BarrierBypass: Out-of-Sight Clean Voice Command Injection Attacks through Physical Barriers

Payton Walker
prw0007@tamu.edu
Texas A&M University
College Station, Texas, USA

Tianfang Zhang
tz203@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Cong Shi
cs1421@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

Nitesh Saxena
nsaxena@tamu.edu
Texas A&M University
College Station, Texas, USA

Yingying Chen
yingche@scarletmail.rutgers.edu
Rutgers University
New Brunswick, New Jersey, USA

ABSTRACT

The growing adoption of voice-enabled devices (e.g., smart speakers), particularly in smart home environments, has introduced many security vulnerabilities that pose significant threats to users' privacy and safety. When multiple devices are connected to a voice assistant, an attacker can cause serious damage if they can gain control of these devices. We ask where and how can an attacker issue *clean* voice commands stealthily across a *physical barrier*, and perform the first academic measurement study of this nature on the command injection attack. We present the BarrierBypass attack that can be launched against three different barrier-based scenarios termed *across-door*, *across-window*, and *across-wall*. We conduct a broad set of experiments to observe the command injection attack success rates for multiple speaker samples (TTS and live human recorded) at different command audio volumes (65, 75, 85 dB), and smart speaker locations (0.1-4.0m from barrier).

Against Amazon Echo Dot 2, BarrierBypass is able to achieve 100% wake word and command injection success for the across-wall and across-window attacks, and for the across-door attack (up to 2 meters). At 4 meters for the across-door attack, BarrierBypass can achieve 90% and 80% injection accuracy for the wake word and command, respectively. Against Google Home mini BarrierBypass is able to achieve 100% wake word injection accuracy for all attack scenarios. For command injection BarrierBypass can achieve 100% accuracy for all the three barrier settings (up to 2 meters). For the across-door attack at 4 meters, BarrierBypass can achieve 80% command injection accuracy. Further, our demonstration using drones yielded high command injection success, up to 100%. Overall, our results demonstrate the potentially devastating nature of this vulnerability to control a user's device from outside of the device's physical space, and its limitations, *without* the need for complex and error-prone command injection.

WiSec '23, May 29-June 1, 2023, Guildford, United Kingdom

ACM ISBN 978-1-4503-9859-6/23/05…$15.00
https://doi.org/10.1145/3558482.3581772

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

IoT, speech recognition, physical barrier, command injection attack

ACM Reference Format:
Payton Walker, Tianfang Zhang, Cong Shi, Nitesh Saxena, and Yingying Chen. 2023. BarrierBypass: Out-of-Sight Clean Voice Command Injection Attacks through Physical Barriers. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec '23), May 29–June 1, 2023, Guildford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3558482.3581772

1 INTRODUCTION

As voice assistant (VA) devices such as Amazon Echo and Google Home smart speakers are approaching ubiquity, we are forced to become more aware of the inherent security risks associated with these devices. VA devices typically act as a central hub of control for a multitude of connected smart devices such as smart locks, lights, cameras, thermostats, appliances, and garage doors. Each of these devices can be controlled in some way by issuing voice commands to the VA device. But these commands also introduce new types of risks. The ability to control such devices with vocal commands opens up a lot of attack possibilities that did not exist before. Among the different types of attacks that can be performed, the potential for home/office/hotel/dorm intrusion is one of the most severe and threatening.

Media coverage on this subject reveals the growing concern for the security vulnerabilities in a smart home environment [13, 24, 25, 29, 30]. While much of the concern is centered around the vulnerability to hacking that comes with connecting a multitude of devices, many professionals agree, for the purposes of home intrusion, there is a very low chance that an attacker would attempt to perform complex hacking as opposed to simply brute forcing there way in [14]. However, the ability to issue simple vocal commands to a voice assistant in order to control a lock or door is one vulnerability that requires no hacking and could potentially be favored by attackers who want to gain access to a space.

Aside from commands being accidentally issued through television advertisements [9, 26, 35], in an ISTR special report from Symantec, the author discusses the "mischievous man next door attack" which involves a neighbor issuing voice assistant commands

either with ultrasonic frequencies, or by waiting until you leave and simply shouting a command through the door [34]. The report touches on the significant security risk that is introduced if you have smart locks or a garage door that can be controlled by your voice assistant because it would allow an attacker to gain entry into your home. While home invasion is a serious concern when a command injection attack is possible, it is important to note that there are many other scenarios that can cause harm if an attacker can control the smart devices of a home. For example, turning on the stove can cause a gas leak or become a fire hazard.

Another form of command injection attack that has emerged in academia in recent years is hidden voice commands that obfuscate command audio so it is unrecognizable to humans, but recognizable by VA devices. However, hidden voice command attacks have limited applicability and their accuracy is generally low. They are also very sensitive to noise because of how specially they are crafted to begin with. Also, even after the past several years of research on these attacks [8], the vendors have not really come up with defenses to such attacks. This is perhaps because the vendors are likely ignoring them as being rather impractical or uneventful. Indeed, the recent work by Abdullah et al. [8] revealed that many of the hidden voice command attacks presented in research are not truly feasible in real-world settings due to their low accuracy and lack of transferability to different systems. Another recent work on command injection by Sugawara et al. [31] introduces the LightCommands attack which uses laser-based injection of the audio signal. The main drawbacks to this attack are that it requires a line of sight, is very complicated to setup/launch, and can be error prone.

In this paper, we aim to address most of the aforementioned problems with the existing voice command injection attacks from the literature or practice. We focus on an attack model, BarrierBypass, in which a loudspeaker issues *clean vocal commands* — through a physical barrier — to a voice assistant or other voice controllable technology that is located inside a home, office, or hotel room. While this attack model eliminates many of the complications of hidden command injection, it does introduce its own limitations. For example, because this attack injects clean commands and requires louder volumes, the attack would likely only be launched in certain scenarios such as when the user is not present in the space (such as during work hours). We consider three different barrier types which serve as the entry points for the attacker to inject such out-of-sight voice commands:

(1) ***Window Barrier***: The attacker injects a command through a window to target a voice assistant in the room. The attacker can launch this attack in-person or remotely-controlled via drone technology which can target multiple homes in a neighborhood or even high rise buildings with condos or offices. We demonstrate the feasibility of both attack scenarios in this work.

(2) ***Door Barrier***: The attacker injects a command through a door that connects to the space with the victim voice assistant. This barrier is likely most susceptible due to the *thin gap* beneath the door above the flooring which is sufficient for the sound waves to pass into the space easily.

(3) ***Wall Barrier***: The attacker is located in an adjacent space and injects a command through an interior wall. This barrier is applicable to housing setups such as dorms, hotels, or apartments where adjacent units share a wall.

Is issuing voice assistant commands across a physical barrier possible? What types of barriers can be attacked? How can such an attack be achieved and what particular settings are required in order to bypass the barrier? What are the limitations of this attack? These are the main research questions that we consider during this work and seek to answer. We perform extensive experimentation to evaluate the BarrierBypass attack in different parameter settings such as command audio loudness and location of the voice assistant device to determine when this type of command injection attack is possible in a real-world scenario. To our knowledge, a broad study on the voice assistant command injection attack, across physical barriers, has yet to be conducted in academia. This work demonstrates when the BarrierBypass attack is practical and it can be used to inform future research directions on the subject.

Main Contributions and Results: We summarize our key contributions and results below:

(1). Design of Clean Voice Barrier-based Attacks: We designed three different barrier-based command injection attacks to represent common materials/objects that may act as a physical barrier between an attacker and the victim's voice assistant during a command injection attack. Specifically, we define the BarrierBypass attack in the *across-door*, *across-window*, and *across-wall* scenarios and assess the effect of each barrier type on the attack's success. We present an attack that circumvents the sophistication and complexity of hidden voice command or laser-based command attacks, achieving the same goal with high accuracy in certain scenarios.

(2). Measurement Study Evaluating the Effect of Multiple Parameters: We present a measurement study and conduct an array of experiments to evaluate the effect of different barriers, under different attack settings, on command injection attack success. We test different speakers, loudness levels, voice assistant models (Amazon Echo Dot 2 and Google Home mini), device distances from the barriers, and observe the effect of different across-wall constructions (with and without insulation). BarrierBypass is able to achieve 100% injection accuracy for both the wake word and command under certain conditions and selecting the highest performing speaker.

(3). Demonstration of Drone-based Attack: We utilized two drone models equipped with Bluetooth speakers to demonstrate the potential for executing the BarrierBypass attack via drones. Our experimental simulations of the attack reveal high command injection success when using a drone that has a low operating loudness, or when the command audio is increased by 10 dB to compensate for a higher operating loudness.

(4). Informed Suggestions to Increase Attack Robustness and Defense Potential: Compiling the knowledge gained from our multiple experiments and attack demonstrations, we devise a set of suggestions that could be applied by an attacker in order to improve the potential for this attack under realistic conditions. Conversely, this information can be used to inform defensive mechanisms.

2 BACKGROUND

2.1 Sound Passage Through Barriers

As sound waves hit a physical barrier, they will lose energy and attenuate as they pass through the solid material. This occurs because the sound is either reflected off of the material (causing echo) or absorbed by it. Therefore, sounds on one side of a barrier played

at a particular loudness (decibel) level, will be quieter when heard or recorded on the other side because the decibels are reduced. The transmission loss of sound across a barrier can be affected by many factors attributed to the barrier's material and construction. Thickness, density, and air space within the barrier are all factors that can either increase or decrease the level of sound transmission. For example, in double paned windows, thicker glass and greater air space in the middle are desired to optimize sound blockage [2]. A barrier's ability to block sound is measured using different rating values such as Sound Transmission Class (STC) and Noise Reduction Coefficient (NRC). We provide further detail on these values and what they represent in the following subsections.

2.2 Rating Values for Sound Propagation

Sound Transmission Class: Sound Transmission Class (STC) is an established rating system for how much sound is blocked by a particular assembly [3]. It is an integer rating that roughly equates to the dB reduction in sound across a particular barrier. For example, a wall that reduces a 100 dB noise on one side, to a 60 dB noise on the other side would have an STC rating of 40. It is the most commonly used metric in the US for describing sound blockage potential and allows for direct comparison between different products (i.e., walls, doors, windows, etc.) and manufacturers. Specifically, the STC rating is calculated as the average noise blockage, in dB, for 18 different frequency values and has a logarithmic scale. This rating is based on the ASTM E413-16 standard [4].

Since our work is mostly concerned with the amount of sound that is able to persist through a barrier and into the space on the other side, the STC rating is most relevant. The STC ratings for the different barrier setups that we consider include: STC of 20 for the door-barrier [27], STC of around 33 for the window-barrier [2], and STCs of 30 and 34 for the wall-barrier without insulation and with insulation, respectively. We will revisit these values later on when interpreting our experimental results.

Noise Reduction Coefficient: The Noise Reduction Coefficient (NRC) measures the amount of noise that a material absorbs [6]. Where the STC is a rating that describes how much noise can pass through a barrier, the NRC describes the amount of noise that is left within a space. Therefore, two materials with the same NRC does not imply that the same amount of noise is transmitted through the other side for each of them. NRC values are on a scale of 0 to 1, where 0 indicates the material will reflect back all of the sound that hits it, and a value of 1 indicates that all of the sound is absorbed by the material (e.g., none of it is reflected back). The NRC provides a single-value approximation of the noise absorption of a material by averaging the sound absorption coefficient values at four 1/3 octave frequencies (250, 500, 1000 and 2000 hertz) and is rounded to the nearest 0.05 increment. This rating is based on the ASTM C423-17 standard [5].

3 ATTACK & THREAT MODEL

In this section we define three BarrierBypass attacks based on different types of barriers (Door, Window, Wall), depicted in Figure 1, as well as describe our threat model.

3.1 Barrier-Based Attacks

Across-Door Attack: The first barrier that we consider is a standard interior door. We define the *across-door* attack to represent all situations where an attacker may attempt to inject a command to a victim's VA that is located across an interior door. If the door is locked, hindering the attacker from gaining direct access into the room, there is still the potential for the attacker to issue a command across the door barrier in order to achieve their goal (i.e., unlock the door's smart lock or control some other connected smart device). In this situation, the gap that exists between the bottom of the door and the floor can be considered a vulnerability that may be exploited by this attack. The presence of a small gap will significantly increase the audio propagation in the room and increase the potential for attack success.

Across-Window Attack: The next barrier that we consider is a standard window. We define the *across-window* attack to represent the more likely attack situation that an attacker is attempting to issue a command from outside the victim's home or office. Often the attacker will have no access to desired space, or even to an adjacent room, so issuing a command through a window may be their only option. Again, if the user can issue a command from this location, they may be able to gain access by issuing commands to other smart devices that are linked to the voice assistant (i.e., smart locks on the doors, smart garage door). The window used in our experiments was a builder's grade, double-pane window that was located on the balcony of a third floor apartment.

Across-Wall Attack: The last barrier that we consider in this study is an interior wall. We define the *across-wall* attack to represent the situations where an attacker may be in an adjoining room. This would be a common barrier for attackers in adjacent living arrangements such as apartment complexes, dorm rooms, or hotels. An attacker could easily set up the speaker equipment for their attack in their own space next door and not be disturbed. To allow for greater experimental control, we decided to simulate the across-wall scenario using a soundproof box and wall inserts that we constructed. We consider two typical constructions of interior walls that are still present today 1) without insulation and 2) with insulation. The details on the construction of the soundproof box and the wall inserts are provided in Sections 4.2 and 4.3, respectively.

3.2 Threat Model

In our threat model, the attacker does not need prior knowledge of the target VA device or its settings. Through a process of initial testing with different wake words, an attacker can learn what device is in the victim space and how to activate it (e.g., the Amazon Echo only has four possible wake word settings so each could be tested). Also, depending on the placement of the target device, the attacker could look through a window of the target room (either in person or automated with a camera) and identify the device that is being used. The attacker is equipped with a portable loudspeaker device that is pre-loaded with some voice commands that they would like to issue. The command audio can be recorded by the attacker themselves, generated using Text-to-Speech software, sourced from publicly accessible repositories of human speech samples, or recorded/synthesized samples of the victim's voice. Since

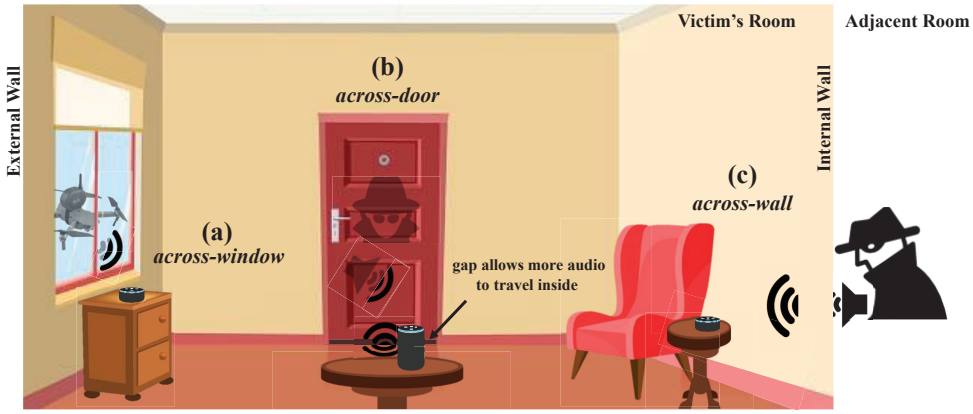


Figure 1: The BarrierBypass attack in the three barrier-based scenarios that we explore in our study including the (a) window barrier, (b) door barrier, and (c) wall barrier. The attacker is located on one side of the barrier, either in person such as an adjoining room or remotely using a drone, and attempts to inject an audible command to control the voice assistant located on the other side.

modern voice assistants are not voice specific by default, the attacker does not necessarily need command audio that is in the victim's voice, making this attack easier to conduct. In fact, the attacker can run initial testing in their own space to identify a particular voice sample that performs the best for targeting a specific voice assistant device or passing through a specific barrier. BarrierBypass is designed as an untargeted attack that can be executed independent of the victim. The attacker can use any speech audio so there is no dependence on acquiring the user's speech. Therefore, the same attack setup can also be launched against many different victims successively in a short period of time. There is also a lot of freedom for the attack to target any available barrier separating them from the victim voice assistant (i.e., they can issue the command across all available windows or walls). In particular, the attack could launch BarrierBypass remotely using a drone device equipped with a loudspeaker. The drone can fly around to inject the command audio and could target all the windows in a home and even multiple homes (i.e., an entire street or neighborhood) and "leave" the scene very quickly if they suspect detection. They could also target apartments/condos in a high-rise building by flying the drone up to a window. Drones can be purchased cheaply and can come already equipped with a speaker [38] for $150, or the speaker device can be purchased by itself for $50 [19] and attached to any drone. While the BarrierBypass attack is fully functional as an untargeted attack, there is some potential for a more targeted approach against a specific victim. Using a replay or synthesis attack, an attacker can fool speaker recognition on a virtual assistant device and achieve even more severe attack capabilities.

While BarrierBypass is intended to be launched when the user is not home, there are some scenarios where it can be launched with the victim present in the space. Because the command audio loses a lot of power and becomes quieter as it passes through a physical barrier, there is potential for the injected command to go unnoticed. In some cases the victim may be occupied doing some task or activity that may draw their attention away from their voice assistant (i.e., taking a shower, napping, watching TV in another room). During these times, the attack can still launch the attack successfully while avoiding detection.

Since the goal of the attack is to issue a command, we consider both parts of a voice assistant command audio, the wake word and the command itself. We recognize that wake word injection is foremost crucial for the attack because it activates the device to accept commands. Additionally, injecting the wake word alone can open up new attack possibilities. When a voice assistant is woken up, a recording is made that is sent over the internet for processing and is typically stored in a command history log. Therefore, an attacker could inject the wake word with the intent of allowing the device to make an unauthorized recording of the audio in the space (i.e., user speech, audio from a television, music playing). The attacker may then compromise the online repository of VA recordings to learn private user information. *While we evaluate the BarrierBypass attack on voice assistant devices, it is important to note that the attack is applicable to any voice controllable system.*

4 METHODOLOGY

4.1 Experimentation

Parameters: To generalize the results from our experimental attack simulations, we consider multiple parameters and values. Aside from the three types of barriers and different setups for each, we also test VA command audio samples from Male and Female speakers that are generated using text-to-speech or recorded from live human speakers. We consider different loudness levels for the injected audio including 65 dB to represent normal conversational loudness, 75 dB to represent loud speech, and even 85 dB for very loud audio achievable using a loudspeaker device. We tested different distances of the VA device from the barrier including 0.1 and 0.5 meters for the across-window attack, and 0.1, 0.5, 1, 2, and 4 meters for the across-door attack. Lastly, we ran experiments using two different types of VA smart speakers.

Experimental Setup: For each experiment we recreate a realistic attack setup with the portable loudspeaker placed on one side of the barrier (attacker side), and the target smart speaker on the other side of the barrier (victim side) at certain distances. We ensure that the loudspeaker and smart speaker devices are aligned directly across from each other with the loudspeaker facing the

barrier. We use the digital sound level meter on the attacker side to set the SPL of the command audio from the loudspeaker to the appropriate loudness. As a representative example of a command an attacker may attempt to issue, we selected the single-word, "Disarm" command. We consider the scenario where an attacker may be attempting to enter a victim's home and needs to disable the security system that is linked to their smart home environment (i.e., smart speakers). However, we believe our results are representative of other types of single-word commands. For each experimental parameter setting, we attempted the attack 10 times and recorded the number of successful injections of the wake word and the command portions. With 12 speaker samples, 3 SPL levels, 10 barrier/distance combinations, and 2 smart speaker devices, conducted a total of 7,200 attack simulations as part of our evaluation.

Command Audio Samples: We created a set of command audio samples consisting of both Text-to-Speech (TTS) samples and recordings of Live Human (LH) speakers saying the single-word command, "Hey Google/Alexa, Disarm". This command represents an attacker's attempt to turn off a user's home security system so that the attacker may gain access. We do not make any claims that our results are representative of other single-word commands, but we do believe that more complex commands would make the attack more difficult. Specifically, we use samples from three Male speakers (M1-M3) and three Female speakers (F1-F3), for both sample types, for a total of 12 different speaker samples. The TTS samples were generated using a free online text-to-speech generator [1], and the LH speech samples were recorded directly from volunteers. Prior to our experimentation, we confirmed that all of the command audio samples that we collected achieved 100% recognition success in the non-malicious setting (when there is no ambient noise or physical barriers present).

Equipment: In our experiments we use a cheap and low-end Sony SRS-XB2 portable loudspeaker to play the command audio. Notably, more powerful speakers can improve attack success. For the victim voice assistant, we use both the Amazon Echo Dot 2 and Google Home mini smart speakers. In order to ensure the command audio was played at the correct sound pressure level (SPL) we use a Rolls SLM305 digital sound level meter. Additionally, we built our own soundproof box and wall inserts for the across-wall scenario.

4.2 Soundproof Box

For the across-wall attack, we construct a soundproof box in order to self contain the experiments in a highly controlled space that allows us to test different wall constructions. This approach allows to select specific building materials with sound blockage ratings that we know beforehand and to ensure that the command audio is only able to reach the VA device by passing through the wall. We found this approach easier than attempting to learn what materials were used in the walls of a real environment. To build the soundproof box (pictured in Appendix Figure 4a) we followed the instructions outlined in [16]. We lined a cardboard box with foam board using 3M Super77 Spray Adhesive. Next, we added a layer of 1/4" thick Dynamat Dynaliner (Self-Adhesive Sound Deadener). Lastly we added a layer of 3" Acoustic Foam Egg Crate Panels using Auralex Foamtak adhesive spray. The different layers of the soundproof box are shown in Appendix Figure 4b.

4.3 Wall Inserts

To experiment with different across-wall barriers, we constructed two wall inserts to fit inside the soundproof box, pictured in Appendix Figure 3. These inserts are constructed to the exact measurements that allow the insert to fit inside the soundproof box with a tight seal around all edges. Appendix Figure 4c shows the setup for the across-wall attack experiments. The inserts were built with and without insulation [32]. Both inserts have a 2"x4" wood frame and are encased in 5/8" drywall panels that are cut to the exact dimensions of the frame. One of the inserts contains R13 Fiberglass insulation inside the stud frame, while the other insert was left empty. The stud frames were connected using 1 1/2" wood screws, and the drywall was attached with drywall glue and screws.

5 ATTACK RESULTS

In this section we report the BarrierBypass attack results from our experiments. We recorded and present both wake word and command injection success for all audio samples. Appendix Tables 4 & 5 show the *wake word* injection rates for the Amazon Echo Dot and Google Home mini smart speakers, respectively. And Tables 1 & 2 show the *command* injection rates. The values represent the percentage of successful injection out of 10 attempts. We present results for the standard implementation of BarrierBypass using non-specific voice audio for command injection, and discuss our investigation of the targeted implementation for fooling speaker recognition. To save space, we condensed the tables to include only the rows that showed instances of injection success. Therefore, any command SPLs or distances tested that were not included in these tables had no injection success for any of the speaker samples.

5.1 Standard BarrierBypass:

Across-Wall Attack: (Amazon Echo Dot 2) From our experiments for the across-wall attack, we observe that both wake word and command injection success was only possible when the audio was played at the loudest SPL level, 85 dB, when attacking the Amazon Echo Dot 2 device. If we compare the average injection success rates for both types of speakers for the across-wall attack with no insulation, we get 22% success for the live speaker samples, and 50% success for the TTS samples. And if we look at the across-wall attack with insulation we find that the wake word injection success completely diminishes to 0% for the live speakers, and slightly decreases to 47% for the TTS speakers. Comparing the injection success averages from the command injection results we see a similar trend. With no insulation, the live speaker samples have average injection success of 15% and the TTS speaker have 38%. And when insulation is added we again find the live speaker sample injection success drops to 0% and the TTS speaker samples slightly decreases to 35%. However, part of our threat model is that the attacker can perform preliminary testing and select the best performing command sample to launch their attack. ***Choosing sample TTS-M1, the attack achieves 100% success rates for wake word and command injection, at 85 dB for both types of walls, when targeting the Echo Dot .***

(Google Home mini) For the Google Home mini we observed wake word and command injection success at 75 dB and 85 dB, and much greater success rates overall compared to the Amazon Echo

Table 1: Command injection success rates, for attacking the Amazon Echo Dot 2, for each Barrier scenario. *Table is condensed to include only rows that showed some injection success.

Attack Scenario	Distance (m)	Cmd SPL (dB)	Live Speaker Recorded Samples						Text-to-Speech Samples					
			LS-F1	LS-F2	LS-F3	LS-M1	LS-M2	LS-M3	TTS-F1	TTS-F2	TTS-F3	TTS-M1	TTS-M2	TTS-M3
Across-Wall (Not Insulated)	0.1	85	0%	0%	0%	**10%**	**50%**	**30%**	0%	**50%**	**80%**	**100%**	0%	0%
Across-Wall (Insulated)	0.1	85	0%	0%	0%	0%	0%	0%	0%	**40%**	**70%**	**100%**	0%	0%
Across-Window	0.1	85	**10%**	0%	0%	0%	0%	0%	0%	**90%**	0%	0%	**80%**	0%
Across-Door	0.1	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**100%**	**50%**	0%
		85	**100%**	**100%**	**30%**	**100%**	**100%**	**80%**	**20%**	**100%**	**100%**	**100%**	**100%**	**100%**
	0.5	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**100%**	**20%**	0%
		85	0%	**30%**	0%	**100%**	**100%**	0%	**10%**	**100%**	**100%**	**100%**	**100%**	**100%**
	1	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**20%**	0%	0%
		85	0%	**10%**	0%	**50%**	**80%**	0%	0%	**80%**	**70%**	**100%**	**100%**	**100%**
	2	85	0%	0%	0%	**20%**	0%	0%	0%	**90%**	**50%**	**100%**	**100%**	**70%**
	4	85	0%	0%	0%	0%	0%	0%	**10%**	**10%**	0%	**70%**	**80%**	0%

Table 2: Command injection success rates, for attacking the Google Home mini, for each Barrier scenario. *Table is condensed to include only rows that showed some injection success.

Attack Scenario	Distance (m)	Cmd SPL (dB)	Live Speaker Recorded Samples						Text-to-Speech Samples					
			LS-F1	LS-F2	LS-F3	LS-M1	LS-M2	LS-M3	TTS-F1	TTS-F2	TTS-F3	TTS-M1	TTS-M2	TTS-M3
Across-Wall (Not Insulated)	0.1	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**100%**	0%	0%
		85	**80%**	**50%**	**60%**	**70%**	**40%**	**60%**	**30%**	**90%**	**90%**	**100%**	**50%**	**40%**
Across-Wall (Insulated)	0.1	75	0%	0%	0%	0%	0%	0%	0%	**10%**	0%	**100%**	0%	0%
		85	**60%**	0%	**50%**	**40%**	**20%**	0%	0%	**90%**	**80%**	**100%**	**20%**	**20%**
Across-Window	0.1	85	0%	0%	0%	0%	**30%**	0%	0%	**100%**	0%	**10%**	**60%**	0%
Across-Door	0.1	75	0%	0%	0%	0%	**100%**	**10%**	**20%**	0%	0%	**100%**	**90%**	0%
		85	**100%**	**70%**	0%	**100%**	**100%**	**20%**	**100%**	**100%**	**90%**	**100%**	**100%**	**80%**
	0.5	75	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	**40%**	0%	0%
		85	**20%**	0%	0%	**20%**	**100%**	0%	**100%**	**10%**	0%	**100%**	**100%**	0%
	1	75	0%	0%	0%	0%	**90%**	0%	0%	0%	0%	0%	0%	0%
		85	**20%**	0%	0%	**10%**	**100%**	0%	**20%**	0%	0%	**100%**	**100%**	0%
	2	75	0%	0%	0%	0%	**60%**	0%	0%	0%	0%	0%	0%	0%
		85	0%	0%	0%	0%	**100%**	0%	0%	0%	0%	**60%**	0%	0%
	4	85	0%	0%	0%	0%	**80%**	0%	**10%**	0%	0%	**40%**	0%	0%

Dot 2. Again, comparing the average wake word injection success rates for both types of speakers we find at 75 dB the average live speaker sample success is 95%, outperforming the average TTS sample success of 68%. At 85 dB, both live speaker and TTS samples achieve 100% wake word injection success. When insulation is added the average success rates slightly decrease. At 75 dB, the live speaker and TTS sample success rates decrease to 85% and 57%, respectively. And at 85 dB the success rates decrease from 100% for both speaker types with live speaker samples achieving 93% and TTS samples achieving 72%. Like the Amazon Echo Dot 2 results, we see a large decrease in command injection success compared to the wake word injection. At 75 dB, the live speaker samples had 0% command injection success, and the TTS samples had 17% command injection success. When the audio was played at 85 dB these average success rates increase to 60% and 67% for the live speaker and TTS samples, respectively. When the insulation was added, we see very similar success rates at the 75 dB level of 0% and 18% for live speakers and TTS samples, respectively. However, at 85 dB, we see a decrease in injection success (compared to no insulation) with live speaker samples dropping to 28% and TTS samples dropping to 52%. ***Choosing sample TTS-F3 or TTS-M1 achieves 100% success rates for injecting the wake word at both SPL levels, and TTS-M1 achieves 100% success for injecting the command at both SPL levels, when targeting the Google Home.***

Across-Window Attack: In the across-window attack we observed injection success at 0.1 meters. Increasing the distance to 0.5 and 1 meter completely diminished injection success for all speaker samples and audio SPL levels. For both smart speakers we observed injection success at 75 dB and 85 dB for the wake word, and at 85 dB for the command.

(Amazon Echo Dot 2) At 75 dB we observe no wake word injection success for the live speaker samples, and only two instances of injection success (3% average) for the TTS samples. When the audio was increased to 85 dB the average success rates increased to 15% for the live speaker samples and 48% for the TTS samples. No wake command injection success was observed at the 75 dB level, but at 85 dB we observed one instance of successful injection (2% average) for the live speaker samples. The TTS samples showed greater success with an average of 28% command injection success. ***Selecting sample TTS-F2 or TTS-M1 allows the attack to achieve 100% success rates for injecting the wake word at the 85 dB SPL level, and keeping TTS-F2 achieves 90% success for injecting the command at the 85 dB SPL level, targeting the Amazon Echo Dot.***

(Google Home mini) In the results for the Google Home mini we observed nearly identical wake word injection success rates for the live speaker and TTS samples at both the 75 dB and 85 dB SPL levels. At 75 dB the live speaker samples had no wake word injection success and the TTS samples had only one instance of success (2% average). When the SPL level was increased to 85 dB, both the live speaker samples and TTS samples showed an average of 78% wake word injection success. Looking at the success rates for command injection, we find that the TTS samples were more successful. The live speakers samples had an average command injection success rate of 5% while the TTS samples achieved 28%. ***Selecting any of the samples LS-F1, LS-F3, LS-M1, LS-M2, TTS-F2 or TTS-F3 achieves 100% success rates for injecting the wake word at the 85 dB SPL level, and sticking with the TTS-F2 sample achieves 100% success for injecting the command at the 85 dB SPL level, targeting the Google Home.***

Table 3: Command injection success rates for the drone experiments.

Drone	Smart Speaker	Speaker ID	Command SPL	Attack Success
Toys-Sky S167 Quadcopter (operating SPL = 85 dB)	Amazon Echo Dot	TTS-M1	90 dB	50%
		TTS-M2	90 dB	50%
		TTS-F2	90 dB	0%
		TTS-M1	95 dB	80%
		TTS-M2	95 dB	100%
		TTS-F2	95 dB	60%
Holy Stone HS700 (operating SPL = 73 dB)	Amazon Echo	TTS-M1	85 dB	90%
		TTS-M2	85 dB	100%
		TTS-F2	85 dB	100%
	Google Home	TTS-M1	85 dB	70%
		TTS-M2	85 dB	90%
		TTS-F2	85 dB	40%

(Drone Attack) Table 3 depicts the results for each drone-based scenario tested. Firstly, we found that using a drone with an operating loudness of 85 dB (S167) required command audio be played at 90+ dB. Specifically, at 90 dB we observed command injection success only up to 50%. However, increasing the command audio to 95 dB allowed us to observe attack success up to 100%. Since the operating loudness of the S167 was equal to the volume of audio used in our original experiments, the command audio in the presence of the drone had to be increase by at least 5 dB to overcome the added noise and achieve an SNR closer to 1.0 for successful command injection.

For our remaining experiments using the HS700 drone with a much lower operating loudness level, we observed high rates of command injection success, similar to what was observed in prior experiments when no drone was used. Because the operating loudness of the HS700 is only 73 dB, the 85 dB command audio level was not hindered by the added noise because it maintained a similarly high SNR. When targeting the Echo Dot, we observed attack success up to 100%. And when targeting the Google Home we observed attack success up to 90%.

Across-Door Attack: In the across-door attack we observed wake word and command injection success rates at both the 75 dB and 85 dB SPL levels for most of the distances tested. Compared to other barriers, the results confirm that the door is easiest to compromise.

(Amazon Echo Dot 2) For the live speaker samples, we observed wake word injection success at the 75 dB for the 0.1-meter distance only, achieving an average of 60% injection success. At all other distances there was no wake word injection success at 75 dB. When the audio was raised to 85 dB, we observed a greater range of success across the different distances tested. On average, the live speaker samples achieved 97%, 80%, 62%, and 12% wake word injection success for the 0.1, 0.5, 1, and 2-meter distances, respectively. In comparison, the TTS samples showed greater success for both SPL levels and all distances. At 75 dB, the TTS samples achieved average wake word injection success rates of 55%, 37%, 30%, and 17% for the 0.1, 0.5, 1, and 2-meter distances, respectively. And at 85 dB, we observe wake word injection success rates of 100% for 0.1 and 0.5 meters, and 88%, 80%, and 50% for 1, 2, and 4-meter distances.

Looking at the results for command injection we again find decreased success rates compared to the wake word. For the live speaker samples we did not observe any command injection success at the 75 dB SPL level. At 85 dB, we observe average success rates of 85%, 38%, 23% and 3% for the 0.1, 0.5, 1, and 2-meter distances, respectively. Like the wake word results, we found that the TTS samples performed better for command injection. At 75 dB we observed average command injection success rates of 25%, 20%, and 3% for the 0.1, 0.5, and 1-meter distances, respectively. And when the audio was raised to 85 dB, we observed average success rates of 87%, 85%, 75%, 68%, and 28% for the 0.1, 0.5, 1, 2, and 4-meter distances, respectively. While multiple speaker samples showed very high success rates at certain SPL levels and distances, there are two samples that outperformed the rest. ***By choosing TTS-M1 the attack can achieve 100% success at both SPL levels up to 2 meter distances, and achieves 90% success at the 4-meter distance. And choosing TTS-M1 or TTS-M2 for command injection allows the attack to achieve up to 100% success rates at the 75 dB SPL level of distances up to 0.5 meters. When the SPL level is increased to 85 dB, the attack can achieve 100% success for distances up to 2 meters, and 80% success at 4 meters when launching the attack against the Amazon Echo Dot.***

(Google Home mini) We observed greater wake word injection success with the Google Home mini. For both 75 dB and 85 dB levels we see instances of wake word injection success at all distances that were tested. At 75 dB, the live speaker samples achieved average injection success rates of 85%, 52%, 53%, 13%, and 8% for the 0.1, 0.5, 1, 2, and 4-meter distances, respectively. The TTS samples showed similar success rates of 67%, 53%, 30%, 33%, and 3% for the 0.1, 0.5, 1, 2, and 4-meter distances, respectively. When the audio was increased to 85 dB, the average success rates increased. The live speaker samples achieved 100% success for the 0.1, 0.5, and 1-meter distances, and achieved 78% and 53% for the 2 and 4-meter distances, respectively. The TTS samples achieved 100% success for the 0.1 and 0.5-meter distances, and 98%, 83%, and 42% success at 1, 2, and 4-meter distances.

For command injection, we observed a decrease in the average success rates for both speaker types. However, instances of success were still observed for both the 75 dB and 85 dB SPL levels for distances up to 2 meters. At 75 dB, the average command injection success rates for the live speaker samples were 18%, 17%, 15%, and 10% for the 0.1, 0.5, 1, and 2-meter distances. The TTS samples had less success at the larger distances and only achieved accuracies of 35% and 7% for the 0.1 and 0.5-meter distances. When the audio was played at 85 dB, both speaker types showed command injection success at all distances. The live speaker samples achieved average success rates of 65%, 23%, 22%, 17%, and 13% for the 0.1, 0.5, 1, 2, and 4-meter distances, respectively. The TTS samples outperformed the live speaker samples at the shorter distances achieving success rates of 95%, 52%, 37%, 10%, and 8% for the 0.1, 0.5, 1, 2, and 4-meter distances, respectively. ***Choosing LS-M1, LS-M2, TTS-F1, or TTS-F3 will allow the attack to achieve 100% success for wake word injection at both SPL levels up to 2-meter distances. At the 4-meter distance the attack can achieve 50% and 100% accuracy at the 75 and 85 dB levels, respectively. And isolating LS-M2 for command injection allows the attack to achieve up to 100% success for both SPL levels up to 0.5 meters. At the 75 dB SPL level the attack can achieve 90% and 60% success for the 1 and 2-meter distances, respectively. And when the SPL level is raised to 85 dB the attack can achieve 100% success at distances up to 2 meters and 80% success at 4 meters when attacking the Google Home.***

5.2 Targeted BarrierBypass

Replay Attack: To investigate the potential for BarrierBypass to launch a replay attack across physical barriers, we performed a set of experiments using three speaker samples in the across-door attack setup. Specifically, we trained a voice profile for the LS-M1 speaker on both the Amazon Echo Dot 2 and Google Home mini. We recorded samples of the command "Alexa/Hey Google, what's my name?" from the live speakers LS-M1 and LS-M2, as well as a generated samples of the command using the text-to-speech speaker TTS-M1. We selected these speakers because they all achieved 100% wake word and command recognition in the across-door attack. Playing each command audio at 85 dB we recorded the number of times out of 10 attempts that the voice assistant identified the trained speaker's voice. We found that the Amazon Echo Dot 2 was 100% accurate at identifying the trained speaker and denying the other speakers. For the Google Home mini we observed the device was 80% accurate at identifying the trained speaker, and 100% accurate at not identifying the untrained speakers. These experiments demonstrated that 1) speaker recognition can identify a valid user without a barrier present, 2) it will still accept a command from a random speaker (e.g., from attacker) across a barrier, and 3) it can identify a replayed voice of the valid user across a barrier.

Synthesis Attack: Synthesis attacks generate fake speech using a model trained on an original voice such that the synthesized voice matches the original. We performed another side investigation to observe the potential for successful synthesis attacks through physical barriers. We used the voice synthesis model SV2TTS [22] from [23] to generate the "Alexa/Hey Google, what's my name?" command in a live speakers voice. That same live speaker trained voice profiles on Amazon Echo Dot 1 and Google Home smart speakers. In the across-door attack setup, we played the synthesized command at 85 dB and recorded the number of times out of 10 attempts that the voice assistants identified the synthesized audio as coming from the legitimate user. We found that the synthesized commands were 100% successful at fooling the speaker recognition function on both of the smart speakers. This further broadens the threat level and devastating potential of the BarrierBypass attack because it demonstrates that fake commands synthesized in a user's voice are sufficient enough to fool speaker recognition, even across physical barriers.

6 SIGNAL ANALYSIS

In order to improve our understanding of why certain speech samples were successfully injected across the barriers we investigated what frequencies were most affected by the barriers and whether we could identify certain frequency characteristics in our command audio samples that may explain the different levels of success.

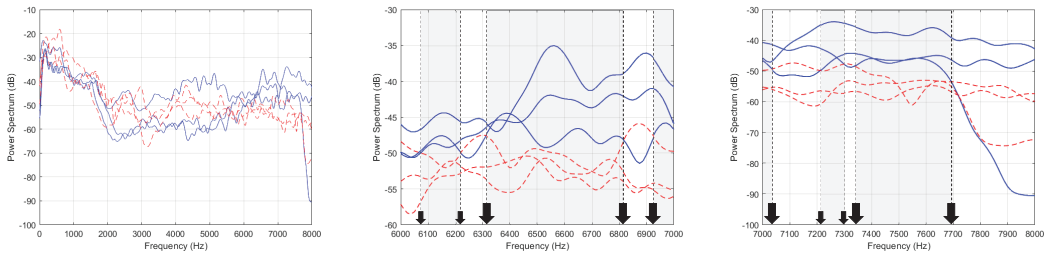
Power Spectrum We generated power spectrum graphs that overlay the spectrums for each of the command audio samples, isolating the wake word specifically, in order to compare frequency distributions and identify specific characteristics. We chose to investigate the wake word portion of the commands because 1) command injection cannot occur unless the wake word is successfully issued, and 2) the injection attack experiment results showed greater success/failure distinction, for the wake word, between different speakers. In Figure 2 we show the power spectrums of the wake word portion of the command audio samples for each individual TTS speaker. In the graphs, the solid blue lines indicate power spectrums of speaker samples that were successful at injection, while the red dashed lines indicate power spectrums of speaker samples that were not successful. From these graphs, we identify certain frequency characteristics that are consistent among the successful samples. Figure 2a shows the full power spectrum of frequencies from 0 to 8 kHz. Looking at this graph we find there are certain frequencies in the upper range that have consistencies between the successful and failing samples. Figures 2b & 2c show power spectrums that zoom into the frequency ranges of 6 to 7 kHz and 7 to 8 kHz, respectively. In these graphs we can identify five different frequency ranges (6.08-6.22 kHz, 6.32-6.82 kHz, 6.93-7.04 kHz, 7.21-7.30 kHz, and 7.34-7.69 kHz) where we find that all samples that showed successful injection have stronger frequencies in these ranges than the samples that were not successful.

While more sophisticated exploration is needed to make final conclusions, we have a few hypotheses about why certain audio samples performed better than others. First, it is possible that audio samples that showed greater success utilized more bass in the part of the wake word that are required for recognition. Therefore, as the audio passes through the physical barriers and those components of the audio are strengthened, the audio maintains a higher potential for successful recognition. Second, voice detection is trained to differentiate human speech from environmental noise and the highest frequency range captured (6-8 kHz) may be unique to human speech played through a loudspeaker, and less likely to occur naturally in an environment. Lastly, there are certain consonants that are important for speech intelligibility that appear in the upper frequency range (2-4 kHz) when recorded by a microphone [28]. The difference in frequency power within this range could also attribute to why some audio samples remain more intelligible (i.e., the samples with greater variance of power within that range).

7 SUMMARY AND DISCUSSION

Amazon vs. Google Observations: We observed some interesting trends between the two smart speaker devices that were used. Since Amazon and Google have their own speech processing services, it is reasonable to assume that different types of speaker samples will show different levels of success. If we consider the average wake word injection rates for both speaker types against the Amazon Echo Dot 2, we find that the TTS samples outperformed the live speaker recorded samples in all but one of the scenarios (Across-Door, 0.1 meters, 75 dB). Similarly, we find that the TTS samples outperformed the live speaker samples for command injection in all scenarios. This indicates that TTS samples are more effective for launching the BarrierBypass attack against Amazon devices.

Another interesting observation that became apparent when comparing the injection success rates was that wake word injection was consistently more successful when attacking the Google Home mini device. By averaging the success rate of all speaker samples for each scenario, we find that there was more success at injecting the wake word to the Google device than the Amazon device across all scenarios that were tested. We also observed through our experiments that the Google device had significantly more instances of mis-recognized commands compared to the Amazon



(a) 0-8 kHz (b) 6-7 kHz (c) 7-8 kHz

Figure 2: Power spectrum graphs of the wake word from each command audio sample that showed injection success (blue) and failure (red) in the across-wall scenario (without insulation).

device. At lower SPL levels or larger distances where the Amazon device would simply disregard the audio that it heard, the Google device would make some attempt at recognition and provide some type of response, although often incorrect.

Lastly, our work demonstrates the feasibility of BarrierBypass when launched in scenarios without environmental noise. Naturally, this would be the most ideal setting to launch the attack and ensure no other audio in the environment interferes with the injection of the command. However, we believe that some environmental noise may be manageable and still allow for a successful attack. With the inbuilt noise cancellation capabilities of modern day VA devices, any environmental noise that is quieter than the injected command audio (after it passes through the barrier) will likely be filtered out by the device and the command will still be recognized.

Sound Rating Values: We compare our observed results to the known Sound Transmission Class (STC) and Noise Reduction Coefficient (NRC) values for each of the barriers that we tested. We chose these rating values because they are both based on ASTM standards. If we consider the STC values for each of the barriers, we can see that our results are inline with the known values (33 for across-window, 30/34 for across-wall, and 20 for across-door). Now if we look at the NRC values for the different barrier materials, we find that both glass and gypsum board (i.e., drywall) have NRC values of 0.05 and wood has an NRC value of 0.10-0.15. All of these values are very low on the [0,1] scale indicating that none of the materials reflect much of the command audio back.

Drone-based Attack: Our drone experiments clearly demonstrate the feasibility of launching the BarrierBypass with drones. Specifically, an attacker could utilize a drone with a low operating loudness that does not impact the required SPL of the command audio to be injected. And by selecting the best performing command audio samples they may achieve up to 100% command injection success. This method of launching the attack provides an attacker the benefits of remote command injection and the ability to target multiple (potential) victim devices in the same area without having to physically relocate or move their attack setup. Additionally, an attacker could utilize the Wi-Peep [21] exploit to initially locate the location of the target device before launching the attack.

Improving Attack Robustness: From our experimental and analysis results we have deduced a few ways to increase the robustness of the BarrierBypass attack. First, while our results at 85 dB demonstrate the feasibility of the attack, using even louder command audio will increase the chances of attack success. An attacker can launch the BarrierBypass attack while the person is away from the home or they are in a situation where the loud audio will not cause detection. Higher volumes outside should not cause a problem, especially in scenarios with high rise buildings. An attacker could also plant a small wireless speaker onto a door or window that they could use to inject a command remotely. These devices can be very small and cheap [18], allowing the attacker to remain discrete.

Learning the type of voice assistant device that the user has before launching the attack would also help improve the chances of success. Our results demonstrate that different speaker types can have different levels of effectiveness for different devices. As a general observation, using TTS speaker samples would likely be the most effective for the BarrierBypass attack. Our analysis revealed that samples with stronger frequencies in the upper range are the most successful, so specifically choosing TTS samples that contain these qualities will improve attack success.

Limitations: The results that we observed for the BarrierBypass attack are somewhat limited to the particular settings that we controlled in our experiments. Firstly, all of our experiments were conducted in quiet spaces where the only audio present was played from the loudspeaker device for the purposes of the attack. In a real-world scenario it is likely that there are other sources of noise in the environment which would affect the overall success of this attack. Second, since our attack uses plain, audible commands for the injection, the BarrierBypass attack is dependent on the user being away from the device and in another area. Otherwise they would easily recognize the command injection attempt. Lastly, as our results demonstrated, there is a distance requirement between the victim's device and the barrier (in the across-wall and across-window scenarios) for the attack to be successful at the SPL levels we tested. While louder command audio would surely increase the attack range, it also increases the chance of discovery. Therefore, the BarrierBypass attack is limited to scenarios where the victim's device is in close proximity to the barrier being targeted.

Potential Defenses: The potential defenses against the BarrierBypass attack are largely based on hindering the physical phenomenon that would allow command audio to bypass physical barriers. One potential defense against our attack would be to use materials with higher STC and NRC values. To defend against the attack presented in this work, an STC of 50 or higher would be required. This can be achieved using concrete masonry walls, doubling layers of drywall,

or using specialized materials such as sound deadening paint or noise blocking curtains. Another potential defense is placing the smart speaker device at the furthest location from any accessible barriers. We demonstrate that distances of 4 meters become difficult for the attack even for an interior door. Another solution is to build a machine learning classifier that can differentiate between audio played through a barrier and audio played normally. As our analysis demonstrated, there are certain frequencies that are affected/blocked by the different barrier types. Blue et al. [11] achieve this effect by identifying sub-bass over-excitation which is a characteristic of audio played from loudspeaker devices and is not present in human speech. This would also be effective against BarrierBypass because as the command audio passes through the physical barrier, the bass/sub-bass components of the audio will become stronger. Another solution presented by Blue et al. [10] could also be effective at identifying the BarrierBypass attack. In their 2MA work, the authors present a two microphone authentication solution that provides source localization by determining the direction of arrival. This approach, combined with a predetermined knowledge of the VA devices placement, could be used to identify when a command is coming from the other side of a barrier.

8 RELATED WORK

Replay Attacks: Among all the spoofing attacks, replay attacks may be the most accessible to adversaries because it simply involves recording and replaying a victim's voice commands. Existing studies have shown that such attacks are effective against state-of-the-art speaker verification systems [15, 17], under scenarios of replaying over the internet or within the physical space of the victim. Other than directly replaying the recorded speech, recent studies also reveal the potential ways of enhancing the stealthiness and effectiveness of the attack. VMask [40] designed adversarial machine learning techniques to generate subtle perturbations to make any recorded speech pass speaker verification systems. To improve the stealthiness, Guo *et al.* [20] exploited a loudspeaker array to make the sound emission focus on the microphone of the VA system. To bypass existing defense schemes, Yoon *et al.* [37] leveraged a mouth simulator instead of a loudspeaker to replay the recorded speech. However, replay attacks using commands in the victim's voice are not always necessary. Many of the current VA devices available (such as those used in our study) do not employ strict speaker verification. If the audio is understandable via speech recognition, the device will execute any command that is given.

Laser-based Injection: Laser-based injection has also been utilized for signal and command injection targeting smart speakers. Recently, Light Commands [31] has brought up a new security issue, which is a new class of signal injection attacks targeting microphones of the smart speakers by physically converting a light signal to sound signal. The attacker can inject arbitrary audio signals to a target microphone by aiming a specially designed amplitude-modulated light at the microphone's aperture. By means of Light Commands [31], the attacker can obtain control over some commodity smart speakers, such as Amazon's Alexa, Apple's Siri, and Google Assistant, at distances up to 110m, which provide a brand new perspective for attacking smart speakers. One drawback to this form of attack is that it requires a direct line-of sight between the attacker and victim's device. Therefore, simply closing the blinds or moving your device to a location out of view will thwart this attack. Our BarrierBypass attack does not require this line of sight and is much more accurate in practical settings.

Ultrasonic/Hidden Audio Injection: In addition to conventional attacks through replaying human-sounding speech, researchers also show the potential of generating unintelligible or even inaudible attack sounds. Particularly, DolphinAttack [36] modulates the recorded voice commands onto the ultrasonic frequency range, which can be demodulated by the microphone due to their non-linearity. Hidden voice attacks [7, 12] convert recorded speech into obfuscated voice commands, which are recognizable to the speech recognition models while remaining unintelligible to humans. Recent studies also demonstrate the possibilities of embedding such commands into background music [39] or the audio channel of video streams [41]. By combining hidden voice commands with live speech, the hybrid commands can even bypass the state-of-the-art defense schemes [33]. While hidden voice commands introduce new approaches to evade detection, they are often very complicated to produce and are not feasible for real-world attack settings. Our attack does not obfuscate the command itself, but rather injects the clear-text command through a barrier. Hidden voice commands are obfuscated and are often misrecognized or not effective. In a recent work by Abdullah et al. [8], the authors survey current research works that present hidden voice command type attacks and demonstrate that most of them will not be successful when launched against real-world systems. In this work we evaluate BarrierBypass against live implementations of VA devices and bypass real barriers with greater ease and feasibility than hidden voice commands.

9 CONCLUSIONS AND FUTURE WORK

In this work, we present the BarrierBypass attack that issues audible voice commands to smart speakers across physical barriers. Our attack demonstrates the settings in which clean command injection can be successful and what barrier types are at risk. This attack can be launched in person or remotely via drones or other controlled devices, and allow an attacker to gain full control over a victim's VA device when the device is placed near a barrier and the scenario allows for loud command audio to be played. Compared to other command injection attacks, BarrierBypass exploits the lack of speaker verification present on modern smart speaker devices and bypasses physical barriers that would hinder other types of attacks. We evaluated the attack in multiple settings that test different command audio SPL levels and distances. Our experiments tested three different barrier-based attack scenarios using two live implementations of smart speaker devices and demonstrate that 100% wake word and command injection accuracy can be achieved when selecting the highest performing speaker samples and under certain conditions.

ACKNOWLEDGMENTS

We would like to thank the reviewers and shepherd for guiding the final revisions of this paper. This work was partially funded by the National Science Foundation (NSF) under grants CNS-1714807, CNS-2030501, CNS-2139358, CNS2114220, CNS2120396.

REFERENCES

[1] [n. d.]. *Free Text-To-Speech and Text-to-MP3 for US English.* https://ttsmp3.com/.

[2] Soundproof Windows Inc. [n. d.]. *STC Ratings.* Soundproof Windows Inc. https://www.soundproofwindows.com/stc-ratings/.

[3] Soundproofing Company Inc. [n. d.]. *Understanding STC and STC Ratings.* Soundproofing Company Inc. https://www.soundproofingcompany.com/soundproofing_101/understanding-stc-and-stc-ratings.

[4] 2016. *ASTM E413-16: Classification for Rating Sound Insulation.* https://www.astm.org/e0413-16.html.

[5] 2017. *ASTM C423-17: Standard Test Method for Sound Absorption and Sound Absorption Coefficients by the Reverberation Room Method.* https://www.astm.org/c0423-17.html.

[6] Archtoolbox 2021. *Architectural Acoustics - Controlling Sound.* Archtoolbox. https://www.archtoolbox.com/materials-systems/architectural-concepts/acoustics.html.

[7] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph N. Wilson. 2019. Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems. *ArXiv* abs/1904.05734 (2019).

[8] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. 2021. SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. In *IEEE Symposium on Security and Privacy*. 730–747.

[9] Mae Anderson. 2017. *Burger King's Ad Exposed Voice Assistants' Hackability.* https://www.inc.com/associated-press/burger-king-ad-voice-assistants-siri-alexa-google.html.

[10] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2MA: Verifying Voice Commands via Two Microphone Authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (Incheon, Republic of Korea) *(ASIACCS '18)*. Association for Computing Machinery, New York, NY, USA, 89–100. https://doi.org/10.1145/3196494.3196545

[11] Logan Blue, Luis Vargas, and Patrick Traynor. 2018. Hello, Is It Me You're Looking For? Differentiating Between Human and Electronic Speakers for Voice Interface Security. In *Proceedings of the 11th ACM Conference on Security and Privacy in Wireless and Mobile Networks* (Stockholm, Sweden) *(WiSec '18)*. Association for Computing Machinery, New York, NY, USA, 123–133. https://doi.org/10.1145/3212480.3212505

[12] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Michael E. Sherr, Clay Shields, David A. Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *USENIX Security Symposium*.

[13] Ziv Chang. 2019. *Inside the Smart Home: IoT Device Threats and Attack Scenarios.* https://www.trendmicro.com/vinfo/us/security/news/internet-of-things/inside-the-smart-home-iot-device-threats-and-attack-scenarios.

[14] Jon Chase. 2022. *The Best Smart Locks.* https://www.nytimes.com/wirecutter/reviews/the-best-smart-lock/.

[15] Héctor Delgado, Nicholas W. D. Evans, Tomi H. Kinnunen, Kong-Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md. Sahidullah, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. 2021. ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. *ArXiv* abs/2109.00535 (2021).

[16] Domininc. 2021. *How To Soundproof A Cardboard Box At Home In 8 Simple Steps.* https://soundproofcentral.com/soundproof-cardboard-box/.

[17] Serife Seda Kucur Ergunay, Elie el Khoury, Alexandros Lazaridis, and Sébastien Marcel. 2015. On the vulnerability of speaker verification to realistic voice spoofing. *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2015), 1–6.

[18] EverythingPromo. 2021. *MINI PORTABLE WIRELESS SPEAKER.* https://www.everythingpromo.com/mini-portable-wireless-speaker.

[19] Fully Needed. 2021. *Wireless Drone Megaphone Aerial Broadcasting Speaker.* https://www.fullyneeded.com/products/wireless-drone-megaphone-aerial-broadcasting-speaker.

[20] Juan Guo, Liang Chen, Haoran Sun, Aidong Xu, Zeguang Li, Yinwei Zhao, Yixin Jiang, Tengyue Zhang, and Yunan Zhang. 2021. A Defense Method Based on a Novel Replay Attack. *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)* (2021), 277–281.

[21] Janet Hurley. 2022. *Waterloo researchers make device that can 'see' through walls. What does it mean for your privacy?* https://www.thestar.com/news/gta/2022/11/05/waterloo-researchers-make-device-that-can-see-through-walls-what-does-it-mean-for-your-privacy.html.

[22] Corentin Jemine. 2022. *Real-Time Voice Cloning.* https://github.com/CorentinJ/Real-Time-Voice-Cloning.

[23] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 4485–4495.

[24] Grayson Kemper. 2020. *What Are the Main Security Risks With Smart Home Automation?* https://www.byteant.com/blog/what-are-the-main-security-risks-with-smart-home-automation/.

[25] Ainsley Lawrence. 2021. *Threats to Smart Home Security and How to Counter Them.* https://techspective.net/2021/12/27/threats-to-smart-home-security-and-how-to-counter-them/.

[26] Emily Long. 2019. *How to Keep Commercials from Activating Your Smart Speaker.* https://lifehacker.com/how-to-keep-commercials-from-activating-your-smart-spea-1831437974.

[27] Ryan Mchugh. [n. d.]. *A Guide to Interior Door Sound Ratings.* Home Advancement. https://www.homeadvancement.com/doors/interior-doors/sound-ratings.

[28] Mic University. 2021. *Facts About Speech Intelligibility.* https://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility.

[29] Rambus. [n. d.]. *Smart Home: Threats and Countermeasures.* https://www.rambus.com/iot/smart-home/.

[30] Manasa Reddigari. 2019. *The 10 Biggest Security Risks in Today's Smart Home.* https://www.bobvila.com/slideshow/the-10-biggest-security-risks-in-today-s-smart-home-53081.

[31] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems. In *USENIX Security Symposium*.

[32] Don Vandervort. 2021. *Soundproofing Walls and Ceilings.* https://www.hometips.com/diy-how-to/soundproofing-insulation-walls-ceilings.html.

[33] Yi Wu, Xiangyu Xu, Pa Walker, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. HVAC: Evading Classifier-based Defenses in Hidden Voice Attacks. *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security* (2021).

[34] Candid Wueest. 2017. *A guide to the security of voice-activated smart speakers.* https://docs.broadcom.com/doc/istr-security-voice-activated-smart-speakers-en.

[35] Candid Wueest. 2017. *Everything You Need to Know About the Security of Voice-Activated Smart Speakers.* https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/security-voice-activated-smart-speakers.

[36] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2021. The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants. *IEEE Transactions on Dependable and Secure Computing* 18 (2021), 1108–1124.

[37] Sung-Hyun Yoon, Min-Sung Koh, Jae han Park, and Ha jin Yu. 2020. A New Replay Attack Against Automatic Speaker Verification Systems. *IEEE Access* 8 (2020), 36080–36088.

[38] Young Ninos. 2021. *2021 Drone with Camera and Speaker.* https://youngninos.com/products/2019-drone-with-camera-and-speaker.

[39] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. 2018. CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition. In *USENIX Security Symposium*.

[40] Lei Zhang, Yan Meng, Jiahao Yu, Chong Xiang, Brandon Falk, and Haojin Zhu. 2020. Voiceprint Mimicry Attack Towards Speaker Verification System in Smart Home. *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications* (2020), 377–386.

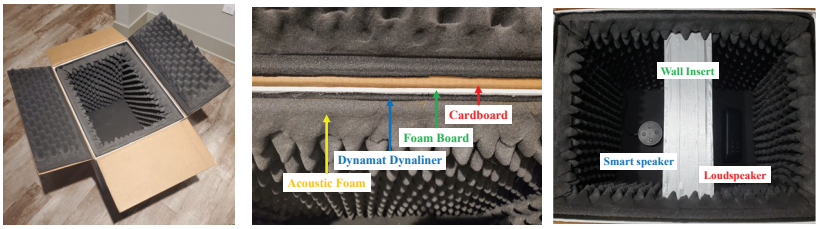
[41] Man Zhou, Zhan Qin, Xiu Lin, Shengshan Hu, Qian Wang, and Kui Ren. 2019. Hidden Voice Commands: Attacks and Defenses on the VCS of Autonomous Driving Cars. *IEEE Wireless Communications* 26 (2019), 128–133.

A APPENDIX

A.1 Additional Images

(a) No Insulation (b) Insulated

Figure 3: Inserts constructed for the Wall-Barrier.



(a) Soundproof Box (b) Box Layers (c) Experiment Aerial View

Figure 4: Images of soundproof box construction and experimental setup.

A.2 Additional Tables

Table 4: Wake Word injection success rates, for attacking the Amazon Echo Dot 2, for each Barrier scenario. *Table is condensed to include only rows that showed some injection success.

Attack Scenario	Distance (m)	Cmd SPL (dB)	Live Speaker Recorded Samples						Text-to-Speech Samples					
			LS-F1	LS-F2	LS-F3	LS-M1	LS-M2	LS-M3	TTS-F1	TTS-F2	TTS-F3	TTS-M1	TTS-M2	TTS-M3
Across-Wall (Not Insulated)	0.1	85	0%	0%	0%	**30%**	**70%**	**30%**	0%	**100%**	**100%**	**100%**	0%	0%
Across-Wall (Insulated)	0.1	85	0%	0%	0%	0%	0%	0%	0%	**90%**	**90%**	**100%**	0%	0%
Across-Window	0.1	75	0%	0%	0%	0%	0%	0%	0%	**10%**	0%	**10%**	0%	0%
		85	**30%**	0%	0%	0%	**60%**	0%	**10%**	**100%**	0%	**100%**	**80%**	0%
Across-Door	0.1	75	**50%**	**100%**	**90%**	0%	**100%**	**20%**	**90%**	**20%**	**10%**	**100%**	**90%**	**20%**
		85	**100%**	**100%**	**100%**	**100%**	**100%**	**80%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	0.5	75	0%	0%	0%	0%	0%	0%	**50%**	0%	0%	**100%**	**70%**	0%
		85	**60%**	**100%**	**80%**	**100%**	**100%**	**40%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	1	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**100%**	**80%**	0%
		85	**20%**	**100%**	**50%**	**100%**	**100%**	0%	**80%**	**80%**	**70%**	**100%**	**100%**	**100%**
	2	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	**100%**	0%	0%
		85	0%	**10%**	**10%**	**30%**	**20%**	0%	**50%**	**90%**	**50%**	**100%**	**100%**	**90%**
	4	85	0%	0%	0%	0%	0%	0%	**30%**	**10%**	0%	**90%**	**80%**	**90%**

Table 5: Wake Word injection success rates, for attacking the Google Home mini, for each Barrier scenario. *Table is condensed to include only rows that showed some injection success.

Attack Scenario	Distance (m)	Cmd SPL (dB)	Live Speaker Recorded Samples						Text-to-Speech Samples					
			LS-F1	LS-F2	LS-F3	LS-M1	LS-M2	LS-M3	TTS-F1	TTS-F2	TTS-F3	TTS-M1	TTS-M2	TTS-M3
Across-Wall (Not Insulated)	0.1	75	**100%**	**70%**	**100%**	**100%**	**100%**	**100%**	**100%**	**20%**	**100%**	**100%**	**80%**	**10%**
		85	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
Across-Wall (Insulated)	0.1	75	**100%**	**60%**	**80%**	**100%**	**100%**	**70%**	**80%**	0%	**100%**	**100%**	**60%**	0%
		85	**100%**	**80%**	**100%**	**100%**	**90%**	**90%**	**90%**	**40%**	**100%**	**100%**	**70%**	**30%**
Across-Window	0.1	75	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	**10%**	0%
		85	**100%**	**70%**	**100%**	**100%**	**100%**	0%	**80%**	**100%**	**100%**	**70%**	**70%**	**50%**
Across-Door	0.1	75	**100%**	**70%**	**40%**	**100%**	**100%**	**100%**	**100%**	0%	**100%**	**100%**	**100%**	0%
		85	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	0.5	75	**100%**	0%	**10%**	**100%**	**100%**	0%	**100%**	**20%**	**100%**	**100%**	0%	0%
		85	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	1	75	**100%**	0%	**20%**	**100%**	**100%**	0%	**70%**	0%	**100%**	**10%**	0%	0%
		85	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**90%**	**100%**	**100%**	**100%**	**100%**
	2	75	0%	0%	0%	**20%**	**60%**	0%	**100%**	0%	**100%**	0%	0%	0%
		85	**100%**	**100%**	**70%**	**100%**	**100%**	0%	**100%**	**30%**	**100%**	**80%**	**90%**	**100%**
	4	75	0%	0%	0%	0%	**50%**	0%	0%	0%	**20%**	0%	0%	0%
		85	0%	**50%**	**70%**	**100%**	**100%**	0%	**100%**	**30%**	**80%**	**40%**	0%	0%