRIDE model and will be developed further using intelligence graphs in policy form to ensure all threats are understood and quantified. [18] When looking at the objective of secure and trusted communication across an untrusted media, it is first helpful to consider the conceptual model of a public key cryptography system.



Threat	Property Violated	Mitigations
Spoofing	Authentication	User pre-enrolled keys, user bound devices, multi-factor authentication
Tampering	Integrity	Use of session keys and hashes used to check validity of messages
Repudiation	Non-Repudiation	Use of user-bound devices (biometrics), record of location, and session key
Information Disclosure	Confidentiality	Public key encryption used data in transit, access control on infrastructure
Denial of Service	Availability	Additional authentication factor removes need for account lockout
Escalation of Privilege	Authorisation	User can only execute commands associated with their ID privilege

Figure III-2 Development of the Security Model

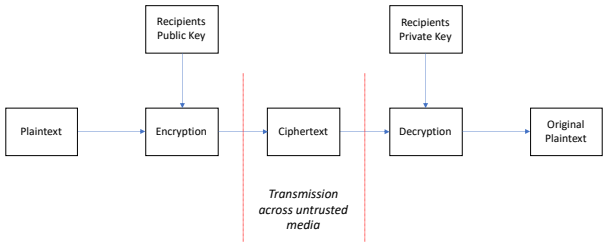


Figure IV-1 Public Key Cryptography Model

In this scenario the sender can send a secret message to a recipient without the need to exchange a secret key. Building on these concepts a number of scenarios are considered for applications where an audio interface is in use.

A. Audio Public Key Infrastructure – Use Case Scenario

The first scenario to consider is a proposal that audio sound can be used as a public key in authentication and encryption schemes. This would be desirable as a way of securing communications for applications involving audio interfaces and where low friction is required.

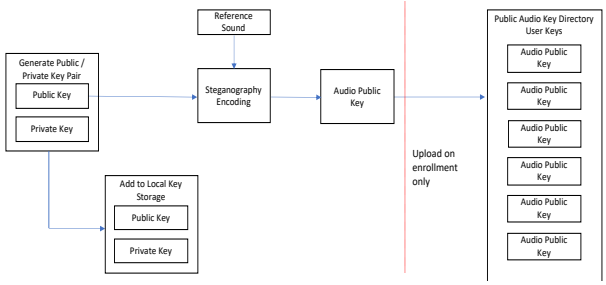


Figure IV-2 Novel Audio PKI Conceptual Model Enrolment Process

Firstly, to build an equivalent of a public key infrastructure, a number of processes and components are required. The end user needs to create a key pair and encode the public key into a sound using steganography. There then needs to be a way of enrolling into the scheme such that a recipient can lodge their audio public key with the equivalent of a Public Key Infrastructure (PKI) provider or Certificate Authority (CA). An enrolment process is shown in Figure IV-2 Novel Audio PKI Conceptual Model Enrolment Process. In this process, the public keys are hidden not to provide security, rather for the purpose of not interfering with the audio quality of the sound used to convey them. This is much the same process as a website offering its public key in a certificate in a browser, not as a set of ASCII values visible to the user but as a simple padlock. This obfuscation is about audio usability and the security in the model comes from the underlying cryptography and the out of band confirmation described in the model below. This approach offers a new avenue of audio user experience where the sounds used could be derived from secure sounding clips such as a safe closing, a padlock snapping shut or tech sound. Alternatively the sound used for a public key could actually be branded by the user or company offering the service adding a marketing or personalisation opportunity in a way that certificates or passwords cannot.

Once an infrastructure or cloud service provider has a directory of previously enrolled "Audio Public Keys" the question of how they could be used arises. If the objective is to minimise user authentication friction whilst maintaining or even enhancing security, a number of factors can be considered such as location of the user, verification of a user's enrolled device, authentication of the user and mitigating the impact of threats such as tampering, spoofing, interception, interjection and replay attacks. In Figure IV-2 Novel Audio PKI Conceptual Model Enrolment Process, a model is proposed that could be used to provide mutual trust between the provider or a voice assistant "skill" or service and the end user. The concept works as follows. 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

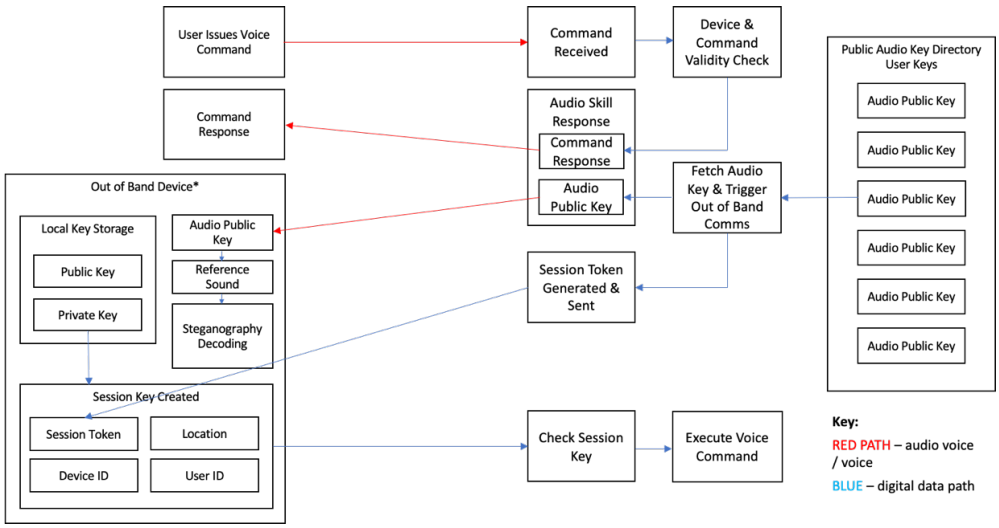


Figure IV-3 Novel Audio PKI Conceptual Model – Command Verification Process

confirmed, the voice assistant responds with two things. The command response 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 out of band device (this could be either a phone or IOT 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 (out of band – digitally) the voice assistant service provider can execute the command and any onward actions.

V. Experimental Work

The experiment was to conduct a lab test and initial analysis to evaluate a steganographic method which will be used to hide key material in sounds. The steganographic method under test was devised as part of earlier related research conducted into "Two-factor authentication for voice assistance in digital banking using public cloud services." [12] Early iterations of this work involved utilising a sound in the form of a ringtone as an additional authentication factor. Whilst that research for that project pivoted away to another direction by utilising an IoT "beacon" device, the challenge of utilising sound in this way remained unanswered. Java code was written to investigate the feasibility of audio steganography for this purpose. [19] This was further developed and was used as a basis to develop a testable lab with adjustable parameters. The purpose of the experiment was to investigate the ability to encode a secret message within a sound, transmit that message to another receiving machine and then decode or recover the message from the transmitted message. The experiment had two key stages, transmission of the sound via digital means and transmission via acoustic "over the air" means. During the development of the code under test, it was noted that environmental factors for testing audio steganography over the air were very susceptible to environmental factors. Consideration was given to facilities such as anechoic chambers and soundproof environments. Further investigation led to the decision for the first controlled over the air tests to be carried out in a recording studio which provides a practical yet controlled and neutral sound characteristic for the testing.

The experiment was also arranged to simulate potential use case scenarios envisaged for later stages of the research. The process from an operational perspective can be described as a series of steps as shown in the following charts:

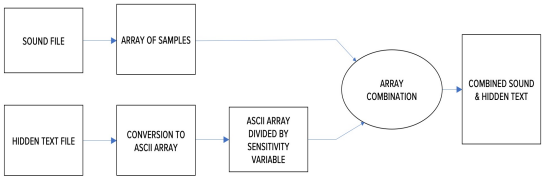


Figure V-1 Encoding Process

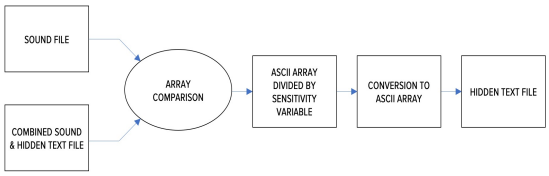


Figure V-2 Decoding Process

The depth of the encoding was the controlled variable used to assess the effectiveness of the encoding and decoding for this experiment. It was anticipated that the depth of encoding would impact reliability, detectability and quality of the sound. By adjusting the depth of encoding this balance was explored. In addition, the method was anticipated to present serious challenges for the recovery of encoded information when transmitted over the air due to factors such as background noise, audio processing, alignment of signals etc. The equipment was first set up as shown *in Figure V-3 Overview of Digital Environment for Audio Steganography Experiment*. The experimental code was run on each audio lab PC and the files were shared via a network storage drive (but could have been any digital transmission method that preserves file integrity).

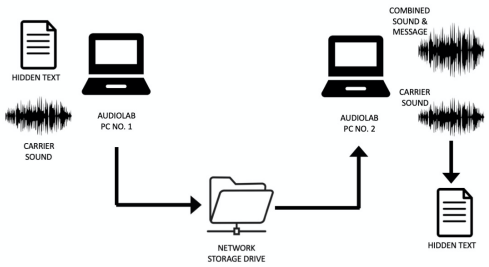


Figure V-3 Overview of Digital Environment for Audio Steganography Experiment

For the over the air testing, the setup was amended to incorporate the new method of sound transmission:

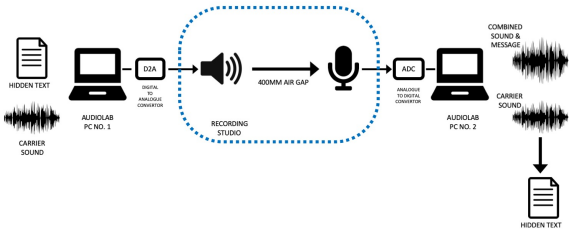


Figure V-4 Overview of "Over the Air" Environment for Audio Steganography Experiment

The equipment selected for this phase of the experiment was specifically selected to ensure the lowest possible noise, distortion or interference with the signals and sounds.

VI. EXPERIMENTAL RESULTS

Two sounds were used in the digital environment test and the over the air test. A music clip and a voice sample. The waveforms are shown in this section for later comparison purposes with the output waveforms which have both the original sound and the hidden encoded text. The testing was carried out by encoding sounds with a series of sensitivity settings using both voice and music samples. As expected, the decoding of the over the air lab test still requires further analysis. Waveform plots were exported from the Sonic Visualiser software. Each sound file was loaded in turn exported to a picture file with the same name as the corresponding source sound file. Once the sound files were loaded, the view was expanded such that the x axis time started at zero seconds and the amplitude of the wave fully visible in the picture. The first series of tests were run using the music sample input file, the digital transmission method, and set of sensitivity variable settings from 100 through to 500,000. The first test was then run by encoding the hidden file into music file, digitally transmitting it to another machine and decoding it. It was possible to recover the hidden text with no error or degradation with message "**this is a hidden message"** perfectly intact with no additional characters. Playing the output sound file however, it was possible to hear a noticeable click at the start of the sample playback which could alert a listener to the presence of hidden data, interference or be considered an error. As can be seen by plotting the output sound file, there is clearly visible deviations from the input sound in the sub 0.2 second range again potentially alerting to the presence of hidden information.

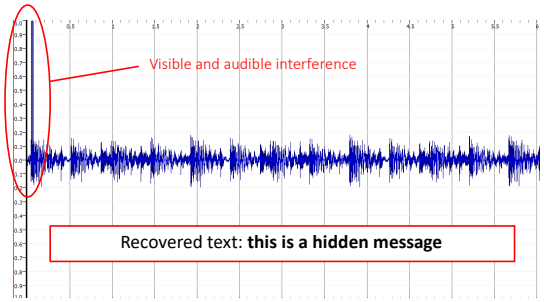


Figure VI-1 Digital Environment Encoded Music Sample – Sensitivity Variable 100

The next test was run as before but with an adjustment of the sensitivity variable to 1000.
As before it was possible to recover the hidden text with no error or degradation with message "**this is a hidden message**" perfectly intact with no additional characters. This time however there was no discernible interference or difference between the input music sample and the output sound file from an auditory perspective. In addition, when plotting the output sound file it was not possible to easily detect the presence of the hidden information or interference.

Figure VI-2 Digital Environment Encoded Music Sample – Sensitivity Variable 1000

The next four tests were with the sensitivity variable set at 50,000, 75,000, 100,000 and 500,000. Although there was no auditory degradation of the sound and no visible artefacts when plotting the sound, none of the hidden text was recoverable as shown below.

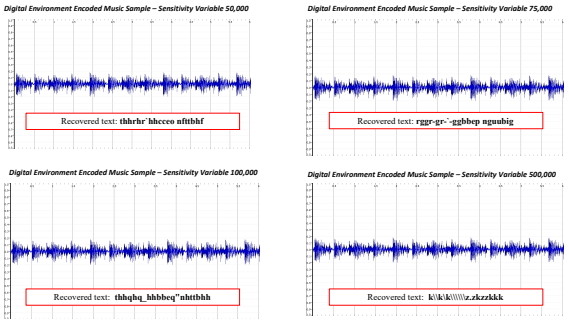


Figure VI-3 Digital Environment Encoded Music Sample – Sensitivity Variable 50k - 500k

A number of attempts were made to recover the hidden text. By increasing the value of the sensitivity variable, it did not allow for any text recovery above the 10,000 mark however the text recovered each time was different. Due to the way the encoding works, the higher the sensitivity variable, the smaller the difference between the reference input sound and the encoded sound so it is expected behaviour that there is a point where this difference is undetectable, and the method of encoding becomes unfeasible. The tests were repeated with a voice sample.

The results of decoding the hidden message in the digital domain tests are shown below. The tables show for each sensitivity variable setting, the outcome of the decoding process, the decoded output itself, and a subject assessment of whether the encoding process impacted the audio signal or was clearly visible in the simple time x amplitude plots. For both the voice and the music sample, the decoding of the hidden text ceased at sensitivities above 10,000.

Table VI-1 Music Sample Tests Results Table

Music Sample Tests				
Sensitivity Variable	Decoded OK?	Decoded Output	Visual Interference?	Audio Interference?
100	Y	this is a hidden message	H	Y
1000	Y	this is a hidden message	L	N
5000	Y	this is a hidden message	L	N
10000	Y	this is a hidden message	L	N
50000	N	thhrhr`hhcceo nfttbhf	L	N
75000	N	rggr-gr-`-ggbbep nguubig	L	N
100000	N	thhqhq_hhbbeq"nhttbhh	L	N
500000	N	k\\k\k\\\\\z.zkzzkkk	L	N

The same procedure was repeated using a voice sample and the results gathered below:

Table VI-2 Voice Sample Tests Results Table

Voice Sample Tests				
Sensitivity Variable	Decoded OK?	Decoded Output	Visual Interference?	Audio Interference?
100	Y	this is a hidden message	H	Y
1000	Y	this is a hidden message	M	Y
5000	Y	this is a hidden message	M	N
10000	Y	this is a hidden message	L	N
50000	N	thhrhr`hhccenlerr`fe	L	N
75000	N	rggr gr ` ggbben-lerr`ge	L	N
100000	N	thhq"hq" "hhbben"keqq ee	L	N
500000	N	zkkz.kz.k.kkkkkz.zkzzkkk	L	N

It was easier to audibly detect interference in the voice sample. As this manifests itself as a short click, it was hard to hear any interference in the music samples. It was noted that the samples that failed to decode (at sensitives above 10,000) had a high incidence of similar characters. At sensitivities 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. It was shown that for digital transmission, an aggressive setting leads to interference whilst a subtle setting leads to decode failure and therefore a working range was established. Overall these results reveal a robust and usable technique if tuned correctly.

A. Embedding Key Material for User Experience

To prove this technique would apply to the overall model developed, further testing was carried out embedding an RSA public key. A stated before, the purpose of the steganography is to embed the key in a pleasant sound, rather than a sound analogous to a modem chatter which would be unpleasant on the ear. The security overall is provided by the model and the application of *existing* public key cryptography techniques. The novelty in this model using audio and an out of band path to complete the overall authentication.

A 2048 bit key pair was generated for the purpose of the test. To confirm the optimum working range of the test was repeated between 100 and 100,000 in 100 increments. The result was then checked, character by character and the percentage of key match to the original recorded. As the scheme in its basic form needs 100% accuracy, the working range can clearly be seen in the plot below:

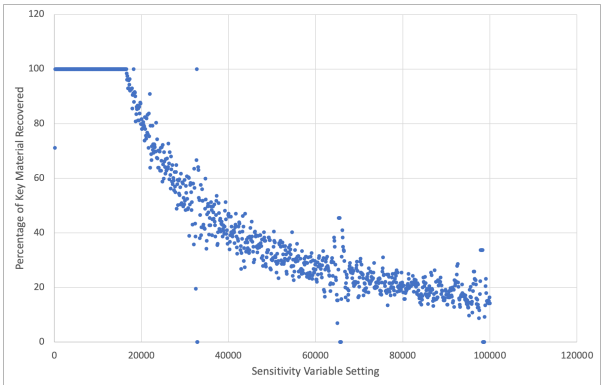


Figure VI-4 Percentage of Recovered Key Material in Audio Sample Test

B. Over the Air Environment Tests

The files that were used for the digital environment test that had the hidden information encoded in them, were each played in turn through the monitor speaker and a second separate laptop with microphone connected via the analogue to digital convertor was used to record the sounds.

The environment was also controlled from an acoustic viewpoint to minimise interference and distortion. The first tests used the music samples; the second set of tests used the voice samples.

The over the air testing showed when the sensitivity variable was set at the more aggressive settings, it could be visibly and audibly detected in the experimental data. Attempts to manually align the original sound and the captured sound manually failed to result in any signals being successfully decoded. This is as expected and during the structured research the scale of the challenge of both aligning the signals for decoding and then compensating for the different amplitudes has to be addressed with some additional processing.

During the experiment, the input and output signals are similar, they would require a degree of normalisation for the technique to work in an over the air environment. Normalisation would require alignment, attenuation or amplification, comparison and correction. Initial attempts to manually achieve this using the graphical tools in Sonic Visualiser and Audacity were not successful. Using techniques such as perceptual hashing, bit error rate mapping and dynamic warping distance (using the empirical data to drive structured research) may deliver a solution and is planned for later research.

Also, to assess detectability of the hidden signal, spectrogram plots were also made. The plots were made using sonic visualiser. Spectrograms are a visual representation of the spectrum of frequencies present in the sound sample as it varies over time. The purpose of plotting the spectrograms were to compare the input sound files with the output sound files (with hidden text) to see if there are visibly any difference between the two. It was anticipated that as the method of encoding was based upon modifying the amplitude of parts of the sound rather than changing the sound itself, that no visible difference would be visible between input and output frequencies present in the digital domain tests. With the over the air environment tests, it was envisaged that some very minor differences would be present due to noise however this would be minimal due to the carefully controlled environment of the lab test.

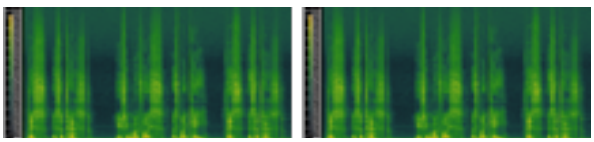


Figure VI-5 Comparison of Input and Output Voice Spectrogram with Sensitivity Variable 1000

For all the samples observed, there were no significant differences between the input and output file spectrograms with the exception of where the sensitivity variable was set lower than 1000. As observed in the time domain, the "spike" seen at the most aggressive sensitivity is as follows:

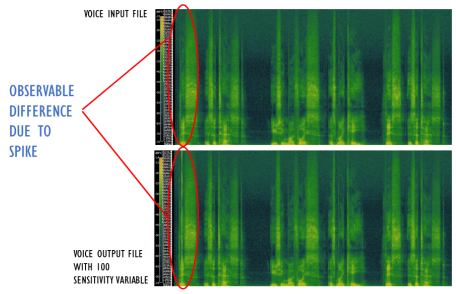


Figure VI-6 Comparison of Input and Output Sample (100 Sensitivity)

VII. Summary Experimental Findings

This lab test generated useful insight into the future direction of the research and in particular the feasibility of this steganography technique in a sound PKI technique. More testing is required to establish the practical soundness of the overall method. The technique as developed so far puts the hidden information in the first few milliseconds of a sample. It also has the benefit of not increasing the size of the overall sound file and if tuned correctly has no impact that can be detected by the human auditory system (an essential quality for audio steganography). It is also simple to implement in its current form and works effectively in the digital domain providing the sensitivity is set correctly. This in itself is a useful finding in its own right. The technique however requires further development if it is to be successful in "over the air" applications. If the sound is to be transmitted "over the air" as an audio signal, the lab test showed that the technique does not yet work without additional processing which will inform the next steps of the research.

The findings from this lab test will also be useful in the scope and development of a conceptual framework of how this technique might be used in a wider context.

VIII. Conclusions & Future Research Direction

Devices that use voice authentication and voice user interfaces continue to grow in acceptance and use. With more smart home devices, cars, voice assistants, smart phones and IOT devices using audio channels for input, the importance of the security and performance of such systems will continue to grow. In this research a new conceptual model for authenticating users has been proposed that uses audio steganography to transmit key material. A path for future research has been outlined as well as establishing a new digital steganography technique. By testing both digital and analogue (over air) transmission methods, the analogue aspects have proved more of a challenge. Future experimental work to be undertaken will include decoding the "over the air" signals using techniques such as perceptual hashing, bit error rate mapping and dynamic warping distance for feature extraction. The focus of the research will continue to be concentrated on finalising the conceptual model utilising empirical and experimental data collected in the next phase of the research. It has been possible using the literature review, analysis and experimentation to begin the creation of a conceptual model – a PKI for Sound.

IX. Acknowledgment

The research reported here was supported with a grant of UK Government under Cyber ASAP Programme of Innovate UK "Beacon-based Authentication" [12] and was initially carried out at the Cyber Security Research Centre of London Metropolitan University with contributions from several members of the Centre - Matthew Lane, Siddhartha Natarajan, Khalid Mohamed, Artur Naciscionis, Viktor Sowinski-Mydlarz and Dr. Mona Ibrahim and has continued as a PhD research project focussing on audio security.

X. References

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

[2] S. Kinkiri, W. Melis and K. Simeon, "Machine learning for voice recognition," in *The Second Medway Engineering Conference on Systems: Efficiency, Sustainability and Modelling, University of Greenwich*, 2017.

[3] K. S. Adewole, A. S. Olaniyi and R. G. Jimoh, "Application Of Voice Biometrics As An Ecological And Inexpensive Method Of Authentication," *International Journal of Science and Advanced Technology,* vol. 1, no. 6, pp. pp 196-201, 2011.

[4] 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].

[5] 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].

[6] 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.

[7] 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.

[8] 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].

[9] 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.

[10] 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].

[11] 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].

[12] 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.

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

[14] D. Khan, The Code-Breakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet, New York: Scribner, 1996, pp. 131-132.

[15] R. J. Anderson and F. A. Petitcolas, "On The Limits of Steganography," *IEEE Journal of Selected Areas in Communications,* vol. 16, no. 4, pp. pp. 474-481, 1998.

[16] J. Fridrich, Steganography in Digital Media: Principles, Algorithms, and Applications, Cambridge: Cambridge University Press 2009, 2009, pp. pp 3-13.

[17] W. Bender, D. Gruhl, N. Morimoto and A. Lu, "Techniques for Data Hiding," *IBM Systems Journal,* vol. 35, no. 3&4, p. 323, 1996.

[18] V. Vassilev, V. Sowinski-Mydlarz, P. Gasiorowski, K. Ouazzane and A. Phipps, "Intelligence Graphs for Threat Intelligence and Security Policy Validation of Cyber Systems," in *Proceedings of International Conference on Artificial Intelligence and Applications*, New Delhi, India, 2020.

[19] S. Natarajan, *Audio Steganography - Project Submission,* London: London Metropolitan University , 2018.

[20] H. Özer, B. Sankur, N. Memon and E. Anarım, "Perceptual Audio Hashing Functions," *EURASIP Journal on Advances in Signal Processing,* vol. 12, pp. 1780-1793, 2005.

[21] J. Lyons, "http://practicalcryptography.com/," 2013. [Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. [Accessed 28 February 2020].

[22] P. Nair, "The dummy's guide to MFCC," Medium, 24 July 2018. [Online]. Available: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd. [Accessed 28 March 2020].

[23] V. Tyagi and C. Wellekens, "On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, 2005.

[24] L. E. J. Frenzel, Handbook of Serial Communications Interfaces, Newnes, 2016, pp. 229-232.

[25] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978.

[26] Ricardo Portilla; Brenner Heintz; Denny Lee;, "Understanding Dynamic Time Warping," Databricks, 30 April 2019. [Online]. Available: https://databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html. [Accessed 27 March 2020].