Your Password Is Music To My Ears: Cloud Based Authentication Using 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 in progress into identifying and addressing the threats faced by voice assistants and audio based digital systems. The popularity of these systems continues to grow as does the number of applications and scenarios they are used in. Smart speakers, smart home devices, mobile phones, telephone banking and even vehicle controls all benefit from being able to be controlled to some extend by voice without diverting the attention of the user to a screen or having to use an input device such as a screen or keyboard. Whilst this removes barriers to use for those with accessibility challenges like visual impairment or motor skills issues and opens up a much more convenient user experience, a number of cyber security threats remain unanswered. This paper details a threat modelling exercise and suggests a model to address the key threats whilst retaining the usability associated with voice driven systems, by using an additional sound based authentication factor.**

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

I. Introduction

More than 200 million smart speakers have been sold in a trend that is showing year on year growth with Amazon's Alexa powered devices accounting for over 100 million units. [1] Other devices such as mobile phones, laptops, home automation products and even cars are exploiting the benefit of providing the user with convenience of speech interaction. The ability to control technology without the distraction of a screen or the use of a tactile input device such as a touch screen or keyboard and mouse make voice a great choice for interactions demanding low cognitive load and no motor skills required on behalf of the user. One of the early applications that has been widely exploited by voice interaction for internet-based devices is search, with an estimated 50% of all adults having used voice for internet search, and over a billion voice searches per month. [2] Other growing uses include receiving the news and weather, streaming music and controlling smart home devices such as lighting and heating. Improvements in the accuracy and responsiveness of voice-based systems continues as better machine learning models are developed and the cost of deployment continues to fall. [3]

Despite the trends that are leading to wider adoption and increased usage, there are security and usability issues still to be addressed when considering this technology for high security applications. Many of the current implementations for voice channel interaction have limited command verification, authentication and identification and the present solutions that attempt to address security concerns have limitations that compromise usability. [4] 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. [5] [6] 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 includes 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. [7] This research sets out to investigate current threats, usability limitations and propose a new method of authentication for users of smart speakers and other digital audio-based systems. The research has outlined a number of threats and through threat modelling outlines a set of mitigations and a conceptual model that provides an additional factor based on sound, for authentication based upon the novel combination of existing technologies such as audio steganography, cryptography, GPS and cloud technology.

II. Background

The most popular voice-controlled systems on the market today are provided across multiple platforms and devices. Amazon Alexa, Microsoft's Cortana, Apple's Siri and Google Assistant are delivered into phones, televisions, laptops, cars and smart speakers. Despite differences in branding and capabilities, they all have common factors and there are moves afoot to standardise the understanding of their architectures. [8] Whilst not all of these systems are the same, the key components used in a typical architecture are as follows:

A. End User Device

The end user device or client is simply the device listening for commands or a "wake" word. In the most basic form, these require at least one microphone, a speaker, power source and internet connectivity usually provided by WiFi. With most implementations speech is turned into text for onward processing. Where interactions require it, this device also responds with audio, generated speech and occasionally visual prompts. Management of dialog and commands is usually specified in pre-determined intents.

B. *Cloud Based API*

The end user device typically connects to a cloud service provision where requests, commands and actions are sent to be processed. These are usually to a tight specification that ensures only the desired actions are permissible and executed.

C. *Voice Assistant Service Provider*

The provider of the service will also provide features such as user account management, user and/or device authentication, natural language processing, provision connections to a third-party application or service provider and an ecosystem that might involve user selectable applications or "skills" to consume and use.

D. *3rd Party Application Service Provider*

Some voice-controlled systems and assistants provide access to third party skills and intents. As well as language processing this might require further API's to service the requested service, data storage or integration with other applications.

E. *Companion App*

It is common for voice assistance and voice-controlled systems to have a companion app. Typically these provide a method of enrolment, access to more complex set up and configuration parameters and integration tools.

In addition to the components used, it is helpful to consider a number of services and concepts upon which these voice controlled systems and voice assistants rely:

F. *Authentication*

Authentication is the process of identifying or proving the identity of a user, device or process. Traditionally, authentication is categorised into something you have, something you know, something you are. An important consideration when devising an authentication scheme is ensuring that the authentication is secure by having a strong, robust and secure enrolment process that binds the user, device or process to the digital version of their identity.

G. *Biometric Authentication*

Biometrics requires the comparison of a person's intrinsic and unique physical or behavioural attributes with a stored value or representation of that attribute. Biometrics require a robust enrolment process to ensure the stored attributes are linked to a digital identity. The stored attributes also need careful protection from manipulation, tampering or theft. 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. [9]

H. *Speaker Verification*

Speaker-verification is the process that authenticates a claim that a person is who he or she is by comparing the speakers voice with a database of reference voiceprints captured during a enrolment process. To be reliable, a speaker verification system must be able to deal factors such as environmental noise and health driven changes to a users voice. To deal with this the enrolment process may require an extensive number of samples. As a result the process is a statistical matching exercise rather than a binary yes or no. [9] There are various types of speaker verification:

Table II-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 responds to prompts to repeat words or numbers
Text-independent verification	Users voice is verified covertly

I. *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 and is more concerned with identifying different speakers in a conversation or for providing enhanced personalisation.

III. Current and Emerging Threats

Automatic Speaker Verification (ASV) and voice biometrics offer a very low friction method of authenticating and controlling devices. There are still however some considerable issues and challenges to overcome before they can be deployed in high security environments and applications. In the UK, a BBC investigation into the use of voice biometrics by reporter Dan Simmons uncovered an issue where his telephone banking account service was accessed by his non-identical twin brother. The reporter enabled his HSBC bank account for access via his voice and it was then subsequently accessed by his brother who after seven failed attempts, managed to access the account by mimicking his brothers voice. The system also allowed for repeat attempts making the job of would-be attackers easier. [4] Many of the challenges and issues encountered by the deployment of ASV and voice biometrics are due to the low maturity of undying technology and can be attacked like any other cyber system. 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. [10]

Table III-2 Classes of Audio Attack

Attack Vector	Technique
Replay Attack	Attacker replays a 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 using trained models in order to be able to respond to verification challenges
Covert and side channel attacks	Attacker uses light, ultrasound, infrasound or fragments of sound to intiate covert commands
Technical Exploitation	Attacker exploits a vulnerability in the implementation of the system

Thanks to the improvements in machine learning, natural language processing and AI, the current state of the art systems are vulnerable to text to speech attacks using freely available tools. The quality of the output of synthetic speech has increased, and the amount of data required to train synthetic voices has decreased. Demonstrations of how such attacks can be carried out using modest resources have been shown at hacking conferences and published by Seymor et al. [11] 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." [12] Another method demonstrated by Seymore et al. which was more robust to factors such as noise and interference was utilising open source tools Tacotron [13] [14] and Wavenet [15] and utilising voice samples freely available to the public. Given samples of sufficient quality audio it was concluded that voice authentication is relatively easy to subvert. New forms of attack are emerging that allow malicious actors to gain covert access to voice-controlled systems and assistants which are inaudible using ultrasonic sounds [16] or in comprehensible to the human owners of such systems using non-sensical word sequences which are interpreted as commands. [17] In addition, recent research has also shown 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. [18]

Table III-1 Simplified STRIDE Analysis of Conceptual Model

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

IV. Developing a Threat Model

In order to address the security concerns associated with the current and emerging threats, simple threat modelling is required to drive the requirements of a secure audio-based authentication system. Threat modelling not only helps understand and prioritise the threats to be addressed but also outlines mitigation strategies that assist in the development of a conceptual model leading to a potential system implementation. [19] One challenge to this approach is that when using the modelling as a design tool for a model, we have to consider a generalised view to existing systems and models as these are many and varied, but with many vulnerable to the same vulnerabilities and classes of attack. For most implementations, the authentication process for voice-controlled systems relies on a binding of the device during enrolment to a service and subsequently authenticating the device rather than the user, voice recognition, or the verification of text as previously discussed. Work has commenced on understanding the threats using both the STRIDE model and is shown in Table III-1 Simplified STRIDE Analysis of Conceptual Model and will be developed further using intelligence graphs in policy form to ensure all threats are understood and quantified. [20]

V. Developing a Conceptual Model

To address the threats, a new model has been developed that has centred on offering an additional authentication factor combined with preserving the ease of use of voice. The model developed thus far is the application of both existing public key cryptography techniques and a new audio steganography technique. When looking at the objective of secure and trusted communication across an untrusted media, it is first helpful to

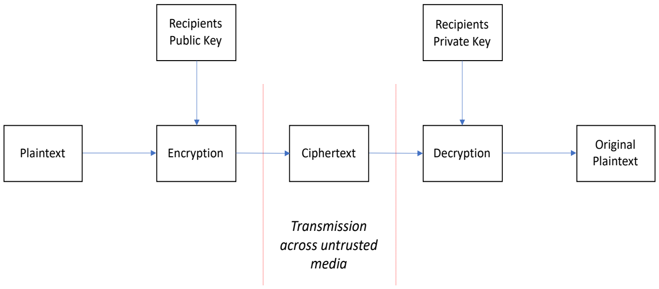

Figure V-1 Public Key Cryptography Model

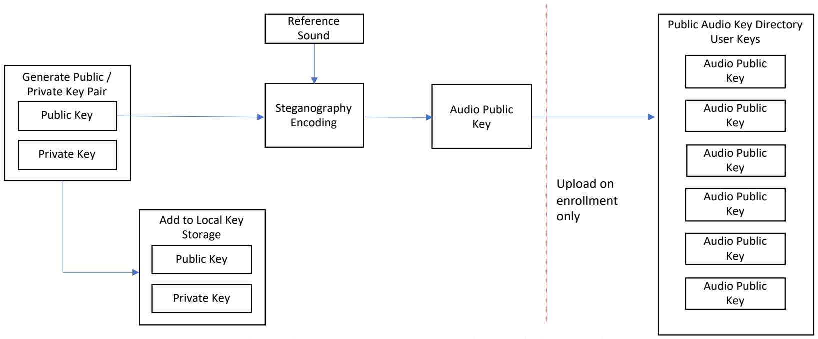

Figure V-2 Novel Audio PKI Conceptual Model Enrolment Process

consider the conceptual model of a public key cryptography system as illustrated in Figure V-1 Public Key Crypto Model. In this scenario the sender can send a secret message to a recipient without the need to exchange a secret key. To build an equivalent of a public key infrastructure with a sound based password, 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 V-2 Novel Audio PKI Conceptual Model Enrolment Process

The novelty in this model using audio and an out of band path such as a sound coming from the users mobile phone to complete the overall authentication. 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 as a public facing website which offers its public key in a certificate in a browser not as a set of ACSII 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 music 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 V-3 Novel Audio PKI Conceptual Model – Command Verification 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 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

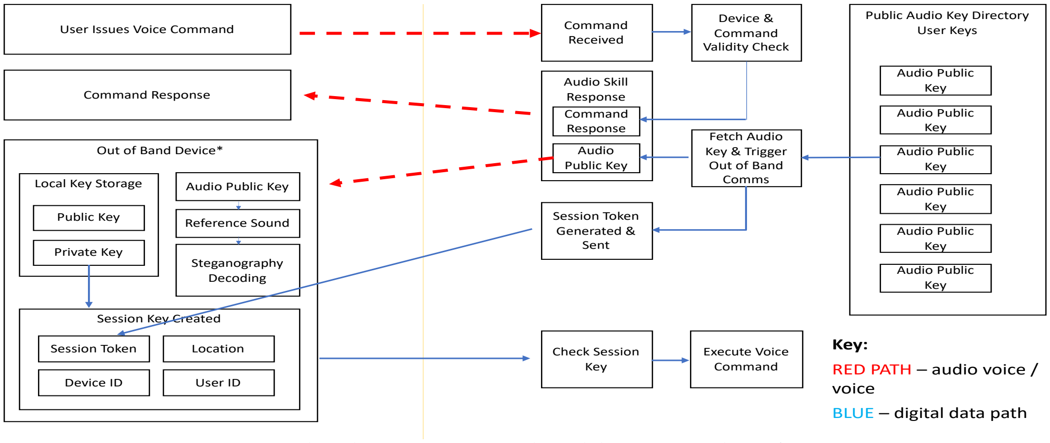

Figure V-3 Novel Audio PKI Conceptual Model – Command Verification Process

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.

VI. Experimental Work

Eexperimental work is in initial stages and has been started to evaluate the proposed model. The steganographic method used to hide the public keys in sounds was devised as part of earlier related research conducted into “Two-factor authentication for voice assistance in digital banking using public cloud services.” [21] [22] This was developed further using MATLAB and a series of tests undertaken as part of a plan to further develop and validate the model with adjustable parameters. The purpose of the experiment was to investigate the ability to encode an RSA 2048 bit public key within a sound, transmit that message to another receiving machine and then decode or recover the message from the transmitted message. 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.

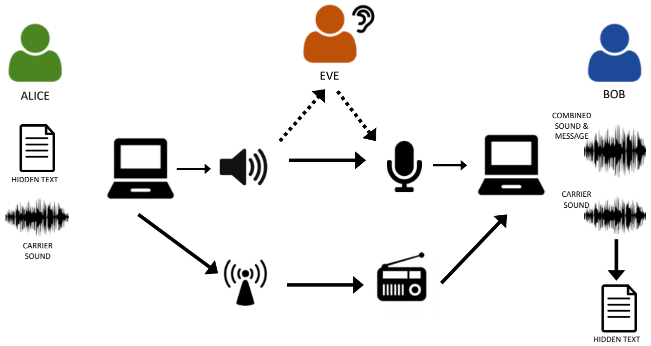

Figure VI-1 Overview of Steganography Key Encoding Test

In the experiment, the key is hidden information in the first few milliseconds of a music sample. It was found to have 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 on its own. The technique however requires further development if it is to be successful in “over the air” applications. 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.

VII. Conclusions & Future Research Direction

With the market for voice driven services growing and the security concerns rising, the demand for extra low friction authentication will increase. This will be especially important if such devices are to gain widespread acceptance in high security applications such as online banking or healthcare. In this research a new conceptual model for authenticating users has been proposed that uses audio steganography to transmit key material. The path for the next phase of research has been outlined for testing and refining the model. 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 for dealing with real world noisy environments. This paper has outlined the vulnerabilities faced by voice controlled systems and outlined preliminary research into both a conceptual model and future experimentation into using audio passwords to secure voice controlled systems.

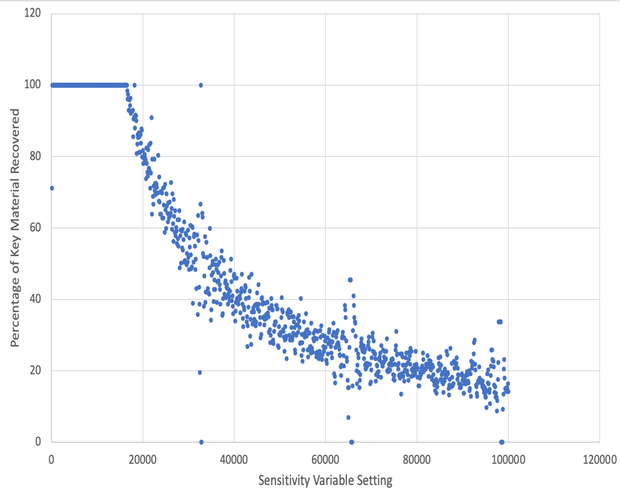

Figure V-4 Percentage of Recovered Public Key Under Test

VIII. Acknowledgment

The initial research that formed the foundation of the first phase of research was supported with a grant of UK Government under Cyber ASAP Programme of Innovate UK "Beacon-based Authentication" [21] 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.

IX. References

[1] G. Sterling, "Marketing Land," Third Door Media, 14 February 2020. [Online]. Available: https://marketingland.com/more-than-200-million-smart-speakers-have-been-sold-why-arent-they-a-marketing-channel-276012#:~:text=Some%20might%20invoke%20the%20cliche,skills%20%E2%80%94%20with%20no%20breakout%20hits.. [Accessed 13 November 2020].

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

[3] 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.

[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] 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.

[6] 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].

[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] 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].

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

[10] 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.

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

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

[13] 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].

[14] 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.

[15] 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].

[16] 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.

[17] 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].

[18] 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].

[19] A. Shostack, Threat Modeling: Designing for Security, Indianapolis: Wiley, 2014.

[20] 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.

[21] 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.

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

[23] 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.

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

[25] 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.

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

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

[28] 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.

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

[30] 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].

[31] 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.

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

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

[34] 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].