Adversarial Voice Recognition: Protection Against Cyber Attacks

Vijaya Babu Kuchipudi[1], Dr. Harsh Lohiya[2], Dr. Laxmaiah Mettu[3]

[1]Research Scholar, Dept. of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore, Bhopal-Indore Road, Madhya Pradesh, India.
[2]Research Guide, Dept. of Computer Science and Engineering, Sri Satya Sai University of Technologyand Medical Sciences, Sehore, Bhopal-Indore Road, Madhya Pradesh, India.
[3]Research Co-Guide, HOD. Dept. of Computer Science and Engineering, CMR Engineering College, Kandlakoya (V), Medchal, Hyderabad.

Abstract

This study presents new configurations of the Internet of Things that use reinforcement learning. The NIST National Vulnerability Database (NVD) independently assessed voice-activated devices at 7.6 out of 10, which is an alarming risk factor. Our investigation of inaudible assaults on these devices validates this. A scenario is shown in our basic network model where an attacker gains unauthorised access to sensitive information on a protected laptop by using inaudible voice instructions. By running a battery of attack simulations on this basic network model, we were able to demonstrate how easily privileged information can be discovered and owned via physical access on a large scale, all without the need for additional hardware or enhanced device capabilities. After testing six different reinforcement learning algorithms in Microsoft's CyberBattleSim framework, we settled on Deep-Q learning with exploitation as the best option, which allowed us to quickly take possession of all nodes with little effort. Because of the proliferation of mobile devices, voice activation, and non-linear microphones that are vulnerable to stealth attacks in the near-ultrasound or inaudible ranges, our research highlights the urgent need to comprehend non-conventional networks and develop new cybersecurity measures to protect them. Since the inaudible attacks originate in the microphone design and digital signal processing, there may be more digital voice assistants than humans by 2024, and the only way to address them is via standard patches or firmware updates.

INTRODUCTION

There has been a meteoric rise in the popularity of speech-activated gadgets like digital voice assistants in the last several years. In 2020, there were over 4.2 billion of these assistants in use worldwide. By 2024, researchers expect that number would have surged to 8.4 billion, exceeding the total population. These numbers demonstrate how integrated this technology is into our daily lives and the vast unrealized potential it has for experiments in cybersecurity and AI-powered personal assistants. More and more consumer electronics come with virtual assistants built in. These assistants facilitate human-device interaction by handling voice requests, delivering information, and controlling other linked devices.

The pervasiveness of these voice-activated gadgets makes them a fascinating test subject for studying their cybersecurity. With their widespread integration and use, these gadgets provide a one-of-a-kind chance to study cybersecurity—a discipline that is becoming more important as these devices penetrate our lives—through investigation, design, and testing. Research into possible cybersecurity risks and the development of strong defences in this setting is an attractive and pressing issue.

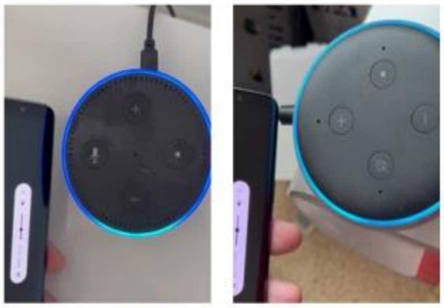

Figure 1. NUIT2 Attacks from Android to Echo Dot Gen 2/3 Devices: "Alexa, what's the weather?"

Related work

Approach to Exploring the Inaudible Command Response Surface

In this study, we model certain C&C attacks on voice-activated systems (VAS). Virtual assistants (VAS) such as Alexa, Siri, Google Assistant, and Cortana from Microsoft are often criticised for not having biometric identification or for failing to authenticate users. The makers of these handy gadgets make them react to voice orders, so anyone within range who can imitate the commands might theoretically use them. These VAS devices can't properly differentiate between speakers, with the exception of iPhones with biometric capabilities. As a result, there aren't any safeguards in place to ensure that only authorised users may deactivate security systems, make unsolicited calls, send messages, or access users' calendars and accounts. According to the ecosystem, the programme permits supplementary abilities or gadgets, such as remote access to actual homes or financial transactions.

Unbeknownst to the user, a second significant attack vector takes use of this authentication hole. Without the need for external amplification, the built-in audio mics can pick up barely audible orders and carry them out. Because it does not need any specialised hardware and takes advantage of a flaw in the microphones that convert near-ultrasonic commands into vocal instructions, this stealth attack greatly increases the danger.

Brief Summary of Previous Inaudible Attack Research

Xia, Chen, and Xu introduced a novel kind of attack in early 2023; they dubbed it Near-Ultrasound Inaudible Trojan (NUIT) and demonstrated how an attacker may use the speaker to compromise a device's microphone [2-3,5]. Among fifty sample orders, 84% of the time the Amazon Echo Dot was able to recognise the wake word and 58% of the time it was able to carry out the stealth instructions as an inaudible trojan.

The DolphinAttack project by Zhang et al. proved in theory that bespoke hardware can control VASs, but it also need less covert specialised amplification gear to activate orders. As an example of how these assaults are becoming more sophisticated, Pandya, Borisaniya, and Buddhadev came up with ShoutIMEI, an ultrasonic covert channel-based attack that targets Android devices. As an additional entry point for exploiting VAS, Mitev, Miettinen, and Sadeghi investigated the potential of skill-based man-in-the-middle attacks on virtual assistants such as Alexa from Amazon.

On top of that, Amazon Alexa on Echo Dot contains serious flaws including command injection, which highlights the severity of these dangers as High Risk, 7.6/10 without known fixes. The inclusion of this attack type in the MITRE ATT&CK architecture (T1202) highlights its continued significance in modern cybersecurity research. It is also known as indirect command execution. Additionally, this has been labelled as a "Common Weakness Enumeration: CWE-77: Improper Neutralisation of Special Elements used in a Command ('Command Injection')"by the MITRE Corporation. A thorough examination of mobile malware was conducted by Pattani and Gautam. Their research included an all-encompassing overview of mobile device security risks and solutions. The research found that covert channels are a major problem in mobile security, especially when coupled with VAS vulnerabilities, which may cause major security lapses.

Adversarial Agents, Reinforcement Learning Frameworks and CyberBattleSim

In the same way that AI-powered voice assistants have profited from advances in speech recognition, we suggest investigating hostile agents using AI-powered simulations. We use CyberBattleSim, developed by Microsoft, as our platform to study primarily lateral motions across various connections between VAS nodes. An open platform that uses reinforcement learning (RL) to simulate and analyse complex cyber assault scenarios is the CyberBattleSim environment, which was created by the Microsoft Defender Research Team. It is a controlled laboratory for studying the dynamics of cyberattacks and testing defence techniques under different situations; its primary purpose is to study the behaviour of automated systems in a restricted network environment. To sum up, CyberBattleSim portrays attackers using agents created using RL algorithms. While evading detection, these agents travel laterally and explore a network graph to take control of nodes . According to the algorithm comparison by Shashkov et al. , this method lays a strong foundation for the advanced research of adversarial agent learning in cybersecurity. Their research presents strong evidence in favour of RL's use to cybersecurity, particularly its ability to reveal hidden entry points and build effective defences.

Kunz et al. provided a Multiagent Cyber Battle Sim for RL Cyber Operation Agents, which is a significant addition to the development of the Cyber Battle Sim platform. Their method highlighted the importance of RL in navigating the cyber battlespace, which becomes even more intricate and deep when several opposing agents are involved. To augment the multi-agent method, Piplai et al. suggested a cyber-attack and defence knowledge-guided two-player RL model. By including a guided RL approach in a two-player situation, their model brought an additional level of intricacy to CyberBattleSim, showcasing the dynamic interaction between attackers and defenders. The use of RL offers a novel and potent framework for studying cyber defences and threats. A plethora of studies using RL platforms highlights the promise and need of delving into the AI-enabled cybersecurity environment, particularly studying advanced threats like inaudible attack surfaces.

METHODOLOGY

The CyberBattleSim is modelled in this study as an RL environment that is compatible with OpenAI's gym architecture. A novel strategy in the field of cybersecurity is the Near-Ultrasound Inaudible Trojan (NUIT). These kinds of attacks use a device's audio capabilities for nefarious purposes, thereby fusing the device's source and target into one. A common NUIT1 situation is a hostile action triggered by an inaudible order, which is beyond the hearing range of humans but may be detected by the device's microphone. A wide range of potentially disastrous outcomes may result from the stealth command, including but not limited to data theft, unauthorised access, and more. An additional layer to this danger is introduced by NUIT2, or NUIT-N, assaults. Here, each device serves as a broadcaster and receiver; hence, NUIT2 represents two devices, whereas NUIT-N has one hosting broadcaster and N receivers. The order is sent by the host, which might be a compromised machine, and then understood and carried out by the receiver, which is another device that is within the ultrasound's range. These assaults may create a complex and evasive attack surface by jumping from device to device, so evading conventional network barriers and safeguards.

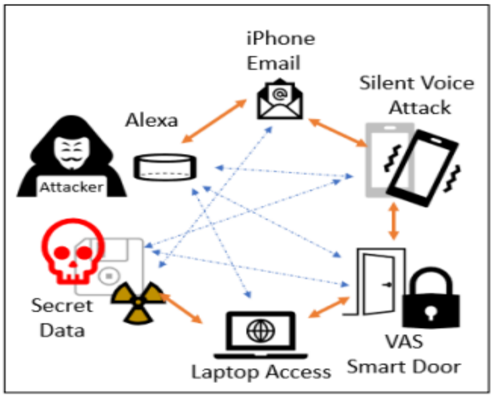

Figure 2. Baseline RL NUIT Scenario to Demonstrate Attack Gaining Secret Information. The orange lines show the best path derived from RL, and the blue lines delay the attack without progressing to the end goal.

An efficient method for simulating such complex and detailed situations is the CyberBattleSim (CBS) environment. In a home network, an Amazon Echo Dot might be the target of a NUIT2 assault, which CBS could model in great detail. A malicious skill is downloaded into the Echo Dot by the attacker, who may have issued the instruction via an inaudible broadcast from an infected device nearby. Figure 2 is an example of how an attacker's malware may be designed to trick users into giving their email credentials. Then, the malware can be enabled by utilising instructions that cannot be heard. Baseline an Attack on RL NUIT to Showcase the Acquiring of Confidential Data. The blue lines stall the assault before reaching the final objective, whereas the orange lines depict the optimal route determined by RL.

"Alexa, enable evil downloader skill called Bank of Bitcoin." After the hacker has access to the victim's email, they may use the stolen credentials to get access to the victim's iPhone. The hacker tricks the victim into giving them control of their iPhone by sending an email with a NUIT1 attack attachment. After taking control of an iPhone, an attacker may use a NUIT1 assault to open a smart door lock, which is not a typical node in this case. If this happens, the assailant may physically enter the building and take a sensitive laptop. The fact that they are able to use the computer they have access to extract important state secrets shows how an apparently harmless audible instruction may lead to a major security violation. The importance of defences that are both inventive and nimble in responding to changing threat environments is shown by these simulations. They show how smart speakers and cellphones, which aren't usually thought of as part of the conventional network, may really be entry points for sophisticated cyberattacks. In order to promote more strong and all-encompassing cybersecurity tactics, CyberBattleSim offers a platform to research, comprehend, and handle these difficulties. As seen graphically in Figure 2, our basic simulation environment incorporates a five-node scenario. We simulate a malicious Alexa skill capable Amazon Echo Dot (node) to illustrate the attack scenario. To begin with, we include a collector—an email account—to determine whether an iPhone is present in the signature line (Find Device Type in Email, remote attack). Then, we collect data from emails and go laterally on the network. A phishing email instructs the next remote assault to unlock the smart door using NUIT1, assuming the attacker knows they can target an iPhone (node). In this scenario, the smart door (node) is the next asset to be modelled. It opens in response to inaudible instructions and exposes physical theft techniques for stealing a classified laptop, including remote attacks. The assailant now has physical control of the sensitive computer system, allowing them to access sensitive data via remote access (Access State Secrets).

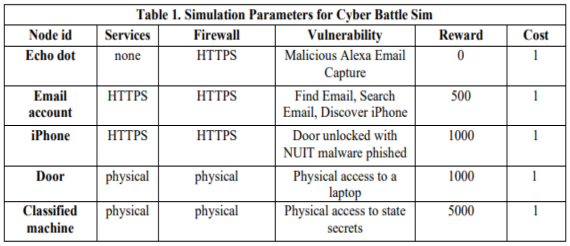

Table 1. Simulation Parameters for Cyber Battle Sim

Node id	Services	Firewall	Vulnerability	Reward	Cost
Echo dot	none	HTTPS	Malicious Alexa Email Capture	0	1
Email account	HTTPS	HTTPS	Find Email, Search Email, Discover iPhone	500	1
iPhone	HTTPS	HTTPS	Door unlocked with NUIT malware phished	1000	1
Door	physical	physical	Physical access to a laptop	1000	1
Classified machine	physical	physical	Physical access to state secrets	5000	1

We construct the network using the example that comes with Microsoft's model Capture-the-Flag simulation in order to set the settings of CyberBattleSim for NUIT. The baseline inaudible attack network's simulation parameters are shown in Table 1. CyberBattleSim's primary target audience is business networks. This is why our model incorporates non-standard nodes that have 'physical' services and firewall restrictions. To specify the desired route across the environment, taking into consideration both physical and virtual interactions, certain enhancements are necessary.

One of the main results of the simulation is that it finds a reasonable strategy to maximise attack rewards. When we compare the depth of conquest (the number of nodes owned at intermediate stages) with the speed of conquest (all nodes owned) as measured in epochs (simulated time) steps, we may rate the complexity of the assault configuration. Increasingly complex network topologies either can't be solved or need more time for simulation. There are five nodes in the baseline, and each of them is shown in many states. From the perspective of the attackers, the agent begins in a network with unknown attributes and connections, goes through phases of discovery and ownership, and then performs successful or unsuccessful operations at the nodes that it has specified in its settings for properties and node ages. From simple credential theft to more complex physical intrusions like data or equipment theft, the agent has nine potential exploits at their disposal. Unauthorised physical entry to a secured smart door is one example of how random actions at a specific node might lead to undesirable results. To help RL (and learning algorithms) strike a balance between exploration and exploitation, this tiny network provides a rich combinatorial environment.

RESULTS AND DISCUSSION

After all nodes have been found (stage 1) and owned (stage 2), the final status of the CyberBattleSim is summarised in Table 3. It is important to keep in mind that the cybersecurity environment is always changing and becoming more complicated. Tools like as CyberBattleSim can assist in modelling, studying, and developing defences for these complex assault scenarios.

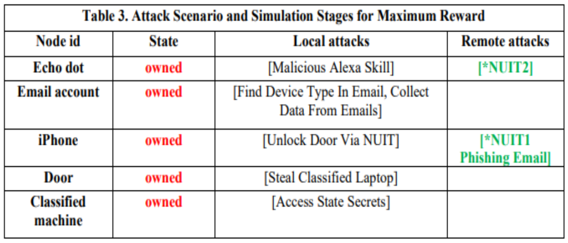

Table 3. Attack Scenario and Simulation Stages for Maximum Reward

Node id	State	Local attacks	Remote attacks
Echo dot	owned	[Malicious Alexa Skill]	[*NUIT2]
Email account	owned	[Find Device Type In Email, Collect Data From Emails]	
iPhone	owned	[Unlock Door Via NUIT]	[*NUIT1 Phishing Email]
Door	owned	[Steal Classified Laptop]	
Classified machine	owned	[Access State Secrets]	

This multi-stage assault demonstrates how different devices, which aren't often thought of as belonging to a network, may be connected to allow for a substantial breach. The resulting approach fails to address the issue of low-reward exploration and fails to achieve its exploitation aims. If an attacker carelessly gains access to the smart lock on the door from the email node, for instance, it defeats the aim of the NUIT attack and squanders their time. A common consequence of simulations including Internet of Things devices and VAS in a typical contemporary house highlights the exponential combinatorics of permitting a wide attack surface. An attacker's attack surface grows exponentially as the number of VAS-enabled devices on a network increases—for example, when Alexa, Siri, Google, and Cortana are all connected. Every VAS node is a breeding ground for new cyber-physical assets and vulnerabilities, as well as a proliferation of access and enabled skills. In this context, we see a cyberattack scenario that involves five separate nodes, all of which interact in a complex network that is crucial to the attack's overall execution. As previously shown, the MITRE ATT&CK matrix corresponds to the assault phases, whereas the MITRE D3FEND matrix represents the defensive remedies.

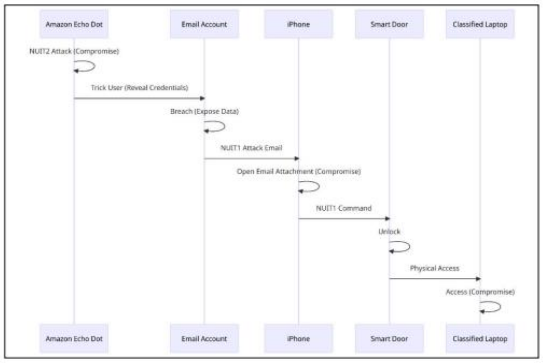

Figure 3. Sequence diagram of the baseline NUIT Attack model under test

• **Echo Dot:** An Amazon Echo Dot is the first victim of a NUIT2 attack, which causes it to download a malicious skill.

Email Account: Once an intruder has access to an Echo Dot, they may steal users' email credentials by using a deceptive skill. After this happens, the user's email account is compromised,

• **iPhone:** after gaining access to the email account, the hacker attaches a NUIT1 attack to an email. The user's email is sent to their iPhone via syncing. The user's iPhone gets infected when, unknowingly, they open the attachment and activate the NUIT1 payload.

• **Smart Door**: Once the intruder has control of the iPhone, they may activate the door by sending another NUIT1 instruction to the smart lock. Keep in mind that the smart door is an unconventional network node, yet it plays a crucial role in this attack chain anyway.

• **Laptop:** An intruder may physically reach the premises, where a classified laptop is kept, thanks to the open door.

The assailant now has access to this laptop, which holds sensitive information. Once the attacker has access to the Echo Dot, they use MITRE ATT&CK (T1202) to execute commands indirectly. These approaches prioritise maximising exploitation over exploring atypical paths or non-local maxima, successfully working out an assault plan and executing the found strategy.

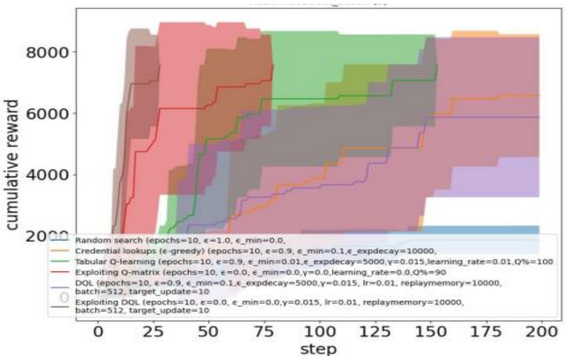

Figure 4. Effects of varying RL algorithm types and rewards accumulated over time. Better algorithms like Deep-Q Learning (DQL) accumulate rewards rapidly (upper left) compared to the worst performer as random search (blue line below rewards of 2000).

While epsilon-greedy provides a happy medium between the two extremes, random search is great at exploring but terrible at exploiting. Both DQL and Tabular Q Learning can achieve this balance, but it all depends on the hyperparameters set in the learning phase's search strategy. Nevertheless, when making choices using the learnt Q-matrix or DQL model, taking advantage of optimum actions maximise rewards. Because it is inefficient, random search struggles to scale as networks become larger and learning speeds increase. Because there are so many state action pairings,

Although these abstract network qualities are useful for comprehending non-conventional (disconnected) networks, they also pose a potential limitation of the Cyber Battle simulation. Figure 4 shows the results of several kinds of RL algorithms and the effects of cumulative rewards. As seen in the top left, better algorithms like Deep-Q Learning (DQL) gather rewards swiftly, while the weakest performance, random search, accumulates awards at a slower rate (blue line below 2000 rewards). vulnerability assessment and assaults on networks. Its relative inflexibility to other phases of an assault, such as reconnaissance and exfiltration, and its limited coverage of lateral motions constitute a second limitation. Our simulation relies heavily on reconnaissance, which we represent as a critical flaw in the VAS that allows an attacker to learn other assets' email addresses and passwords or signature lines. The CyberBattleSim on atypical networks allows for quick coverage of potentially susceptible network topologies in a home IoT scenario, while these simulations maintain an abstract state-action space.

After applying a carrier frequency of 16 kHz to normalise the residual audio signal, the generator raises the input signal's audible range to a higher frequency, approaching ultrasonic (16–22 kHz). For instance, in Figure 4, we may see the original and altered signals in the time domain (waveform) and the frequency and power density domain (spectral). Figure 3 shows spectrograms of the audible and inaudible signals, as well as the archiving process after speech identification. A brief overview of the steps involved in the NUIT process is as follows: the input audio signal is read, filtered, modulated to a near-ultrasonic frequency range using SUSBAM, normalised, processed using a Tukey window, converted to integer format, and finally written to an output WAV file.

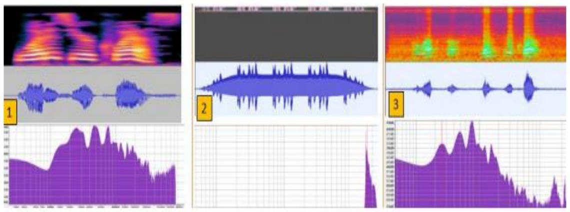

Figure 5. How the server (3) interprets inaudible broadcasts (2) transformed using NUIT from the original audible signal (1)

Figure 5. How the server (3) interprets inaudible broadcasts (2) transformed using NUIT from the original audible signal (1)

CONCLUSIONS

Because they are vulnerable to inaudible assaults, voice-activated gadgets provide a large and growing attack surface, which we tackled in this study. These devices provide complex security difficulties because to the dynamic topologies of the network environment, which includes geographical and temporal outliers as well as ephemeral targets. We used quantitative methods to evaluate the efficacy of stealth orders and found that wake word recognition had a high success rate, whereas command and control execution had a large but lower rate. This kind of assault, which is mostly focused on hardware, may do serious harm to devices that are deployed remotely.

To successfully traverse the ever-changing network topologies, we pioneered the use of reinforcement learning environments such as CyberBattleSim. In such settings, one may explore new nodes, assess risks, and develop mitigation strategies while also making sense of the combinatorial expansion of device kinds and software talents.

REFERENCES

[1] Statista (2023), number of digital voice assistants in use worldwide from 2019 to 2024 (in billions),

[2] Xia, Q., Chen, Q., & Xu, S. (2023). Near-Ultrasound Inaudible Trojan (NUIT): Exploiting Your Speaker to Attack Your Microphone, UseNix Security 2023.

[3] Pattani, K., & Gautam, S. (2021, March). A Comprehensive Study on Mobile Malwares: Mobile Covert Channels—Threats and Security. In Virtual International Conference on Soft Computing, Optimization Theory and Applications (pp. 91-102). Singapore: Springer Nature Singapore.

[4] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017, October). Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security (pp. 103-117).

[5] Microsoft Defender Research Team. (2021). CyberBattleSim. github. com/microsoft. Team, M. D. R. (2021). CyberBattleSim. Created by Christian Seifert, Michael Betser, William Blum, James Bono, Kate Farris, Emily Goren,

[6] Shashkov, A., Hemberg, E., Tulla, M., & O'Reilly, U. M. (2023). Adversarial agentlearning for cybersecurity: a comparison of algorithms. The Knowledge Engineering Review, 38, e3.

[7] Esteban, J. (2022). Simulating Network Lateral Movements through the CyberBattleSim Web Platform (Masters Thesis, Massachusetts Institute of Technology).

[8] Kunz, T., Fisher, C., Novara-Gsell, L., Nguyen, C., & Li, L. (2023). A Multiagent CyberBattleSim for RL Cyber Operation Agents. arXiv preprint arXiv:2304.11052.