Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems*

Takeshi Sugawara
The University of Electro-Communications
sugawara@uec.ac.jp

Benjamin Cyr
University of Michigan
bencyr@umich.edu

Sara Rampazzi
University of Michigan
srampazz@umich.edu

Daniel Genkin
University of Michigan
genkin@umich.edu

Kevin Fu
University of Michigan
kevinfu@umich.edu

***Abstract*—We propose a new class of signal injection attacks on microphones based on the photoacoustic effect: converting light to sound using a microphone. We show how an attacker can inject arbitrary audio signals to the target microphone by aiming an amplitude-modulated light at the microphone's aperture. We then proceed to show how this effect leads to a remote voice-command injection attack on voice-controllable systems. Examining various products that use Amazon's Alexa, Apple's Siri, Facebook's Portal, and Google Assistant, we show how to use light to obtain full control over these devices at distances up to 110 meters and from two separate buildings. Next, we show that user authentication on these devices is often lacking or non-existent, allowing the attacker to use light-injected voice commands to unlock the target's smartlock-protected front doors, open garage doors, shop on e-commerce websites at the target's expense, or even locate, unlock and start various vehicles (e.g., Tesla and Ford) that are connected to the target's Google account. Finally, we conclude with possible software and hardware defenses against our attacks.**

***Index Terms*—Signal Injection Attack, Transduction Attack, Voice-Controllable System, Photoacoustic Effect, Laser, MEMS**

I. Introduction

The consistent growth in computational power is profoundly changing the way that humans and computers interact. Moving away from traditional interfaces like keyboards and mice, in recent years computers have become sufficiently powerful to understand and process human speech. Recognizing the potential of quick and natural human-computer interaction, technology giants like Apple, Google, Facebook, and Amazon have each launched their own large-scale deployment of voice-controllable (VC) systems that continuously listen to and act on human voice commands.

With tens of millions of devices sold with Alexa, Siri, Portal, and Google Assistant, users can now interact with service providers without the need to sit in front of a computer or type on a mobile phone. Responding to this trend, the Internet of Things (IoT) market has also undergone a small revolution. Rather than having each device be controlled via a dedicated piece of software, IoT manufacturers now spend time making hardware, coupled with a lightweight interface to integrate their products with Alexa, Siri or Google Assistant. Thus, users can receive information and control products by the mere act of speaking, without the need for physical interaction with keyboards, mice, touchscreens, or even buttons.

However, while much attention is being given to improving the capabilities of VC systems, much less is known about the resilience of these systems to software and hardware attacks. Indeed, previous works [1, 2] already highlight a major limitation of voice-only user interaction: the lack of proper user authentication. As such, a voice-controllable system can execute an injected command without the need for additional user confirmation. While early command-injection techniques were easily noticeable by the device's legitimate owner, a more recent line of work [3, 4, 5, 6, 7, 8, 9] focuses on stealthy command injection, preventing the user from recognizing or even hearing the injected commands.

The absence of voice authentication has resulted in a proximity-based threat model, where close-proximity users are considered legitimate, while attackers are kept at bay by physical obstructions like walls, locked doors, and closed windows. For attackers aiming to surreptitiously gain control over physically-inaccessible systems, existing injection techniques are unfortunately limited, as the current state of the art [6] has an injection range limited to 25 *ft* (7.62 m) in open space, with physical barriers (e.g., windows) further reducing the distance.

Thus, in this paper we tackle the following questions:
Can commands be remotely and stealthily injected into a voice-controllable system? If so, how can an attacker perform such an attack under realistic conditions and with limited physical access? Finally, what are the implications of such command injections on third-party IoT hardware integrated with the voice-controllable system?

A. Our Contribution

In this paper we present LightCommands, an attack that is capable of covertly injecting commands into voice-controllable systems at long distances.

Laser-Based Audio Injection. We have identified a semantic gap between the physics and specifications of MEMS (micro-electro-mechanical systems) microphones, where such micro-

*Paper version as of November 4, 2019. More information and demonstrations is available at lightcommands.com

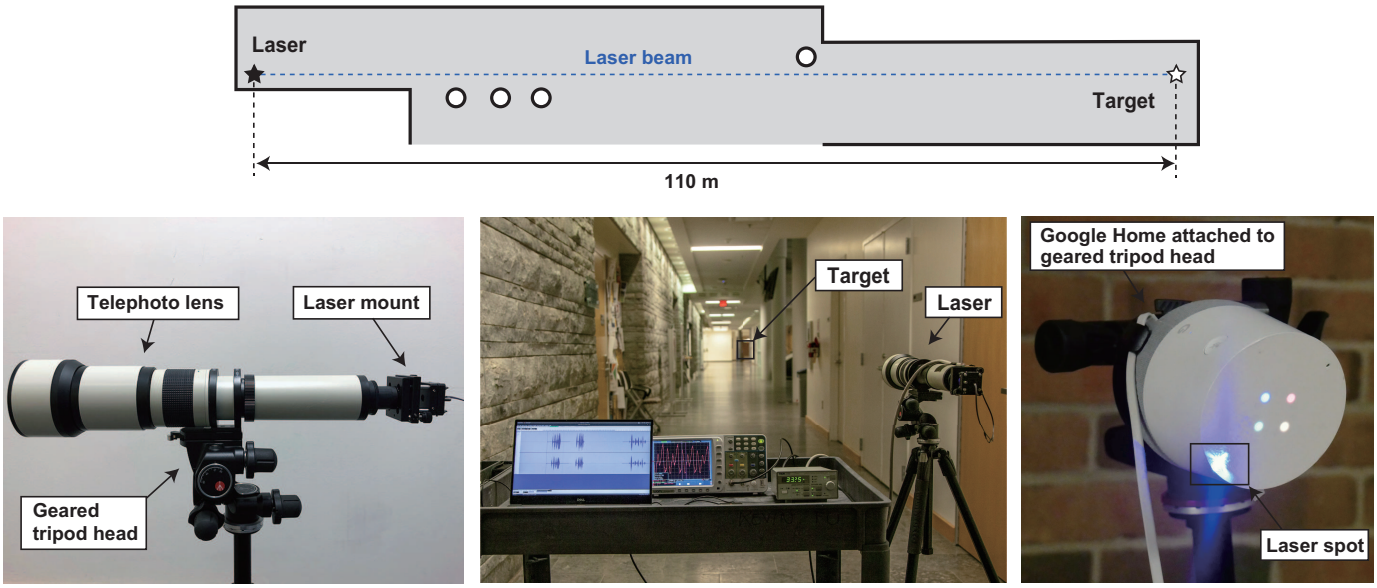

Fig. 1. Experimental setup for exploring attack range. (Top) Floor plan of the 110 m long corridor. (Left) Laser with telephoto lens mounted on geared tripod head for aiming. (Center) Laser aiming at the target across the 110 m corridor. (Right) Laser spot on the target device mounted on tripod.

phones unintentionally respond to light as if it was sound. Exploiting this effect, we can inject sound into microphones by simply modulating the amplitude of a laser light.

Attacking Voice-Controllable Systems. Next, we investigate the vulnerability of popular VC systems (such as Alexa, Siri, Portal, and Google Assistant) to light-based audio injection attacks. We find that 5 mW of laser power (the equivalent of a laser pointer) is sufficient to obtain full control over many popular Alexa and Google smart home devices, while about 60 mW is sufficient for gaining control over phones and tablets.

Long Range. Using a telephoto lens to focus the laser, we demonstrate the first long-range command injection attack on VC systems, achieving distances of up to 110 meters (the maximum available space in our testing area) as shown in Figure 1. We also demonstrate how light can be used to control VC systems across buildings and through closed glass windows at similar distances. Finally, we note that unlike previous works that have limited range due to the use of sound for signal injection, the range obtained by light-based injection is only limited by the attacker's power budget, optics, and aiming capabilities.

Insufficient Authentication. Having established the feasibility of malicious control over VC systems, we investigate the security implications of sound injection attacks. We find that VC systems are often lacking user authentication mechanisms, or if the mechanisms are present, they are incorrectly implemented (e.g., allowing for PIN brute forcing). We show how an attacker can use light-injected voice commands to unlock the target's smart-lock protected front door, open garage doors, shop on e-commerce websites at the target's expense, or even locate, unlock and start various vehicles (e.g., Tesla and Ford) if the vehicles are connected to the target's Google account.

Attack Stealthiness and Cheap Setup. We then show how an attacker can build a cheap yet effective injection setup, using commercially available laser pointers and laser drivers. Moreover, by using infrared lasers and abusing volume features (e.g., whisper mode for Alexa devices) on the target device, we show how an attacker can mount a light-based audio injection attack while minimizing the chance of discovery by the target's legitimate owner.

Countermeasures. Finally, we discuss software and hardware-based countermeasures against our attacks.

Summary of Contributions. In this paper we make the following contributions.

1) Discover a hardware problem with MEMS microphones, making them susceptible to light-based signal injection attacks (Section IV).
2) Investigate the vulnerability of popular Alexa, Siri, Portal, and Google Assistant devices to light-based command injection across large distances and varying laser power (Section V).
3) Investigate the security implications of malicious command injection attacks on VC systems and demonstrate how such attacks can be mounted using cheap and readily available equipment (Section VI).
4) Discuss software and hardware countermeasures to light-based signal injection attacks (Section VII).

B. *Safety and Responsible Disclosure*

Laser Safety. Laser radiation requires special controls for safety, as high-powered lasers might cause hazards of fire, eye damage, and skin damage. We urge that researchers receive formal laser safety training and approval of experimental designs before attempting reproduction of our work. In particular, all the experiments in this paper were conducted under a Standard Operating Procedure which was approved by our university's Safety Committee.

Disclosure Process. Following the practice of responsible disclosure, we have shared our findings with Google, Amazon, Apple, August, Ford, Tesla, and Analog Devices, a major supplier of MEMS microphones. We subsequently maintained contact with the security teams of these vendors, as well as with ICS-CERT and the FDA. The findings presented in this paper were made public on the mutually-agreed date of November 4th, 2019.

II. Background

A. Voice-Controllable System

The term "Voice-Controllable (VC) system" refers to a system that is controlled primarily by voice commands directly spoken by users in a natural language, e.g., English. While some important exceptions exist, VC systems often immediately operate on voice commands issued by the user, without requiring further interaction. For example, when the user commands the VC system to "open the garage door", the garage door is immediately opened.

Following the terminology of [4], a typical VC system is composed of three main components: (i) voice capture, (ii) speech recognition, and (iii) command execution. First, the voice capture subsystem is responsible for converting the sound produced by the user into electrical signals. Next, the speech recognition subsystem is responsible for detecting the wake word in the acquired signal (e.g., "Alexa" for Amazon's Alexa, "OK Google" for Google Assistant, "Hey, Portal" for Facebook's Portal and "Hey Siri" for Apple's Siri) and subsequently interpreting the meaning of the voice command using signal and natural-language processing. Finally, the command-execution subsystem launches the corresponding application or executes an operation based on the recognized voice command.

B. Attacks on Voice-Controllable Systems

Several previous works explored the security of VC systems, uncovering vulnerabilities that allow attackers to issue unauthorized voice commands to these devices [3, 4, 5, 6, 7].

Malicious Command Injection. More specifically, [1, 2] developed malicious smartphone applications that play synthetic audio commands into nearby VC systems without requiring any special operating system permissions. While these attacks transmit commands that are easily noticeable to a human listener, other works [3, 8, 9] focused on camouflaging commands in audible signals, attempting to make them unintelligible or unnoticeable to human listeners, while still being recognizable to speech recognition models.

Inaudible Voice Commands. A more recent line of work focuses on completely hiding the voice commands from human listeners. Roy et al. [5] demonstrate how high frequency sounds inaudible to humans can become recordable by commodity microphones. Subsequently, Song and Mittal [10] and *DolphinAttack* [4] extended the work of [5] by sending inaudible commands to VC systems via word modulation on ultrasound carriers. By exploiting nonlinearities in the microphones, a signal modulated onto an ultrasonic carrier is demodulated to the audible range by the targeted microphone, recovering the original voice command while remaining undetected by humans.

However, both attacks are limited to short distances (from 2 *cm* to 175 *cm*) due to the transmitter operating at low power. Unfortunately, increasing the transmitting power generates an audible frequency component containing the (hidden) voice command, as the transmitter is also affected by the same nonlinearity observed in the receiving microphone. Tackling the distance limitation, Roy et al. [6] mitigated this effect by splitting the signal in multiple frequency bins and playing them through an array of 61 speakers. However, the re-appearance of audible leakage still limits the attack's range to 25 *ft* (7.62 m) in open space, with physical barriers (e.g., windows) and the absorption of ultrasonic waves in air further reducing range by attenuating the transmitted signal.

Skill Squatting Attacks. A final line of work focuses on confusing speech recognition systems, causing them to misinterpret correctly-issued voice commands. These so-called skill squatting attacks [11, 12] work by exploiting systematic errors in the recognition of similarly sounding voice commands to route users to malicious applications without their knowledge.

C. Acoustic Signal Injection Attacks

Several works used acoustic signal injection as a method of inducing unintended behavior in various systems.

More specifically, Son et al. [13] showed that MEMS sensors are sensitive to ultrasound signals, resulting in jamming denial of service attacks against inertial measurement unit (IMU) on drones. Subsequently, Yan et al. [14] demonstrated that acoustic waves can be used to saturate and spoof ultrasonic sensors, impairing the safety of cars. This was further improved by Walnut [15], which exploited aliasing and clipping effects in the sensor's components to achieve precise control over MEMS accelerometers via sound injection.

More recently, Nashimoto et al. [16] showed the possibility of using sound to attack sensor-fusion algorithms that rely on data from multiple sensors (e.g., accelerometers, gyroscopes, and magnetometers) while Blue Note [17] demonstrates the feasibility of sound attacks on mechanical hard drives, resulting in operating system crashes.

D. Laser Injection Attacks

In addition to sound, light has also been utilized for signal injection. Indeed, [18, 19, 14] mounted denial of service attacks on cameras and LiDARs by illuminating victims' photoreceivers with strong lights. This was later extended by Shin et al. [20] and Cao et al. [21] to a more sophisticated attack that injects precisely-controlled signals to LiDAR systems, causing the target to see an illusory object. Next, Park et al. [22] showed an attack on medical infusion pumps, using light to attack optical sensors that count the number of administered medication drops. Finally, [23] show how various sensors, such as infrared and light sensors, can be used to activate and transfer malware between infected devices.

Another line of work focuses on using light for injecting faults inside computing devices, resulting in security breaches.

More specifically, it is well-known that laser light causes soft (temporary) errors in semiconductors, where similar errors are also caused by ionizing radiation [24]. Exploiting this effect, Skorobogatov and Anderson [25] showed the first light-induced fault attacks on smartcards and microcontrollers, demonstrating the possibility of flipping individual bits in memory cells. This effect was subsequently exploited in numerous follow ups, using laser-induced faults to compromise the hardware's data and logic flow, extract secret keys, and dump the device's memory. See [26, 27] for further details.

E. Photoacoustic Effect

Photoacoustics is a field of research that studies the interaction between light and acoustic pressure waves (see [28] for a survey). The first work in this area dates back to 1880, where Alexander Graham Bell [29] invented an optical communication device that uses a vibrating mirror as a mechanical sunlight modulator and a selenium cell to convert the modulated light back to electricity. While the so-called photophone was successful at transmitting voice across distances, the inherent requirement of having a line of sight between the transmitter and receiver made the technology inferior to the radio communication that was emerging at that time. The rise of digital communication technology has made the analog modulation even less attractive, and voice transmission over light had been forgotten for decades.

Recently, researchers rediscovered light-voice transmission as a sophisticated user interface that delivers an audible message to a particular user by using air as the medium. Tucker [30] reported that the U.S. military is developing a device that ionizes molecules in the air using an extremely short-pulse (femtosecond) laser to generate plasma that makes sound. Sullenberger et al. [31] proposed a different way of generating sound using an infrared laser with a particular wavelength that efficiently heats up ambient water vapor, causing an acoustic pressure wave in the air which results in successful sound delivery to a user at 2.5 meters away.

F. MEMS Microphones

MEMS is an integrated implementation of mechanical components on a chip, typically fabricated with an etching process. While there are a number of different MEMS sensors (e.g., accelerometers and gyroscopes), in this paper we focus on MEMS-based microphones, which are particularly popular in mobile and embedded applications (such as smartphones and smart speakers) due to their small footprints and low prices.

Microphone Overview. The first column of Figure 2 shows the construction of a typical backport MEMS microphone, which is composed of a diaphragm and an ASIC circuit. The diaphragm is a thin membrane that flexes in response to an acoustic wave. The diaphragm and a fixed back plate work as a parallel-plate capacitor, whose capacitance changes as a consequence of the diaphragm's mechanical deformations as it responds to alternating sound pressures. Finally, the ASIC die converts the capacitive change to a voltage signal on the output of the microphone.

Microphone Mounting. A backport MEMS microphone is mounted on the surface of a printed circuit board (PCB), with the microphone's aperture exposed through a cavity on the PCB (see the third column of Figure 2). The cavity, in turn, is part of an acoustic path that guides sound through holes (acoustic ports) in the device's chassis to the microphone's aperture. Finally, the device's acoustic ports typically have a fine mesh as shown in Figure 3 to prevent dirt and foreign objects from entering the microphone.

G. Laser Sources

Choice of a Laser. A laser is a device that emits a beam of coherent light that can stay narrow over a long distance and be focused to a tight spot. While many technologies exist for emitting coherent light, in this paper we focus on laser emitting diodes, which are common in consumer laser products such as laser pointers. Next, as the light intensity emitted from a laser diode is directly proportional to the diode's driving current, we can easily encode analog signals via the beam's intensity by using a laser driver capable of amplitude modulation.

Laser Safety and Availability. As strong, tightly focused lights can be potentially hazardous, there are standards in place regulating lights emitted from laser systems [32, 33] that divide lasers into classes based on the potential for injury resulting from beam exposure. In this paper, we are interested in two main types of devices, which we now describe.

Low-Power Class 3R Systems. This class contains devices whose output power is less than 5 mW at visible wavelength (400—700 nm, see Figure 4). While prolonged intentional eye exposure to the beam emitted from these devices might be harmful, these lasers are considered safe to human eyes for brief exposure durations. As such, class 3R systems form a good compromise between safety and usability, making these lasers common in consumer products such as laser pointers.

High-Power Class 3B and Class 4 Systems. Next, lasers that emit between 5 and 500 mW are classified as class 3B systems, and might cause eye injury even from short beam exposure durations. Finally, lasers that emit over 500 mW of power are categorized as class 4 systems, which can instantaneously cause blindness, skin burns and fires. As such, uncontrolled exposure to class 4 laser beams should be strongly avoided.

However, despite the regulation, there are reports of high-power class 3B and 4 systems being openly sold as "laser pointers" [34]. Indeed, while purchasing laser pointers from Amazon and eBay, we have discovered a troubling discrepancy between the rated and actual power of laser products. While the labels and descriptions of most products stated an output power of 5 mW, the actual measured power was sometimes as high as 1 W (i.e., $\times 200$ above the allowable limit).

III. Threat Model

The attacker's goal is to inject malicious commands into the targeted voice-controllable device, without being detected by the device's owner and without having physical device access. More specifically, we consider the following threat model.

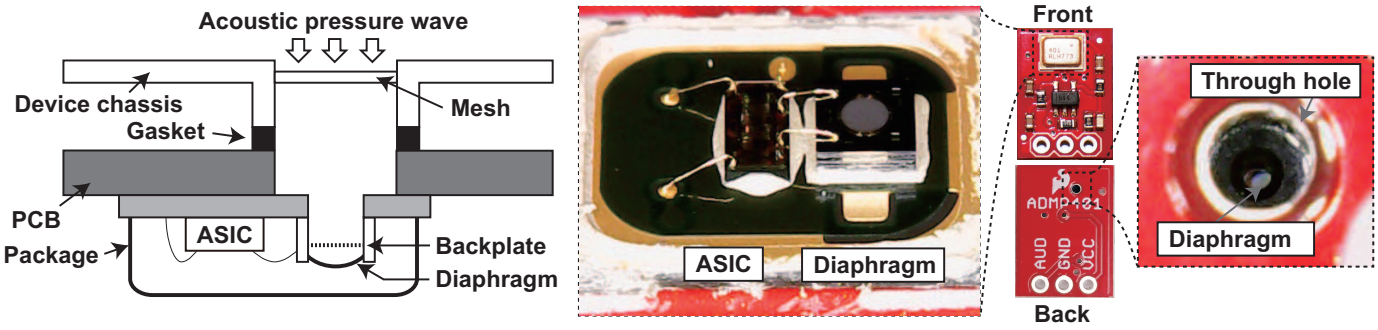

Fig. 2. MEMS microphone construction. (Left) Cross-sectional view of a MEMS microphone on a device. (Middle) A diaphragm and ASIC on a depackaged microphone. (Right) Magnified view of an acoustic port on PCB.

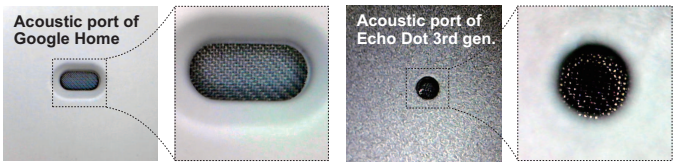

Fig. 3. Acoustic port of (Left) Google Home and (Right) Echo Dot 3rd generation. The ports are located on the top of the devices, and there are meshes inside the port.

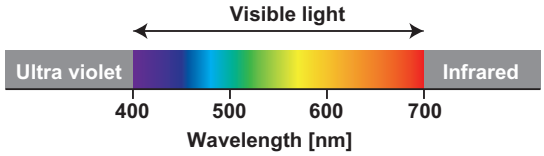

Fig. 4. Wavelength and color of light

No Physical Access or Owner Interaction. While the attacker is free to choose the target, we assume that the attacker does not have any physical access to the device being attacked. Thus, the attacker cannot press any buttons, alter voice-inaccessible settings, or compromise the device's software. Finally, we assume that the attacker cannot make the device's owner perform any useful interaction (like pressing a button or unlocking the screen).

Line of Sight. We do assume however that the attacker has (a remote) line of sight access to the target device and its microphone ports. We argue that such an assumption is reasonable, as voice-activated devices (such as smart speakers, thermostats, security cameras, or even phones) are often left visible to the attacker, including through closed glass windows.

Device Feedback. We note that the remote line of sight access to the target device also allows the attacker to observe the device's LED lights. Next, as these lights come on after the device properly recognized the wakeup word and show an unique patterns once the command was properly recognized and accepted, the attacker to remotely determine if an attack attempt was successful.

Device Characteristics. Finally, we also assume that the attacker has access to a device of a similar model as the target device. Thus, the attacker knows all the target's physical characteristics, such as location of the microphone ports and physical structure of the device's sound path. Such knowledge can easily be acquired by purchasing and analyzing a device of the same model before launching the attacks.

IV. Injecting Sound via Laser Light

A. Signal Injection Feasibility

In this section we explore the feasibility of injecting acoustic signals into microphones using laser light. We begin by describing our experimental setup.

Setup. We used a blue Osram PLT5 450B 450-nm laser diode connected to a Thorlabs LDC205C current driver. We increased the diode's DC current with the driver until it emitted a continuous 5.0 mW laser beam, while measuring light intensity using the Thorlabs S121C photo-diode power sensor. The beam was subsequently directed to the acoustic port on the SparkFun MEMS microphone breakout board mounting an Analog Devices ADMP401 MEMS microphone. Finally, we recorded the diode current and the microphone's output using a Tektronix MSO5204 oscilloscope. See Figure 5 for a picture of our setup. The experiments were conducted in a regular office environment, with typical ambient noise from human speech, computer equipment, and air conditioning systems.

Signal Injection. We used the current driver to modulate a sine wave on top of the diode's current I_t via amplitude modulation (AM), given by the following equation:

$$I_t = I_{DC} + I_{pp} \sin(2\pi f t) \quad (1)$$

where I_{DC} is a DC bias, I_{pp} is the peak-to-peak amplitude, and f is the frequency. In our case, we set I_{DC} = 26.2 mA, I_{pp} = 7 mA and f = 1 kHz, where the sine wave was generated using an on-board DAC on a laptop computer, and was supplied to the modulation port on the current driver through an audio amplifier (Neoteck NTK059 Headphone Amplifier). As the light intensity emitted by the laser diode is directly proportional to the current provided by the laser driver, this resulted in having the 1 kHz sine wave be directly encoded in the intensity of the light emitted by the laser diode.

Observing the Microphone Output. As can be seen in Figure 5, the microphone output clearly shows a 1 kHz sine wave that matches the frequency of the injected signal.

B. Characterizing Laser Audio Injection

Having successfully demonstrated the possibility of injecting audio signals via laser beams, we now proceed to characterize the light intensity response of the diodes (as a function of current) and the frequency response of the microphone to

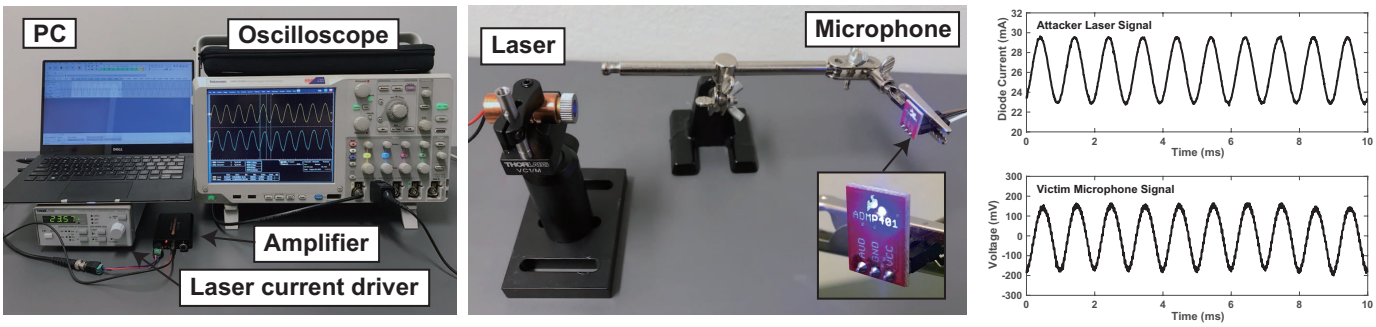

Fig. 5. Testing signal injection feasibility. (Left) A setup for signal injection feasibility composed of a laser current driver, PC, audio amplifier, and oscilloscope. (Middle) Laser diode with beam aimed at a MEMS microphone breakout board. (Right) Diode current and microphone output waveforms.

laser-based audio injection. To see the wavelength dependency, we also examine a 638-nm red laser (Ushio HL63603TG) in addition to the blue one used in the previous experiment.

Laser Current to Light Characteristics. We begin by examining the relationship between the diode current and the optical power of the laser. For this purpose, we aimed a laser beam at our Thorlabs S121C power sensor while driving the diodes with DC currents, i.e., $I_{pp} = 0$ in Equation 1. Considering the different properties of the diodes, the blue and red laser are examined up to 300 and 200 mA, respectively.

The first column of Figure 6 shows the current vs. light (I-L) curves for the blue and red lasers. The horizontal axis is the diode current I_{DC} and the vertical axis is the optical power. As can be seen, once the current provided to the laser is above the diode-specific threshold (denoted by I_{th}), the light power emitted by the laser increases linearly with the provided current. Thus, as $|\sin(2\pi ft)| < 1$, we have an (approximately) linear conversion of current to light provided that $I_{DC} - I_{pp}/2 > I_{th}$.

Laser Current to Sound Characteristics. We now proceed to characterize the effect of light injection on a MEMS microphone. We achieve this by aiming an amplitude-modulated (AM) laser beam with variable current amplitudes (I_{pp}) and a constant current offset (I_{DC}) into the aperture of an Analog Devices ADMP401 microphone, mounted on a breakout board. We subsequently monitor the peak-to-peak voltage of the microphone's output, plotting the resulting signal.

The second column of Figure 6 shows the relationship between the modulating signal I_{pp} and the resulting signal V_{pp} for both the blue and red laser diodes. The results suggest that the driving alternating current I_{pp} (cf. the bias current) is the key for strong injection: we can linearly increase the sound volume received by the microphone by increasing the driving AC current I_{pp}.

Choosing I_{DC} and I_{pp}. Given a laser diode that can emit a maximum average power of L mW, we would like to choose the values for I_{DC} and I_{pp} which result in the strongest possible microphone output signals, while having the average optical power emitted by the laser be less than or equal to L mW. From the leftmost column of Figure 6, we deduce that the laser's output power is linearly proportional to the laser's driving current $I_t = I_{DC} + I_{pp}\sin(2\pi ft)$, and the average power depends mostly on I_{DC}, as $I_{pp}\sin(2\pi ft)$ averages out to zero.

Thus, to stay within the power budget of L mW while obtaining the strongest possible signal at the microphone output, the attacker must first determine the DC current offset I_{DC} that results in the diode outputting light at L mW, and then subsequently maximize the amplitude of the microphone's output signal by setting $I_{pp}/2 = I_{DC} - I_{th}$.*

Characterizing the Frequency Response of Laser Audio Injection. Next, we set out to characterize the response of the microphone to different frequencies of sound signals injected via laser beams. We use the same operating points as the previous experiment, and set the tone's amplitude such that it fits with the linear region ($I_{DC} = 200$ mA and $I_{pp} = 150$ mA for the blue laser, and $I_{DC} = 150$ mA and $I_{pp} = 75$ mA for the red laser). We then record the microphone's output levels while changing the frequency f of the light-modulated sine wave.

The third column of Figure 6 shows the obtained frequency response for both blue and red lasers. The horizontal axis is the frequency while the vertical axis is the peak-to-peak voltage of the microphone output. Both lasers have very similar responses, covering the entire audible band 20 Hz–20 kHz, implying the possibility of injecting any audio signal.

Choice of Laser. Finally, we note the color insensitivity of injection. Although blue and red lights are on the other edges on the visible spectrum (see Figure 4), the levels of injected audio signal are in the same range and the shapes of the frequency-response curves are also similar. Therefore, color has low priority in choosing a laser compared to other factors for making LightCommands. In this paper, we consistently use the 450-nm blue laser mainly because of (i) better availability of high-power diodes and (ii) the advantage in focusing because of a shorter wavelength.

C. Mechanical or Electrical Transduction?

In this section we set out to investigate whether our light-based acoustic signal injection is due to physical movements of the microphone's diaphragm (i.e., light-induced mechanical

*We note here that the subtraction of I_{th} is designed to ensure that $I_{DC} - I_{pp}/2 > I_{th}$, meaning that the diode stays in its linear region thereby avoiding signal distortion.

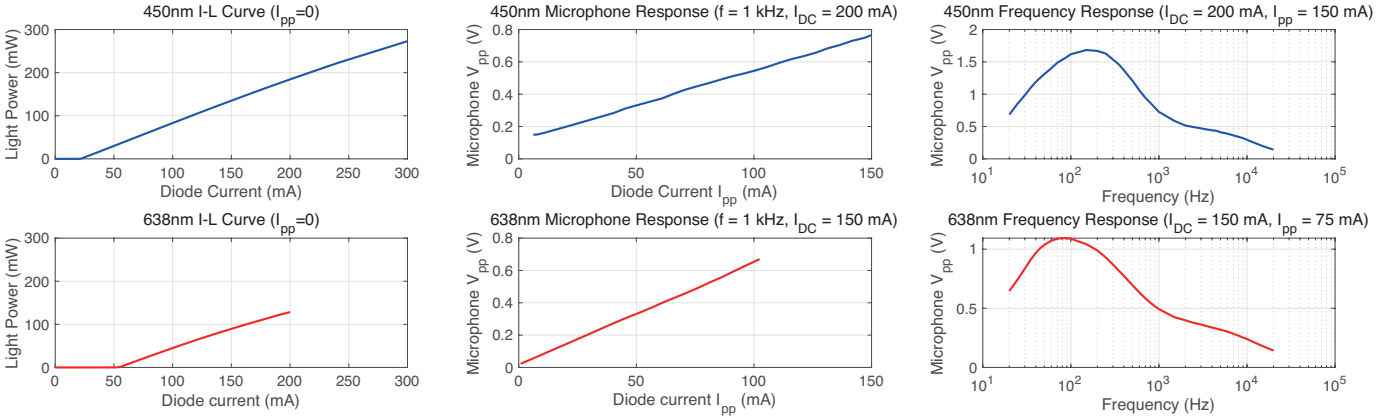

Fig. 6. Characteristics of the 450-nm blue laser (first row) and the 638-nm red laser (second row). (First column) Current-light DC characteristics. (Second column) Microphone response for a 1 kHz tone with different amplitudes. (Third column) Frequency responses of the overall setup for fixed bias and amplitude.

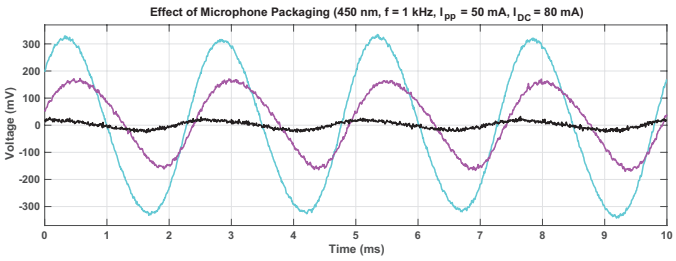

Fig. 7. The microphone's response to laser injection with different mechanical conditions: (Cyan) a baseline measurement without modification, (Magenta) the microphone with its metal package removed, and (Black) a transparent glue on microphone's diaphragm.

vibration), or by another mechanism such as the photoelectric effect. We achieve this via a series of measurements that gradually modify the microphone's mechanical condition while leaving its optical condition intact. Using an ADMP401 microphone, we first take a baseline measurement without any modification. Next, we remove the microphone's pressure reference chamber by opening the package covering the diaphragm and ASIC (as shown in the second column of Figure 2). Finally, we dampen the diaphragm's movement by putting glue directly on the opened and exposed diaphragm[†]. We note that the microphone's optical properties are unchanged as the glue is transparent and applied from the diaphragm's back side, leaving the acoustic port intact.

Figure 7 presents the resulting voltage signals from the microphone in the three conditions illuminated with the same laser beam (the blue laser with $f = 1$ kHz, $I_{pp} = 50$ mA, and $I_{DC} = 80$ mA). As can be seen, the modification decreases the amplitude of the signal detected by the microphone, and the signal after the glue application is less than 10% of the original signal. We thus attribute our light-based signal injection results to mechanical movements of the microphone's diaphragm, which are in turn translated to output voltage by the microphone's internal circuitry.

[†]We used transparent and non-conductive glue (Gorilla Super Glue), and conducted the measurement while the glue is wet since surface tension during the curing process can damage the chip.

V. Attacking Various Voice-Controllable Systems

In this section we evaluate our attack on sixteen popular VC systems. We aim to find out the minimal laser power required by the attacker in order to gain control over the VC system under ideal conditions as well as the maximal distance that such control can be obtained under more realistic conditions.

Target Selection. We benchmark our attack against several consumer devices which have voice control capabilities (see Table I). We aim to test the most popular voice assistants – namely Alexa, Siri, Portal, and Google Assistant. While we do not claim that our list is exhaustive, we do argue that it does provide some intuition about the vulnerability of popular VC systems to laser-based voice injection attacks. Next, to explore how different hardware variations (rather than algorithmic variations) affect our attack performance, we benchmark our attack on multiple devices running the same voice recognition backend: Alexa, Siri, Portal and Google Assistant, as summarized in Table I. For some devices, we examine different generations to explore the differences on attack performance for various hardware models. Finally, we also considered third-party devices with built-in speech recognition, such as the EcoBee thermostat.

A. Exploring Laser Power Requirements

In this section we aim to characterize the minimal laser power required by the attacker under ideal conditions to take control over a voice-activated system. Before describing our experimental setup, we show our selection of benchmarked voice commands and experiment success criteria.

Command Selection. We have selected four different voice commands that represent common operations performed by voice operated systems.

- **What Time Is It?** This command was selected to serve as the baseline of our experiments, as it does not require the device to perform nearly any operation besides correctly identifying the command and accessing the Internet to recover the current time.

- **Set the Volume to Zero.** Here, we demonstrate the attacker's ability to control the output of the VC system. We expect this to be the first voice command issued by the attacker, in an attempt to avoid attracting attention from the target's legitimate owner.
- **Purchase a Laser Pointer.** With this command we show how an attacker can potentially place order for various products on behalf (and at the expense) of users. The attacker can subsequently wait for delivery near the target's residents and collect the purchased item.
- **Open the Garage Door.** Finally, and perhaps most devastatingly, we show how an attacker can interact with additional systems which have been linked by the user to the targeted VC system. While the garage door opener is one such example with clear security implications, we discuss other examples in Section VI.

Command Generation. We have generated audio recordings of all four of the above commands using a common audio recording system (e.g., Audacity). Each command recording was subsequently appended to a recording of the wake word corresponding to the device being tested (e.g., Alexa, Hey Siri, Hey Portal, or OK, Google) and normalized to adjust the overall volume of the recordings to a constant value. We obtained a resulting corpus of 16 complete commands. Finally, for each device, we injected four of the complete commands (those beginning with the device-appropriate wake word) into the device's microphone using the setup described below and observed the device's response.

Verifying Successful Injection. We consider a command injection attempt as successful in case the device somehow indicates the correct interpretation of the command. For devices with screens (such as phones and screen enabled speakers), we considered an attempt successful when the device correctly displayed a transcription of the light-injected voice command. For screen-less devices (e.g., smart speakers), we manually examined the command log of the account associated with the device for the correct command transcription.

Attack Success Criteria. For a given power budget, distance, and command, we consider the injection successful in case the device correctly recognized the command during three consecutive injection attempts. We take this as an indication that the power budget and distance are sufficient for achieving a near-perfect success probability assuming suitable aiming. Next, we consider an attack successful for a given power budget and distance in case all of our four commands where successfully injected to the device during three consecutive injection attempts. Like in the individual command case, we take this as an indication that the considered power budget and distance is sufficient for a high probability successful commands injection. We note that this criteria is conservative, as some commands are easier to inject than others, presumably due to their phonetical properties. As such, the results in this section should be seen as a conservative estimate of what an attacker can achieve for each device assuming good environmental conditions (e.g., quiet surroundings and suitable aiming) while better results in terms of distance and power can be achieved if less than perfect accuracy is considered.

Voice Customization and Security Settings. For the experiments conducted in this section, we left all the device's settings in their default configuration. Next, in embedded Alexa and Google VC systems (e.g., smart speakers, cameras, etc.) voice customization is off by default, meaning that the device will operate on commands spoken by any voice. Meanwhile, for phone and tablet devices (which are typically operated by a single user), we left the voice identification in its default activated setting. For such devices, to ascertain the minimal required power for a successful attack, we trained the VC system with a human voice and subsequently inject the audio recording of the commands spoken using the same voice. Finally, in Section V-C, we discuss bypassing various voice matching mechanisms.

Experimental Setup. We use the same blue laser and Thorlabs laser driver as in Section IV-A, aiming the laser beam at microphone ports of the devices listed in Table I from a distance of about 30 cm. To control the surrounding environment, the entire setup was placed in a metal enclosure, with opaque bottom and sides and with a dark red semi-transparent acrylic top plate, designed to block blue light. See Figure 8. As the goal of the experiments described in this section is to ascertain the minimal required power for a successful attack on each device, we have used a pair of electrically controlled scanning mirrors (40 Kbps high-speed laser scanning system for laser shows) to precisely place the laser beam in the center of the device's microphone port. Before each experiment we manually focused the laser so that the laser spot size hitting the microphone is minimal.

For aiming at devices whose microphone port is covered with cloth (e.g., Google Home Mini shown in Figure 9), the position of the microphone ports can be determined using an easily-observable reference point such as the device's wire connector or LED array. Finally, we note that the distance between the microphone and the reference point is easily obtainable by the attacker either by exploring his own device, or by referring to online teardown videos [35].

Experimental Results. The fifth column of Table I presents a summary of our results. While the power required from the attacker varies from 0.5 mW (Google Home) to 60 mW (Galaxy S9), all the devices are susceptible to laser-based command injection. Finally, we note that the microphone port of some devices (e.g., Google Home Mini) is covered with fabric and / or foam. While we conjecture that this attenuates optical power, as Table I shows, the attack is still possible.

Finally, we note that the experiments done in this section are performed under ideal conditions, at close range and with the aid of electronic aiming mirrors. Thus, in Section V-B we report on attack results under more realistic conditions with respect to distance and aiming.

B. Exploring Attack Range

In this section we set out to explore the effective range of our attack under more realistic attack conditions.

TABLE I
TESTED DEVICES WITH MINIMUM ACTIVATION POWER AND MAXIMUM DISTANCE ACHIEVABLE AT THE GIVEN POWER OF 5 MW AND 60 MW. A 110 M LONG HALLWAY WAS USED FOR 5 MW TESTS WHILE A 50 M LONG HALLWAY WAS USED FOR TESTS AT 60 MW.

Device	Backend	Category	Authen-tication	Minimum Power [mW]	Max Distance at 60 mW [m]	Max Distance at 5 mW [m]
Google Home	Google Assistant	Speaker	No	0.5	50+	110+
Google Home Mini	Google Assistant	Speaker	No	16	20	—
Google Nest Cam IQ	Google Assistant	Camera	No	9	50+	—
Echo Plus 1st Generation	Alexa	Speaker	No	2.4	50+	110+
Echo Plus 2nd Generation	Alexa	Speaker	No	2.9	50+	50
Echo	Alexa	Speaker	No	25	50+	—
Echo Dot 2nd Generation	Alexa	Speaker	No	7	50+	—
Echo Dot 3rd Generation	Alexa	Speaker	No	9	50+	—
Echo Show 5	Alexa	Speaker	No	17	50+	—
Echo Spot	Alexa	Speaker	No	29	50+	—
Facebook Portal Mini	Alexa + Portal	Speaker	No	18	5	—
Fire Cube TV	Alexa	Streamer	No	13	20	—
EcoBee 4	Alexa	Thermostat	No	1.7	50+	70
iPhone XR (Front Mic)	Siri	Phone	Yes	21	10	—
iPad 6th Gen	Siri	Tablet	Yes	27	20	—
Samsung Galaxy S9 (Bottom Mic)	Google Assistant	Phone	Yes	60	5	—
Google Pixel 2 (Bottom Mic)	Google Assistant	Phone	Yes	46	5	—

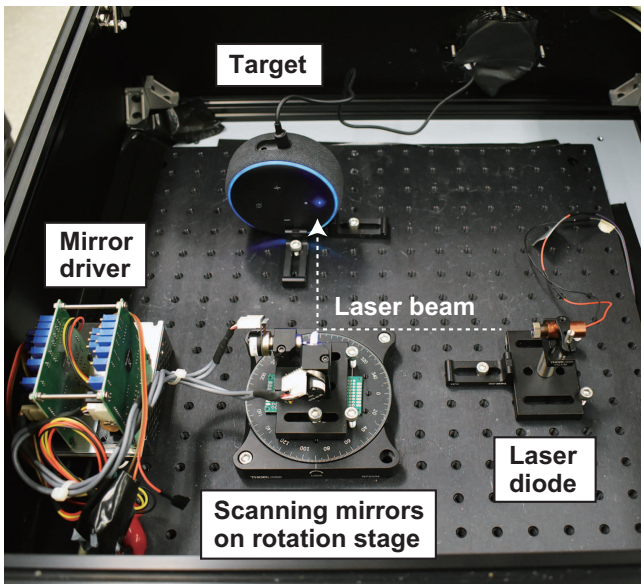

Fig. 8. Setup for exploring laser power requirements: the laser and target are arranged in the laser enclosure. The laser spot is aimed at the target acoustic port using electrically controllable scanning mirrors inside the enclosure. The enclosure's top red acrylic cover was removed for visual clarity.

Experimental Setup. From the experiments performed in Section V-A we note that about 60 mW of laser power is sufficient for successfully attacking all of our tested devices (at least under ideal conditions). Thus, in this section we benchmark the range of our attack using two power budgets.

- **60 mW High-Power Laser.** As explained in Section II-G, we frequently encountered laser pointers whose measured power output was above 60 mW, which greatly exceeds legal 5 mW restrictions. Thus, emulating an attacker which does not follow laser safety protocols for consumer devices, we benchmark our attack using 60 mW lasers, which is sufficient for successfully attacking all of our tested devices in the previous experiment.

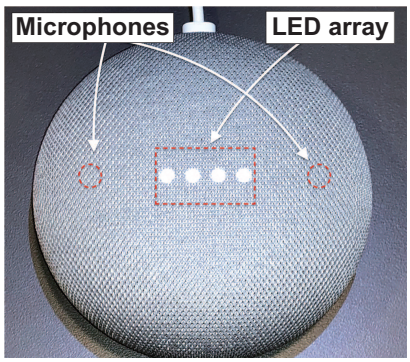

Fig. 9. Google Home Mini. Notice the cloth-covered microphone ports.

- **5 mW Low-Power Laser.** Next, we also explore the maximum range of a more restricted attacker, which is limited to the maximum amount of power allowed in the U.S. for consumer laser pointers, namely 5 mW.

Laser Focusing and Aiming. For large attack distances (tens of meters), laser focusing requires a large diameter lens and cannot be done via the small lenses that are typically used for laser pointers. Thus, we mounted our laser to an Opteka 650-1300 mm high-definition telephoto lens, with 86 mm diameter (Figure 1(left)). Finally, to simulate realistic aiming conditions for the attacker, we avoided the use of electronic scanning mirrors (used in Section V-A) and mounted the lens and laser on a geared camera head (Manfrotto 410 Junior Geared Tripod Head) and tripod. Laser aiming and focusing was done manually, with the target also mounted on a (separate) tripod. See Figure 1 for a picture of our setup.

Test Locations and Experimental Procedure. As eye exposure to a 60 mW laser is potentially dangerous, we blocked off a 50 meter long corridor in our office building and performed the experiments at night. However, due to safety reasons, we were unable to obtain a longer corridor for our high-power tests. For lower-power attacks, we performed

the experiments in a 110 meter long corridor connecting two buildings (see Figure 1(top)). In both cases, we fixed the target at distance and subsequently adjusted the optics, obtaining the smallest possible laser spot. We then regulated the diode current so that the target is illuminated with 5 or 60 mW respectively. Finally, the corridor is illuminated with regular fluorescent lamps at office-level brightness while the ambient acoustic noise in both experiments was about 46 dB (measured using a General Tools DSM403SD sound level meter).

Experimental Results. Table I contains a summary of our distance-benchmarking results. With 60 mW laser power, we have successfully injected voice commands to all the tested devices from a distance of several meters. For devices that reached the maximum 50 meters in the high-power experiment, we also conducted the low-power experiment in the 110 m hallway. Untested devices are indicated by '—' in Table I because of their high minimal activation power.

While most devices require a 60 mW laser for successful command injection (e.g., an non standard complaint laser pointer), some popular smart speakers such as Google Home and Eco Plus 1st and 2nd Generation are particularly sensitive, allowing for command injection even with 5 mW power over tens of meters. Finally, as our attacks were conducted in 50 and 110 meter hallways (for 60 and 5 mW lasers, respectively) for some devices, we had to stop the attack when the maximal hallway length was reached. We mark this case with a + sign near the device's range in the appropriate column.

C. Attacking Speaker Authentication

We begin by distinguishing between speaker recognition features, which are designed to recognize voice of specific users and personalize the device's content, and speaker authentication features which is designed to restrict access control to specific users. While not the main topic of this work, in this section we now discuss both features in the context of light-based command injection.

No Speaker Authentication for Smart Speakers. We begin by observing that for smart speaker type devices (which are the main focus of this work), speaker recognition is off by default at the time of writing. Next, even if the feature is enabled by careful users, smart speakers are designed to be utilized by multiple people. Thus, their speaker recognition features are usually limited to content personalization rather than authentication, treating unknown voices as guests. Empirically verifying this, we found that Google Home and Alexa smart speakers block voice purchasing for unrecognized voices (presumably as they do not know which account should be billed for the purchase) while allowing previously-unheard voices to execute security critical voice commands such as unlocking doors. Finally we note that at the time of writing voice authentication is not available for smart speaker devices, which are common home smart assistant deployments.

Phone and Tablet Devices. Next, while not the main focus of this work, we also investigated the feasibility of light command injection into phone and tablet devices. For such devices, speaker authentication is enabled by default due to the high processing power and single owner use.

Overview of Voice Authentication. After training using samples of owner's voice speaking specific sentences, the tablet of phone continuously listens to the microphone, acquiring a set of voice samples. These are in turn fed into deep learning models which recognize if the voice sample corresponds to assistant-specific wake up words (e.g., "Hey Siri" or "Ok Google") spoken by the owner. Finally, in case of a successful detection of features matching the owner's voice, the phone or tablet proceeds to parse and subsequently execute the voice command.

Bypassing Voice Authentication. Intuitively, an attacker can defeat the speaker authentication feature using authentic voice recordings of the device's legitimate owner speaking the desired voice commands. Alternatively, if no such recordings are available, DolphinAttack [4] suggests using speech synthesis techniques, such as splicing relevant phonemes from other recordings of the owner's voice, to construct the commands.

Wake-Only Security. However, during our experiments we found that speaker recognition is used by Google and Apple to only verify the wake word, as opposed to the entire command. For example, Android and iOS phones trained to recognize a female voice, correctly execute commands where the wake word was spoken by the female voice, while the rest of the command was spoken using a male voice. Thus, to bypass voice authentication, an attacker only needs a recording of the device's wake word in the owner's voice (which can be obtained by recording any command spoken by the owner).

Reproducing Wake Words. Finally, we explore the possibility of using Text-To-Speech (TTS) techniques for reproducing the owner's voice saying the wake words for a tablet or phone based voice assistant. To that aim, we repeat the phone and tablet experiments done in Sections V-A, V-B and Table I, training all the phone and tablet devices with a human female voice. We then used NaturalReader [36], an online TTS tool for generating the wake words specific for each device, hoping that the features of one of the offered voices will match the human voice used for training. See Table II for device-specific voice configurations matching the female voice used for training. Next, we concatenate the synthetically-generated wake word spoken in a female voice to a voice command pronounced by a male native-English speaker. Using these recordings, we successfully replicated the minimal power and maximum distance results as presented in Table I.

We thus conclude that while voice recognition is able to enforce some similarity between the attacker's and owner's voices, it does not offer sufficient entropy to form an adequate countermeasure to command injection attacks. In particular, out of the 18 English voices supported by NaturalReader, we were able to find an artificial voice matching the human female voice used for training for all 4 of the tablet and phone devices considered in this work. Finally, we did not test the ability to match voices for devices other than phones and tablets, as voice authentication is not available for smart speaker devices at the time of writing.

TABLE II
BYPASSING VOICE AUTHENTICATION ON PHONE AND TABLET DEVICES

Device	Assistant	TTS Service	Voice Name
iPhone XR	Siri	NaturalReader	US English Heather
iPad 6th Gen	Siri	NaturalReader	US English Laura
Galaxy S9	Google Assistant	NaturalReader	US English Laura
Pixel 2	Google Assistant	NaturalReader	US English Laura

VI. EXPLORING VARIOUS ATTACK SCENARIOS

The results of Section V clearly demonstrate the feasibility of laser-based injection of voice commands into voice-controlled devices across large attack distances. In this section, we explore the security implications of such an injection, as well as experiment with more realistic attack conditions.

A. A Low-Power Cross-Building Attack

For the long-range attacks presented in Section V-B, we deliberately placed the target device so that the microphone ports are facing directly into the laser beam. While this is realistic for some devices (who have microphone ports on their sides), such an arrangement is artificial for devices with top-facing microphones (unless mounted sideways on the wall).

In this section we perform the attack under a more realistic conditions where an attacker aims from another higher building at a target device placed upright on a window sill.

Experimental Conditions. We use the laser diode, telephoto lens and laser driver from Section V, operating the diode at 5 mW (equivalent to a laser pointer). Next, we placed a Google Home device (which only has top-facing microphones) upright near a window, on a fourth-floor office (15 meters above the ground). The attacker's laser was placed on a platform inside a nearby bell tower, located 43 meters above ground level. Overall, the distance between the attacker's and laser was 75 meters, see Figure 10 for the configuration.

Laser Focusing and Aiming. As in Section V-B, it is impossible to focus the laser using the small lens typically used for laser pointers. We thus mounted the laser to an Opteka 650-1300 mm telephoto lens. Next, to aim the laser across large distances, we have mounted the telephoto lens on a Manfrotto 410 geared tripod head. This allows us to precisely aim the laser beam on the target device across large distances, achieving an accuracy far exceeding the one possible with regular (non-geared) tripod heads where the attacker's arm directly moves the laser module. Finally, in order to see the laser spot and the device's microphone ports from far away, we have used a consumer-grade Meade Infinity 102 telescope. As can be seen in Figure 10 (left), the Google Home microphone's ports are clearly visible through the telescope.[‡]

Attack Results. We have successfully injected commands into the Google Home target in the above described conditions. We note that despite its low 5 mW power and windy conditions (which caused some beam wobbling due to laser movement), the laser beam successfully injected the voice command while penetrating a closed double-pane glass window. While causing negligible reflections, the double-pane window did not cause any visible distortion in the injected signal, with the laser beam hitting the target's top microphones at an angle of 21.8 degrees. We conclude that cross-building laser command injection is possible, at large distances and under realistic attack conditions.

[‡]Figure 10 (left) was taken via a cell phone camera attached to the telescope's eyepiece. Unfortunately, due to imperfect phone-eyepiece alignment, the outcome is slightly out of focus and the laser spot is over saturated. However, the Google Home was in sharp focus with a small laser spot when viewed directly by a human observer.

B. Attacking Authentication

Some of the current generation of VC systems attempt to protect unauthorized execution of sensitive commands by requiring additional user authentication step. For phone and tablet devices, the Siri and Alexa apps require the user to unlock the phone before executing certain commands (e.g., unlock front door, disable home alarm system). However, for devices that do not have other form of inputs beside the user's voice (e.g., voice-enabled smart speakers, cameras, and thermostats) a digit-based PIN code is used to authenticate the user before critical commands are performed.

PIN Eavesdropping. The PIN number spoken by the user is inherently vulnerable to eavesdropping attacks, which can be performed remotely using a laser microphone (measuring the acoustic vibration of a glass window using a laser reflection [37]), or using common audio eavesdropping techniques. Moreover, within an application the same PIN is used to authenticate more than one critical command (e.g., "unlock the car" and "start the engine") while users often re-use PIN numbers across different applications. In both cases, increasing the number of PIN-protected commands ironically increases the opportunity for PIN eavesdropping attacks.

PIN Brute forcing. We also observed incorrect implementation of PIN verification mechanisms. While Alexa naturally supports PIN authentication (limiting the user to three wrong attempts before requiring interaction with a phone application), Google Assistant delegates PIN authentication to third-party device vendors that often lack security experience.

Evaluating this design choice, we have investigated the feasibility of PIN brute forcing attacks on an August Smart Lock Pro, which is the most reviewed smart lock on Amazon at the time of writing. First, we have discovered that August does not enforce a reasonable PIN code length, allowing the user to set a PIN containing anywhere from 1 to 6 digits for door unlocking. Next, we observed that the August lock does not limit the number of wrong attempts permitted by the user, nor does the lock implement a time delay mechanism between incorrect attempts. Thus, all the attacker has to do to unlock the target's door is to simply enumerate all possible PIN codes.

Empirically verifying this, we have written a Python implementation that enumerates all 4-digit PIN numbers using a synthetic voice. After each unsuccessful attempt, the Google home device responded with "Sorry, the security code is incorrect, can I have your security code to unlock the front door?" only to have our program speak the next PIN candidate. Overall, a single unlock attempt lasted about 13 seconds,

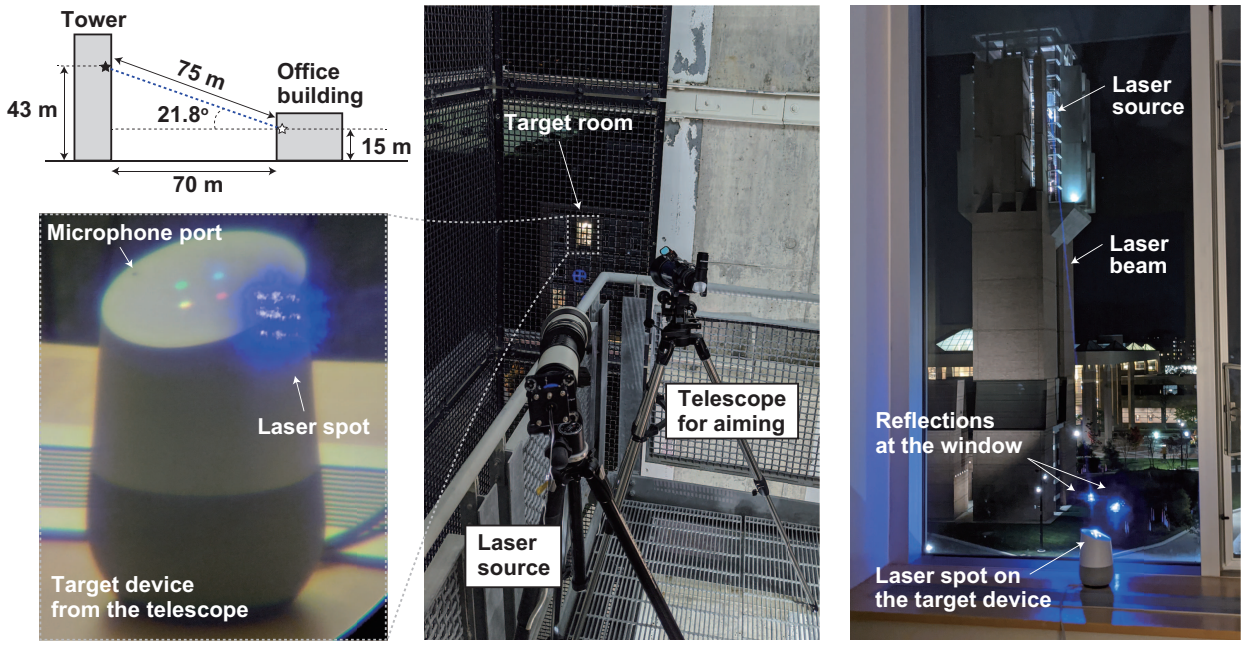

Fig. 10. Setup for the low-power cross-building attack: (Top left) Laser and target arrangement. (Bottom left) Picture of the target device as visible through the telescope, with the microphone ports and laser spot clearly visible. (Middle) Picture from the tower: laser on telephoto lens aiming down to the target. (Right) Picture from the office building: laser spot on the target device.

requiring 36 hours to enumerate the entire 4-digit space (3.6 hours for 3 digits). In both the 3- and 4-digit case, the door was successfully unlocked when the correct PIN was reached.

PIN Bypassing. Finally, we have discovered that while commands like "unlock front door" for August locks or "disable alarm system" for Ring alarms require PIN authentication, other commands such as "open the garage door" using an assistant-enabled garage door opener[§] generally do not require any authentication. Thus, even if one command is unavailable, the attacker can often achieve a similar goal by using other commands.

C. Attacking Cars

Many modern cars have Internet-over-cellular connectivity, allowing their owners to perform certain operations via a dedicated app on their mobile devices. In some cases, this connectivity has further evolved (either by the vendor or by a third-party) in having the target's car be connected to a VC system, allowing voice unlocking and/or pre-heating (which often requires engine start). Thus, a compromised VC system might be used by an attacker to gain access to the target's car.

In this section we investigate the feasibility of such attacks, using two major car manufactures, namely Tesla and Ford.

Tesla. Tesla cars allow their owner to interact with the car using a dedicated Tesla-provided phone app. After installing the app on our phone and linking it to a Tesla Model S vehicle, we have installed the "EV Car"[¶] integration, linking it to the vehicle. While "EV Car" is not officially provided by Tesla, after successful configuration using the vehicle's owner credentials, we were able to get several capabilities. These included getting information about the vehicle's current location[‖], locking and unlocking the doors and trunk, starting and stopping the vehicle's charging and the climate control system. Next, we note that we were able to perform all of these tasks using only voice commands, without the need of a PIN number or key proximity. Finally, we were not able to start the car without key proximity.

Ford Cars. For newer vehicles, Ford provides a phone app called "FordPass", that connects to the car's Ford SYNC system, and allows the owner to interact with the car over the Internet. Taking the next step, Ford also provides a FordPass Google Assistant integration[**] with similar capabilities as the "EV Car" integration for Tesla. While Ford implemented PIN protection for critical voice commands like remote engine start and door unlocking, like in the case of August locks, there are no mechanisms in place to prevent PIN brute forcing. Finally, while we were able to remotely open the doors and start the engine, shifting the vehicle out of "Park" immediately stopped the engine, preventing the unlocked car from being driven.

D. Exploring Stealthy Attacks

The attacks described so far can be spotted by the user of the targeted VC system in three ways. First, the user might notice the light indicators on the target device following a successful command injection. Next, the user might hear the device acknowledging the injected command. Finally, the user might notice the spot while the attacker tries to aim the laser at the target microphone port.

[§]https://www.garadget.com/

[¶]https://assistant.google.com/services/a/uid/000000196c7e079e?hl=en

[‖]Admittedly, the audible location is of little use to a remote attacker who is unable to listen in on the speaker's output.

[**]https://assistant.google.com/services/a/uid/000000ac1d2afd15

While the first issue is a limitation of our attack (and in fact of any command injection attack), in this section we explore the attacker's options for addressing the remaining two issues.

Acoustic Stealthiness. To tackle the issue of the device owner hearing the targeted device acknowledging the execution of voice command (or asking for a PIN number during the brute forcing process), the attacker can start the attack by asking the device to lower its speaker volume. For some devices (EcoBee, Google Nest Camera IQ, and Fire TV), the volume can be reduced to completely zero, while for other devices it can be set to barely-audible levels. Moreover, the attacker can also abuse device features to achieve the same goal. For Google Assistant, enabling the "do not disturb mode" mutes reminders, broadcast messages and other spoken notifications. For Amazon Echo devices, enabling "whisper mode" significantly reduces the volume of the device responses during the attack to almost inaudible levels.

Optical Stealthiness. Next, to avoid having the owner spot the laser light aimed at the target device, the attacker can use an invisible laser wavelength. Experimentally verifying this, we replicated the attack on Google Home device from Section V-A using a 980-nm infrared laser (Lilly Electronics 30 mW laser module). We then connected the laser to a Thorlabs LDC205C driver, limiting its power to 5 mW. Finally, as the spot created by infrared lasers is invisible to human eyes, we aimed the laser using a smartphone camera (as these typically do not contain infrared filters).

Using this setup, we have successfully injected voice commands to a Google Home at a distance of about 30 centimeters in the same enclosure as Section V-A. The spot created by the infrared laser was barely visible using the phone camera, and completely invisible to the human eye. Finally, not wanting to risk prolonged exposure to invisible (but eye damaging) laser beams, we did not perform range experiments with this setup. However, given the color insensitivity described in Section IV-A, we conjecture that results similar to those obtained in Section V-B could be obtained here as well.

E. Avoiding the Need for Precise Aiming

Another limitation of the attacks described so far is the need to aim the laser spot precisely on the target's microphone ports. While we achieved such aiming in Section VI-A by using geared camera tripod heads, in this section we show how the need for precise aiming can be avoided altogether.

An attacker can use a higher-power laser and trade its power with a larger laser spot size, which makes aiming considerably easier. Indeed, laser modules higher than 4,000 mW are commonly available on common e-commerce sites for laser engraving. Since we could not test such a high-power laser in an open-air environment for a safety concerns, we decided to use a laser-excited phosphor flashlight (Acebeam W30 with 500 lumens), which is technically a laser but sold as a flashlight with beam-expanding optics.

To allow for voice modulation, we modified the flashlight by removing its original current driver and connecting its diode terminals to the Thorlabs LDC240C laser driver (see Figure 11). Then, the experimental setup of Section V-B is replicated except that the laser diode and telephoto lens is replaced with the flashlight. Using this setup, we successfully injected commands to a Google Home device at a range of about 10 meters, while running the flashlight at an output power of 1 W. Next, as can be seen in Figure 11, the beam spot created by the flashlight is large enough to cover the entire target (and its microphone ports), without the need to use additional focusing optics and aiming equipment. However, we note that while the large spot size helps for imprecise aiming, the flashlight's quickly diverging beam also limits the attack's maximal distance.

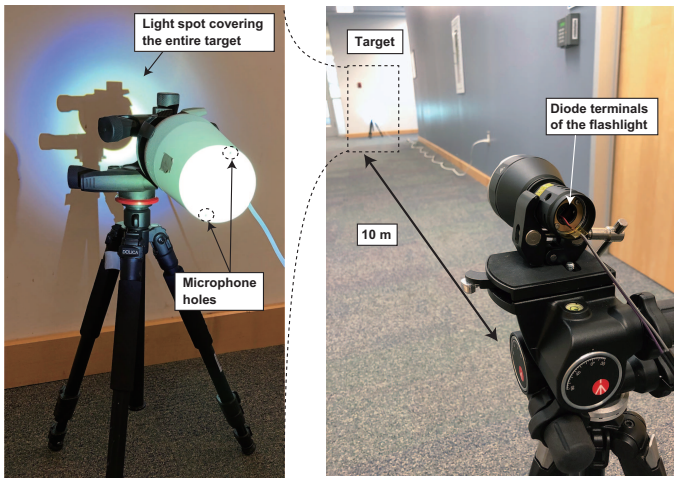

Fig. 11. Setup with laser flashlight to avoid precise aiming. (Left) Target device illuminated by the flashlight. (Right) Modified laser flashlight mounted on a geared tripod head aiming at the target 10 meters away.

Finally, the large spot size created by the flashlight (covering the entire device surface) can also be used to inject the sound into to multiple microphones simultaneously, thereby potentially defeating software-based anomaly detection countermeasures described in Section VII.

F. Reducing the Attack Costs

While the setups used for all the attacks described in this paper are built using readily available components, some equipment (such as the laser driver and diodes) are intended for lab use, making assembly and testing somewhat difficult for a non-experienced user. In this section we present a low-cost setup that can be easily constructed using improvised means and off-the-shelf components.

Laser Diode and Optics. Modifying off-the-shelf laser pointers can be an easy way to get a laser with collimation optics. In particular, cheap laser pointers often have no current regulators, having their anodes and cathodes directly connected to the batteries. Thus, we can easily connect a current driver to the pointer's battery connectors via alligator clips. Figure 12 shows a cheap laser pointer based setup, available at $18 for 3 pieces at Amazon.[††]

[††]https://www.amazon.com/gp/product/B075K69DTQ

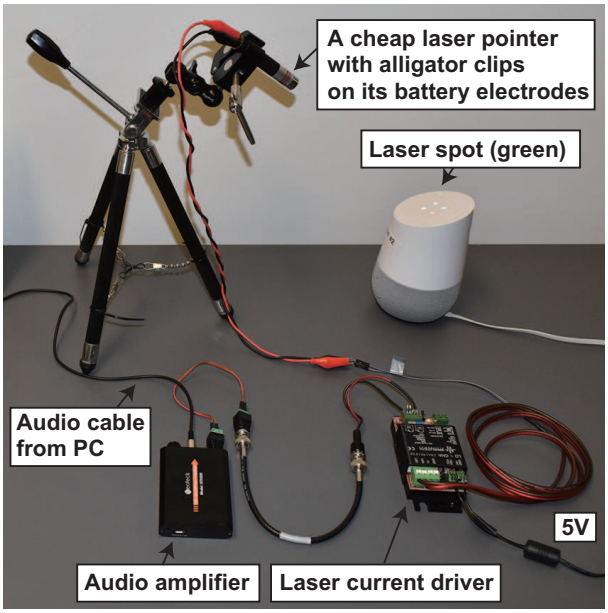

Fig. 12. Setup for low-cost attack: a laser current driver connected to a laser pointer attacking a Google Home device.

Laser Driver. The laser current driver with analog modulation port is the most specialized instrument in the attacker's setup. We used the scientific-grade laser drivers that cost about $1,500, however, there are cheaper alternatives such as the Wavelength Electronics LD5CHA current driver available at a cost of about $300.

Sound Source and Experimental Results. Finally, the attacker needs a method for playing recorded audio commands. We used an ordinary on-board laptop sound card (Dell XPS 15 9570), amplified using a Neoteck NTK059 Headphone Amplifier ($30 on Amazon). See Figure 12 for a picture of a complete low-cost setup. We have experimentally verified successful command injection using this setup into a Google Home target, located at a distance of 40 meters (with the main range limitation being the laser focusing optics and an artificially-limited power budget of 5 mW for safety reasons).

VII. Countermeasures and Limitations

A. Software-Based Approach

As discussed in Section VI-B, an additional layer of authentication can be effective at somewhat mitigating the attack. Alternatively, in case the attacker cannot eavesdrop on the device's response (for example since the device is located far away behind a closed window), having the VC system ask the user a simple randomized question before command execution can be an effective way at preventing the attacker from obtaining successful command execution. However, we note that adding an additional layer of interaction often comes at a cost of usability, limiting user adoption.

Finally, manufacturers can attempt to use sensor fusion techniques [38] in the hopes of detecting light-based command injection. More specifically, common VC systems often have multiple microphones, which should receive similar signals due to the omnidirectional nature of acoustic waves propagation. Meanwhile, when the attacker uses a single laser, only a single microphone receives a signal while the others receive nothing. Thus, manufacturers can attempt to detect such anomalies, ignoring the injected commands. However, we note that attackers can defeat such comparison countermeasures by simultaneously injecting lights to all the device's microphones using wide beams, see Section VI-E.

Finally, LightCommands are very different compared to normal audible commands. For sensor-rich devices like phones and tablets, sensor-based intrusion detection techniques [39] can potentially be used to identity and subsequently block such irregular command injection. We leave further exploration of this direction to future work.

B. Hardware-Based Approach

It is possible to reduce the amount of light reaching the microphone's diaphragm using a barrier that physically blocks straight light beams, while allowing acoustic pressure waves to detour around it. Performing a literature review on proposed microphone designs, we have found several such suggestions, mainly aimed to protect microphones from sudden pressure spikes. For example, the designs in Figure 13 have a silicon plate or movable shutter, both of which eliminate the line of sight to the diaphragm [40]. It is important to note however, that such barriers should be opaque to all light wavelengths (including infrared and ultraviolet), preventing the attacker from going through the barrier using a different colored light. Finally, a light-blocking barrier can be also implemented at the device level, by placing a non-transparent cover on top of the microphone hole, which attenuates the amount of light hitting the microphone.

However, we note that such physical barriers are only effective to a certain point, as an attacker can always increase the laser power in an attempt to compensate for the cover-induced attenuation. Finally, in case such compensation is not possible, the attacker can always use the laser to burn through barriers, creating his own light path.

C. Limitations

Hardware Limitations. Being a light based attack, LightCommands inherits all the limitations of light-related physics. In particular, light does not properly penetrate opaque obstacles which might be penetrable to sound. In addition, unlike sound, LightCommands requires careful aiming and line of sight access. Finally, while line of sight access is often available for smart speakers visible through windows, the situation is different for mobile devices such as smart watches, phones and tablets. This is since unlike static smart speakers, these devices are often mobile, requiring an attacker to quick aim and inject commands. When combined with the precise aiming and higher laser power required to attack such devices, successful LightCommands attacks might be particularly challenging. We thus leave the task of systematically exploring such devices to future work.

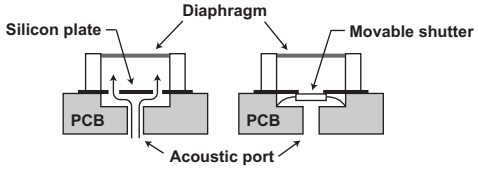

Fig. 13. Designs of MEMS microphone with light-blocking barriers [40]

Liveness Test. As opposed to other attacks, LightCommands' remote threat model and lack of proper feedback channel makes it difficult for the attacker to pass any sorts of liveness checks. Such checks can be as primitive as asking a user a simple questions before performing a sensitive command, or as sophisticated as using data from different microphones [41, 42, 43] or sound reflections [44] in order to verify that the incoming commands where indeed spoken by a live human (as opposed to played back via a speaker). We note that, however, using interactive liveness tests (e.g., questions) typically hurts usability while the works of [41, 42, 43, 44] can only authenticate users at close distances (e.g., tens of centimeters), making it inapplicable to smart speaker devices.

VIII. Conclusions and Future Work

In this paper we presented LightCommands, which is an attack that uses light to inject commands into voice-controllable systems from large distances. To mount the attack, an attacker transmits light modulated with an audio signal, which is converted back to the original audio signal within a microphone. We demonstrated LightCommands on many commercially available voice-controllable systems that use Siri, Portal, Google Assistant, and Alexa, obtaining successful command injections at a maximum distance of more than 100 meters while penetrating clear glass windows. Next, we highlight deficiencies in the security of voice-controllable systems, which leads to additional compromises of third-party hardware such as locks and cars.

Better understanding of the physics behind the attack will benefit both new attacks and countermeasures. In particular, we can possibly use the same principle to mount other acoustic injection attacks (e.g., on motion sensors) using light. In addition, heating by laser can also be an effective way of injecting false signals to sensors.

IX. Acknowledgments

The authors would like to thank John Nees for providing helpful advice regarding laser operation and laser optics.

This research was funded by JSPS KAKENHI Grant Number JP18K18047 and JP18KK0312, by the Defense Advanced Research Projects Agency (DARPA) under contract FA8750-19-C-0531, gifts from Intel, AMD, and Analog Devices, an award from MCity at University of Michigan, and by the National Science Foundation under grant CNS-1330142.

References

[1] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*. ACM, 2014, pp. 63–74.

[2] Y. Jang, C. Song, S. P. Chung, T. Wang, and W. Lee, "A11y attacks: Exploiting accessibility in operating systems," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 103–115.

[3] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands." in *USENIX Security Symposium*, 2016, pp. 513–530.

[4] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible voice commands," in *ACM Conference on Computer and Communications Security*. ACM, 2017, pp. 103–117.

[5] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2017, pp. 2–14.

[6] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, 2018, pp. 547–560.

[7] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "CommanderSong: A systematic approach for practical adversarial voice recognition," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 49–64.

[8] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," *Presented at WOOT*, vol. 15, pp. 10–11, 2015.

[9] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Advances in neural information processing systems*, 2017, pp. 6977–6987.

[10] L. Song and P. Mittal, "Inaudible voice commands," *arXiv preprint arXiv:1708.07238*, 2017.

[11] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, "Skill squatting attacks on Amazon Alexa," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 33–47.

[12] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, "Understanding and mitigating the security risks of voice-controlled third-party skills on amazon alexa and google home," *arXiv preprint arXiv:1805.01525*, 2018.

[13] Y. Son, H. Shin, D. Kim, Y.-S. Park, J. Noh, K. Choi, J. Choi, Y. Kim *et al.*, "Rocking drones with intentional sound noise on gyroscopic sensors." in *USENIX Security Symposium*, 2015, pp. 881–896.

[14] C. Yan, W. Xu, and J. Liu, "Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle," *DEFCON*, vol. 24, 2016.

[15] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "WALNUT: waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks," in *EuroS&P*. IEEE, 2017, pp. 3–18.

[16] S. Nashimoto, D. Suzuki, T. Sugawara, and K. Sakiyama, "Sensor CON-Fusion: Defeating kalman filter in signal injection attack," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, AsiaCCS 2018*, 2018, pp. 511–524.

[17] C. Bolton, S. Rampazzi, C. Li, A. Kwong, W. Xu, and K. Fu, "Blue note: How intentional acoustic interference damages availability and integrity in hard disk drives and operating systems," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2018, pp. 1048–1062.

[18] J. Petit, B. Stottelaar, M. Feiri, and F. Kargl, "Remote attacks on automated vehicles sensors: Experiments on camera and LiDAR," *Black Hat Europe*, vol. 11, p. 2015, 2015.

[19] J. Petit and S. E. Shladover, "Potential cyberattacks on automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 546–556, 2015.

[20] H. Shin, D. Kim, Y. Kwon, and Y. Kim, "Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications," in *Cryptographic Hardware and Embedded Systems - CHES 2017 - 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*, 2017, pp. 445–467.

[21] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on LiDAR-based perception in autonomous driving," 2019.

[22] Y.-S. Park, Y. Son, H. Shin, D. Kim, and Y. Kim, "This ain't your dose: Sensor spoofing attack on medical infusion pump." in *WOOT*, 2016.

[23] A. S. Uluagac, V. Subramanian, and R. Beyah, "Sensory channel threats to cyber physical systems: A wake-up call," in *2014 IEEE Conference on Communications and Network Security*. IEEE, 2014, pp. 301–309.

[24] D. H. Habing, "The use of lasers to simulate radiation-induced transients in semiconductor devices and circuits," *IEEE Transactions on Nuclear Science*, vol. 12, no. 5, pp. 91–100, 1965.

[25] S. P. Skorobogatov and R. J. Anderson, "Optical fault induction attacks," in *Cryptographic Hardware and Embedded Systems - CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002, Revised Papers*, 2002, pp. 2–12.

[26] D. Karaklaji, J. Schmidt, and I. Verbauwhede, "Hardware designer's guide to fault attacks," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 12, pp. 2295–2306, 2013.

[27] J.-M. Dutertre, J. J. Fournier, A.-P. Mirbaha, D. Naccache, J.-B. Rigaud, B. Robisson, and A. Tria, "Review of fault injection mechanisms and consequences on countermeasures design," in *2011 6th International Conference on Design & Technology of Integrated Systems in Nanoscale Era (DTIS)*. IEEE, 2011, pp. 1–6.

[28] S. Manohar and D. Razansky, "Photoacoustics: a historical review," *Advances in Optics and Photonics*, vol. 8, no. 4, pp. 586–617, December 2016.

[29] A. G. Bell, "Upon the production and reproduction of sound by light," *Journal of the Society of Telegraph Engineers*, vol. 9, no. 34, pp. 404–426, 1880.

[30] P. Tucker, "The US military is making lasers that create voices out of thin air," accessed: 2019-08-20.

[31] R. M. Sullenberger, S. Kaushik, and C. M. Wynn, "Photoacoustic communications: delivering audible signals via absorption of light by atmospheric H2O," *Opt. Lett.*, vol. 44, no. 3, pp. 622–625, 2019.

[32] I. S. of Conformity Assessment Schemes for Electrotechnical Equipment and Components, "IEC 60825-1:2014 safety of laser products - part 1: Equipment classification and requirements." [Online]. Available: https://www.iecee.org/index.htm

[33] U. D. of Health, F. Human Services, C. f. D. Drug Administration, and R. Health, "Laser products conformance with IEC 60825-1 ed. 3 and IEC 60601-2-22 ed. 3.1 (laser notice no. 56) guidance for industry and food and drug administration staff." [Online]. Available: https://www.fda.gov/media/110120/download

[34] S. M. Goldwasser and B. Edwards, "Hidden menace: Recongnizing and controlling the hazards posed by smaller and lower power lasers," http://www.repairfaq.org/sam/laser/ILSC_2011-1303.pdf, 2011, accessed: 2019-08-20.

[35] IFIXIT, "Google home mini teardown," https://www.ifixit.com/Teardown/Google+Home+Mini+Teardown/102264, accessed: 2019-08-25.

[36] N. Ltd., "Naturalreader," https://www.naturalreaders.com/online/, accessed: 2019-08-25.

[37] N. Melena, N. Neuenfeldt, A. Slagel, M. Hamel, C. Mackin, and C. Smith, "Covert IR-laser remote listening device," The University of Arizona Honors Thesis https://repository.arizona.edu/handle/10150/244475, accessed: 2019-08-20.

[38] D. Davidson, H. Wu, R. Jellinek, T. Ristenpart, and V. Singh, "Controlling UAVs with sensor input spoofing attacks," in *Proceedings of the 10th USENIX Conference on Offensive Technologies*, ser. WOOT'16. Berkeley, CA, USA: USENIX Association, 2016, pp. 221–231. [Online]. Available: http://dl.acm.org/citation.cfm?id=3027019.3027039

[39] A. K. Sikder, H. Aksu, and A. S. Uluagac, "6thsense: A context-aware sensor-based attack detector for smart devices," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 397–414.

[40] Z. Wang, Q. Zou, Q. Song, and J. Tao, "The era of silicon MEMS microphone and look beyond," in *2015 Transducers - 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, June 2015, pp. 375–378.

[41] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive:

A phoneme localization based liveness detection for voice authentication on smartphones," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1080–1091.

[42] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. ACM, 2017, pp. 57–71.

[43] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1466–1474.

[44] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 447–460, 2019.