This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

Audio Hotspot Attack: An Attack on Voice Assistance Systems Using Directional Sound Beams and its Feasibility

Ryo Iijima*†, Shota Minami*, Yunao Zhou*, Tatsuya Takehisa†, Takeshi Takahashi†, Yasuhiro Oikawa*, Tatsuya Mori*†‡
*Waseda University †National Institute of Information and Communications Technology ‡RIKEN AIP

Abstract—We propose a novel attack, called an "Audio Hotspot Attack," which performs an inaudible malicious voice command attack, by targeting voice assistance systems, e.g., smart speakers or in-car navigation systems. The key idea of the approach is to leverage directional sound beams generated from parametric loudspeakers, which emit amplitude-modulated ultrasounds that will be self-demodulated in the air. Our work goes beyond the previous studies of inaudible voice command attack in the following three aspects: (1) the attack can succeed on a long distance (3.5 meters in a small room, and 12 meters in a long hallway), (2) it can control the spot of the audible area by using two directional sound beams, which consist of a carrier wave and a sideband wave, and (3) the proposed attack leverages a physical phenomenon i.e.,non-linearity in the air, to attack voice assistance systems. To evaluate the feasibility of the attack, we performed extensive in-lab experiments and a user study involving 20 participants. The results demonstrated that the attack was feasible in a real-world setting. We discussed the extent of the threat, as well as the possible countermeasures against the attack.

Index Terms—Voice assistance systems, Voice commands attack, Ultrasonic, Security, Acoustics, Internet of Things

1 Introduction

Voice assistance systems, such as Siri [2], Google Assistant [3], and Amazon Alexa [4] have become increasingly popular as a means to establish user-friendly human–computer interactions. Voice assistance systems are now supported on various devices, e.g., smartphones/tablets, smart speakers, automobiles, smart homes, smart watches, smart TVs, media boxes, and laptops/desktops. Voice assistance systems can integrate speech recognition to demonstrate various skills such as providing recommendations to restaurants, reading out schedules, and even purchasing products when an appropriate voice command is given.

While these voice assistance systems have clear benefits in daily life activities, they also raise intrinsic security and privacy concerns. One of the most serious security issues related to the use of voice assistance systems is the lack of a rigorous mechanism to guarantee the trustworthiness of the voice source that operates the system. As previous studies have demonstrated [5], [6], voice assistance systems are vulnerable to "inaudible voice command attacks." Here, an attacker can issue voice commands to a voice assistance device unbeknownst to the device owner. For instance, if an attacker generates an inaudible voice command that adjusts the volume of the music player set in a car to its maximum, the driver may be surprised or momentarily distracted, thus increasing the likelihood of an accident. Recent studies have leveraged existing vulnerabilities of the device or software. In Ref. [6], the authors found that ultrasound can be used to convey inaudible voice command attacks, by using the vulnerability of the amplifier. Hidden voice commands [5] used the vulnerability of machine learning models that incorrectly recognize noise as normal commands.

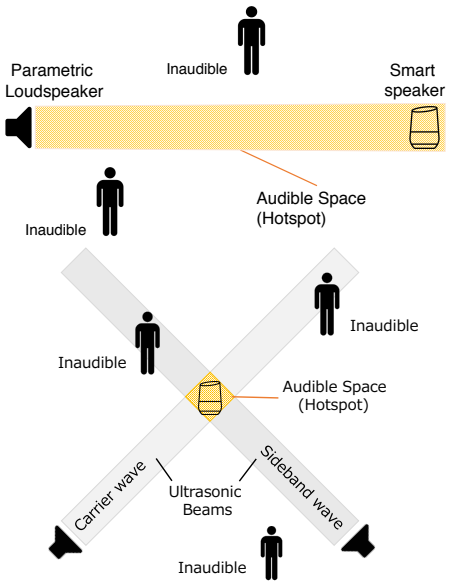


Fig. 1. Overview of the Audio Hotspot Attack. Top: Attack with one parametric loudspeaker (linear attack). Bottom: Attack with two parametric loudspeakers (cross attack). In the yellow colored area, you can hear the sound.

We propose a novel inaudible voice attack, named Audio Hotspot Attack, which leverages the *physical phenomena*. In

A preliminary version of this work appeared as a poster presentation at ACM CCS 2018 [1]

this attack, attackers attempt to input directional sound to voice assistance systems as shown in Figure 1. Directional sound is generated by using the nonlinearity of ultrasonic waves in the air. When the modulated ultrasound passes through the air, which acts as a nonlinear medium, the signal is demodulated into audible sound even if a demodulation circuit is *not* prepared. It is well known that the demodulated sound signals exhibit higher directivity than those emitted from a normal loudspeaker [7], [8]. To generate directional sound, we make use of a parametric loudspeaker, which composes of an array of ultrasound transducers.

The attack proposed in this paper is different from previously proposed attacks in that it leverages physical phenomena that cannot be modified or eliminated. As the previous attacks use vulnerabilities associated with hardware or software, they can be fixed, e.g., by modifying the machine learning algorithm or eliminating the nonlinearity of the microphone. In contrast, the nonlinearity of air is a natural phenomenon, and it is impossible to take measures against it using conventional approaches.

Furthermore, the adoption of parametric loudspeakers enables an attacker to perform a unique form of the attack, called a *cross attack*. As shown at the bottom of Figure 1, an attacker sets two parametric loudspeakers in different places and transmits directional sound beams to the target voice assistance device. The two sound beams are inaudible because each sound beam consists of a carrier wave or sideband wave with ultrasound frequency. The sound beams become audible where the two beams cross at a point; i.e., they become an AM sound wave. An attacker can take control of the cross point by adjusting the sources of the two sound beams.

To evaluate the feasibility of the attack, we pose the following research questions:

RQ1: *Is the Audio Hotspot Attack feasible at long distance with off-the-shelf voice assistant devices?*
RQ2: *Does the Audio Hotspot Attack succeed in noisy practical environments?*
RQ3: *Is the attack stealthy for nearby people and unrecognizable for them?*

We aim to answer these questions through extensive experiments and user studies involving 20 participants.

The contributions of this work can be summarized as follows:

- We proposed a novel inaudible voice command attack that targets voice assistance systems, leveraging the directional sound beams to create a "hotspot" of the attack success area (Section 3).
- We carefully designed and controlled our experiments. We used a room and equipment dedicated to acoustic experiments (Section 4).
- We demonstrated that the attack could succeed at a long distance. We discovered that the attacks were tolerant of environmental noise. For both devices, the attack success rate remained high at a noise sound pressure level. We showed that the cross attack was also feasible (Section 5).
- Through the extensive user studies, we demonstrated that people could not recognize the attacker's voice (Section 6).
- We discussed potential threats that may arise in the future as well as the possible countermeasures against the attack (Section 7).

To the best of our knowledge, this work is the first to make use of directional sound beams as a means of attacking voice-controlled systems. This perspective sheds new light on security and privacy issues for systems that make use of sound.

2 BACKGROUND

In this section, we describe the three key technologies that constitute our attack: the voice assistance system, parametric loudspeakers, and voice presentation attack.

2.1 Voice Assistance Systems

Currently, a typical voice assistance system has two action phases for device operation: activation and recognition. In the first phase, a user speaks a specific wake-up word to activate the system, e.g., "OK Google" for Google Assistant, "Alexa" for Amazon Alexa, and "Hey Siri" for Apple Siri. In the second phase, a user transmits a voice command to the system. The system applies speech recognition to the received voice data and executes a command extracted from this data. The available voice commands include common operations such as turning on a light, answering questions, reading the news, or privacy-sensitive operations that access personal resources such as reading out schedules, sending a text message, making a phone call, or purchasing a product.

Many of the smart speakers today offer speaker recognition functionality so that each person in the household can enjoy the device in a customizable way. For instance, each person using the Amazon Echo can link their own Amazon account to the device. The device identifies each person by leveraging voiceprints to employ biometric verification. To be enrolled in the device's speaker recognition, an owner of the device first needs to register his or her voiceprint, typically by saying a wake-up word multiple times. By comparing the wake-up word against a previously created voiceprint, the voice assistance system verifies a person's identity. Although a third person who is not registered can still attempt to use the device, his or her usage will be limited to non-personalized common services such as reading news or weather forecasts.

As we will discuss in Section 3, speaker recognition technology is vulnerable to voice presentation attacks [9]. These attacks attempt to bypass voice authentication using voice replay/synthesis/conversion technique fraudulently (See Section 2.3).

2.2 Mechanism of parametric loudspeakers

A parametric loudspeaker can generate directional sound using ultrasound. It consists of an array of many ultrasound transducers installed in parallel [10]. Figure 2 presents a parametric loudspeaker used throughout the experiments.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

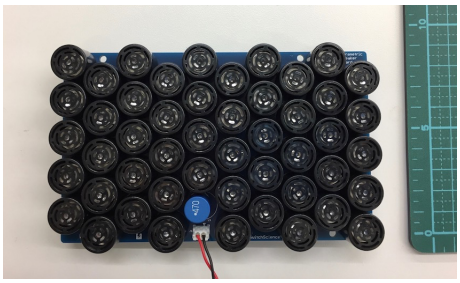


Fig. 2. A parametric loudspeaker. This loudspeaker can generate directional sound. It consists of an array of ultrasonic-emitting loudspeakers arranged in a grid. A parametric loudspeaker emits sounds on a narrow spatial range containing a targeted device.

Each ultrasonic transducer transmits ultrasound that modulates the original sound wave with amplitude modulation (AM). The generated ultrasound is self-demodulated in the air and becomes audible even if we do not prepare a demodulation circuit (called self-demodulation [7]). Next, we present the self-demodulation mechanism, also known as the *parametric phenomenon*.

Let $p = p(x,t)$ be the sound pressure caused by sound wave originating from a parametric loudspeaker, where x is the distance from the loudspeaker and t is time. As the sound wave is AM-modulated, it has three major frequencies, i.e., carrier frequency, f_c, and adjacent sideband, $f_{s\_}$, f_{s+} where $f_{s\_} = f_c - f_m$, $f_{s+} = f_c + f_m$. f_m represents the frequency of the sound wave to be injected by an attacker. We focus on lower sideband to simplify. Primary wave p is expressed as

$$p = p_c \sin\left(2\pi f_c t'\right) + p_{s\_} \sin\left(2\pi f_{s\_} t'\right) \tag{1}$$

p_c and $p_{s\_}$ are the amplitudes of the carrier wave and the sideband wave, respectively. where $t' = t - x/c_0$ is a *retarded time*; the retarded time is the time when the sound wave began to propagate from the sound source.

Burger's equation is one of the fluid models that represents the nonlinear dynamics of sound waves [11]. The dynamics of ultrasound generated from an array of transducers can be modeled with Burger's equation:

$$\frac{\partial p}{\partial x} = \frac{\beta}{\rho_0 c_0^3}\frac{\partial}{\partial t'}p^2 + \frac{\delta}{2c_0^3}\frac{\partial^2 p}{\partial t'^2}, \tag{2}$$

where β is the coefficient of nonlinearity, ρ_0 is the density of air, and c_0 is the sound speed. The first term on the right side has nonlinearity. By substituting Eq. 1 into p, we have

$$\begin{aligned}\frac{\partial}{\partial t'}p^2 \quad = \quad & \frac{\partial}{\partial t'}[p_c^2 \sin^2(2\pi f_c t') + p_{s\_}^2 \sin^2(2\pi f_{s\_} t') \\ & +2p_c p_{s\_} \sin(2\pi f_c t')\sin(2\pi f_{s\_} t')],\end{aligned} \tag{3}$$

For simplicity, we calculate only the third term of Eq. 3, from which, we can derive f_m. [1]

$$\begin{aligned}&\frac{\partial}{\partial t'}(2p_c p_{s\_} \sin(2\pi f_c t')\sin(2\pi f_{s\_} t')) \\ &= 2[2\pi f_{s\_} p_c p_{s\_} \sin(2\pi f_c t')\cos(2\pi f_{s\_} t') \\ &\quad +2\pi f_c p_c p_{s\_} cos(2\pi f_c t')\sin(2\pi f_{s\_} t')], \\ &= -2\pi p_c p_{s\_}[(f_c + f_{s\_})\sin(2\pi(f_c + f_{s\_})t') \\ &\quad +(f_c - f_{s\_})\sin(2\pi(f_c - f_{s\_})t'), \\ &= -2\pi p_c p_{s\_}[(f_c + f_{s\_})\sin(2\pi(f_c + f_{s\_})t') \\ &\quad +f_m \sin(2\pi f_m t')],\end{aligned} \tag{4}$$

Eq. 4 contains two terms. The first term, which contains $\sin(2\pi(f_c+f_{s\_})t')$, will be removed by low–pass filter. Thus, remaining term is a sine function with the frequency of the original modulation wave, f_m. By substituting Eq. 4 into Eq. 2, we derive that $\partial p/\partial x$ contain the following term,

$$\frac{2\beta\pi p_c p_{s\_} f_m}{\rho_0 {c_0}^3}\sin(2\pi f_m t') \tag{5}$$

By integrating the term with respect to x, we derive that p contains the following term

$$\frac{2\beta\pi p_c p_{s\_} f_m}{\rho_0 {c_0}^3} x \sin(2\pi f_m t') \tag{6}$$

which indicates that the observed sound pressure includes the component of the original modulation wave. This is how the nonlinearity of the air demodulates the modulated sound wave.

Figure 3 presents an overview of the parametric phenomenon. After emitted from a parametric loudspeaker, the sound pressure of the audible sound wave, f_m, gradually increases. Although both the audible sound wave and inaudible ultrasound wave are to be attenuated over time, inaudible ultrasound waves attenuate faster due to the fact that in the air, high frequency sound wave attenuates faster compared to low frequency sound waves. The parametric phenomenon is observed only along the direction in which the ultrasound was emitted because the waves have the same phase along the path.

Finally, we show the intuitive explanation of the formation of directional sound beam. The demodulated sound traveling in the forward direction is amplified because the phase is aligned. On the other hand, sound traveling in a direction other than the forward direction is not amplified because the phase is not aligned. The mathematical description of the theory can be found in Refs [7], [8].

2.3 Voice Presentation Attack

In the ISO/IEC standard, presentation attacks are defined as "presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system. [12]" There have been several approaches for evading speaker recognition or, more broadly, voice authentication. These attacks are known as voice presentation attacks [9]. Well-known voice presentation attacks include

1. If we compute the partial differentiation of the first and second terms in a way like Eq 4, sine functions with the frequencies of $2f_{s\_}$, $2f_c$, and so on, appear. Because these frequencies are not associated with f_m and will be removed by the low–pass filter on the microphone, all these sine functions can be omitted in the remaining calculation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

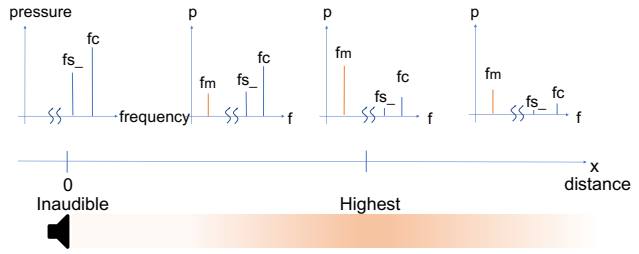


Fig. 3. Illustration of the demodulation in the air. f_c is a carrier frequency and $f_{s\_}$ is a sideband frequency, where $f_{s\_} = f_c - f_m$ and f_m represents a frequency of the sound wave to be injected by an attacker. In a short distance, the sound pressure of the demodulated sound, f_m will increase in proportion to the distance, x, following Eq. 6. However, due to the attenuation of the ultrasonic wave, the sound pressure of the demodulated sound will decrease over a long distance.

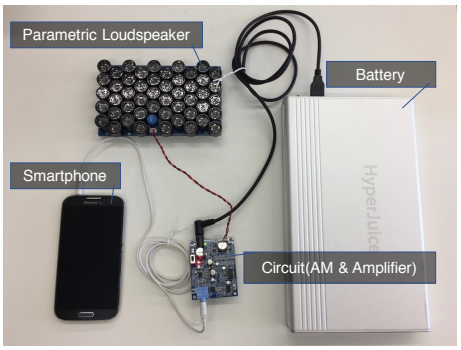


Fig. 4. An example of device setup. We use a battery to allow attackers to use this device anywhere. The circuit contains amplifier and amplified modulator. The details of the circuit are presented in Fig. 5.

the replay attack [13], [14], speech synthesis attack [13], and voice conversion attack [15].

During a replay attack, an attacker pre-records the speech of the victim in advance. The attacker then replays the recorded speech to the target device. Distinguishing between genuine and replayed speech from the time-domain and spectrum-domain representations of speech data is difficult task [16]. The drawback of a replay attack is that an attacker needs to pre-record speech, including voice commands for both activation and recognition. *Speech synthesis* and *Voice conversion* are techniques that alleviate this limitation. Speech synthesis (Text-to-speech, TTS) is a technique to generate natural speech sound from the text. Wavenet [17] is one example that creates synthesized voices by using deep learning models. Voice conversion aims to convert an attacker's voice to a victim's voice in real time. We do not need to prepare text, unlike in TTS. These attacks offer an effective way to generate synthetic speech in a manner such that the generated output is perceived as a sentence uttered by a target. In [15], the author demonstrated that an attacker can successfully execute a voice impersonation attack by using an off-the-shelf voice-conversion tool, even against state-of-the-art voice verification systems. They reveal that the attacker can convert his/her voice if they collect just a few minutes' worth of audio.

While these attack techniques aim at impersonating the victim's voice, our goal focuses on the different attack vector, i.e., secretly delivering the voice signal to the target voice assistant device. As our attack is agnostic to the voice content, voice presentation attack techniques can be directly mounted on our attack.

3 Threat Model and Assumptions

In this section, we describe the Audio Hotspot Attack threat model by making several assumptions to evaluate the threat.

Target of the attack

The goal of an attacker is to manipulate the target voice assistance device without being noticed by people. Although the attack is applicable to various voice assistance systems in principle, a smart speaker is used herein as an example of the target device. Because smart speakers can control smart home devices, the attack vector ranges are widespread. We evaluated the attack using two smart speakers, Amazon Echo and Google Home. For these devices, an attacker must activate the device with a wake-up word, and then transmit a voice command. In this study, we assume that the target device is not moving (i.e., it is set on a fixed place, for example, on the table). This assumption is natural in the case of smart speakers.

Attacker's equipment

As shown in Figure 1, the *Audio Hotspot Attack* has the two attack modes: *linear attack* and *cross attack*. An attacker needs to setup a parametric loudspeaker for the linear attack, and two parametric loudspeakers for the cross attack. The parametric loudspeaker that performs the attack is small and portable. The attacker also needs to carry a smartphone in order to generate malicious voice commands from the parametric loudspeakers. Figure 4 shows an example of a device setup used by an attacker to execute an attack.

Speaker recognition

As mentioned in Section 2, modern devices equipped with voice assistance systems such as smartphones or smart speakers have increasingly adopted the speaker recognition functionality. If the owner of a device has turned on this functionality, an attacker may not be able to succeed in the attack even when he/she has successfully transmitted an inaudible voice command to the target device.

Here, the attacker collects voice samples by being in close physical proximity to the target, by making a phone call, or by searching for clips online. For the purposes of this work, we assumed that an attacker was able to bypass the speaker recognition by leveraging voice presentation attacks, which are discussed in Section 2.3. As shown in Section 7.2.3, there are some methods that detect presentation attacks (PAD method). We assume that the voice assistance systems do not have a PAD method. We confirm that presentation attacks are successful on practical devices, i.e., Google Home and Amazon Echo, before the experiments.

4 Experimental Setup

In this section, we describe the design of our experiments, including details pertaining to the devices, equipment, and software used, together with their settings.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

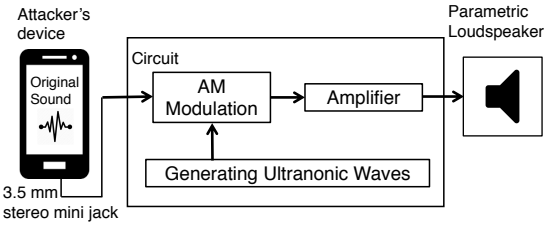


Fig. 5. Circuit diagram. The circuit first applies AM to the input soundwave, using the generated ultrasonic wave as a carrier wave. Next, the sound pressure of the AM wave will be amplified. The amplified soundwave will be the output for the parametric loudspeaker.

TABLE 1
A list of equipment used for the experiments.

Equipment	manufacturer / model number
Parametric loudspeaker	Switch Science / SSCI-018425 [19]
Amplifier	Accuphase / Power Amplifier PRO-15 [20]
Parametric loudspeaker	Ultrasonic audio technologies / Directional Speaker Acouspade [21]
Dynamic loudspeaker	YAMAHA / MONITOR SPEAKER MS101 III [22]
Sound level meter	RION / NL-32 [23]
Ultrasonic microphone	B&K / 4939-A-011 [24]
Audio Interface	MOTU / UltraLite mk4 [25]

4.1 Materials

4.1.1 Experiment room

Sound wave dynamics depend on the material makeup of the room. As these attacks were performed using sound waves, the choice of the experiment room was key. Otherwise, the obtained results will be valid only for a specific environment. To overcome this concern, we used a room designed for acoustic experimentation. To eliminate the effects of the material makeup of the room, all wall and ceiling surfaces were made of sound-absorbing material (Appendix B, Figure 2).

The average sound pressure level (SPL) of the room was around 12 dB(A). Here, dB(A) denotes A-weighted SPL, which is applied to instrument-measured sound levels. A-weighting is used because the human ear is less sensitive to lower audible frequencies.

4.1.2 Target devices

Following the assumption that the target device is stationary, Google Home and Amazon Alexa are the primary target devices used for the analysis. These devices were chosen because they accounted for more than 95% of the smart speaker market share in 2018 [18].

4.1.3 Equipment used for the experiments

Table 1 shows a list of equipment used for the experiments. While there are several commercial parametric loudspeaker products, we intended to take a white-box approach. That is, as the details of the board and elements are publicly available on the manufacturers' websites, we can obtain the technical specifications of the speaker, such as frequency response. To this end, the Switch Science Super directional speaker [19] was adopted as a primary parametric loudspeaker. The kit comprises two printed circuit boards (PCBs). One PCB has an AM circuit, an amplifier circuit, an audio input (3.5 mm stereo mini jack), and a power input (DC 12V/1A). Figure 5 presents a diagram of the circuits. Another PCB implements 49 ultrasonic ceramic transducers connected in parallel. The first PCB applies the AM to the input sound wave and then amplifies the signal level. The amplified signal is transmitted to the second PCB, i.e., ultrasound transducers. Another parametric loudspeaker—directional speaker ACOUSPADE—is also used, to study the maximum distance at which the attack can succeed. The sound level meter is capable of measuring the SPL of 28–138 dB(A) for a frequency range of 20 Hz to 20 kHz. The meter was used to measure the SPL of several areas in the experiment room under various conditions. The ultrasonic microphone was also used for measuring the ultrasonic components in the measured sound waves.

4.2 Voice generation

To generate a malicious voice speech command, we used Amazon Polly [26], a cloud service that turns text into natural sounding speech. As the basis for the analysis, the voice named "Ivy" was used, which is a female, US English accent. The voice parameters (e.g., speaking rate or fundamental frequency) were set to default values. All voice assistance systems that were tested to check whether they accept synthesized voice commands. As speech synthesis services can change in the future, we plan to make our data available to any researchers who wish to replicate or extend our work.

5 Evaluation of the Attack

We evaluated attack feasibility using the following aspects: maximum successful attack distance, noise tolerance of the attack, and the impact of voice commands. For simplicity, and to evaluate the impact of these factors, we applied a linear attack. For the cross attack, we evaluated attack feasibility using the parameters obtained through the linear attack experiments. The attack success depends on the type of voice command (i.e., activation or recognition). Therefore, for each attack mode, we applied both types of voice commands. In general, activation commands ("wake-up words") are more likely to succeed.

5.1 Distance versus Attack success rate

The aim of this study was to clarify how the distance between the target device and adversary's parametric loudspeaker affected the success rate of the Audio Hotspot Attack. Throughout the experiments, the SPL of the output power from the parametric loudspeaker was fixed. In particular, the audible sound of the parametric loudspeakers was adjusted to 60 dB(A), and the SPL of the ultrasound was 100 dB at a point 3 m away from the parametric loudspeaker. Figure 6 presents the experimental setup. The distance measured was between the parametric loudspeaker and the microphone of the voice assistance systems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

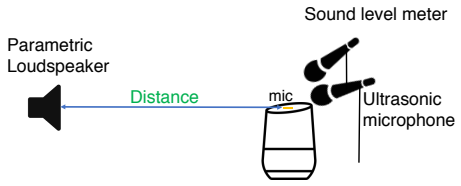


Fig. 6. Experimental setup of distance measurement experiments. The distance measured was between the parametric loudspeaker and the microphone of the voice assistance systems.

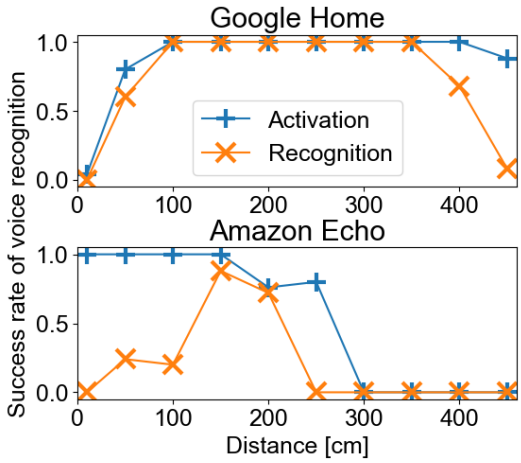


Fig. 7. Distance versus attack success rate. Noise SPL is set to 60 dB(A). For Google Home, the longest distance was 3.5 m. Activation voice commands were more likely to be accepted compared to recognition voice commands.

To measure the distance, we used the experiment room (described in section 4.1.1). We extended the study to three different locations, including a hallway, seminar-room, and outdoors.

5.1.1 Measurement within the experiment room

The distance between the target device and the parametric loudspeaker was altered from 0.1 m to 5 m in increments of 50 cm (i.e., 0.1, 0.5, 1.0,..., 5.0 m). By adjusting the output power of the dynamic speakers, we were able to adjust the SPL of the noise measured in the room to 60 dB(A) with error bounds within 1 dB(A). Notably, a SPL of 60 dB(A) corresponds to an environment where a person's speech is heard at a distance of 1 m. Thus, the noise level was fairly high. This setting was purposively chosen to conservatively evaluate attack success rate (i.e., a higher attack success rate could be expected in quieter settings). We note that the $1/f$ noises better suited to emulate a realistic environment than the white noise because it is natural that signals with the lower or higher frequencies have more or less power respectively.

For a given distance, a pair of activation and recognition voice commands were generated. This process was repeated 25 times. For each voice command, we noted if the command was accepted by the voice assistance system by observing the response of the device. For the activation commands, "*Ok Google*" for Google Home and "*Alexa*" for Amazon Echo were used. For the recognition voice commands, "*What's on my next schedule*?" for Google Home and "*What's on my schedule*?" for Amazon Echo[2] were used.

The attack success rates were calculated, and the results are shown in Figure 7. For a certain range of distances, the attack was highly successful for both devices. This was particularly true for Google Home, the longest distance was 3.5 m. Activation voice commands were more likely to be accepted than recognition voice commands. This makes sense given the fact that the recognition voice commands are much more variable than activation voice commands. In the short distance, the success rate becomes low because the acoustic sound was too loud to be properly processed by the voice assistance systems. Finally, Google Home featured a higher attack success rate than Amazon Echo. As these commercial products are black box in nature, their behaviors can be difficult to interpret. It is possible that circuits and software used for Amazon Echo are somehow resistant to the Audio Hotspot Attack; therefore, they will be investigated in future studies.

5.1.2 Extended measurement in practical environments.

Next, we studied the distances of successful attacks using different locations: a hallway, a seminar room, and outside. The hallway and the room have much higher reverberation compared to the room dedicated for acoustic experiments. We used a commercial parametric loudspeaker product [21], as listed in Table 1. The parametric loudspeaker can emit full frequency-range speech with the audible SPL of 62–63 dB(A) at a distance of 3 m. For reference, the location photos are shown in Appendix B. Note that for these locations, we did not add synthesized noise sounds. The average SPL measured in the hallway was 39.3 dB(A), the seminar room was 55.2 dB(A), and the average outside SPL was 52.5 dB(A). The conditions outside were as follows: the weather on the day was fine, with temperature was 23.2 °C (73.8 °F), a humidity of 36%, and a wind speed of 6 m/s southward. Note that, we do not use synthesized noise in this measurement, to evaluate the effect of noise on realistic environments. The purpose of the experiment was to determine the longest distance at which the attack is still effective, with the effectiveness being determined using the following criteria: if three consecutive voice commands are all accepted for a given distance, the attack is regarded as effective for the distance. For each location, the starting distance was 1 m and the tests were repeated until there was an attack failure. Tables 2 summarize our results.

The hallway experiment demonstrated that the attack was effective at a distance of 10+ m. The seminar room and outside experiment demonstrated that the attack was effective to a distance of 4+ m. The difference in the attack success distances reflects the respective noise levels within each location. These results indicated that the Audio Hotspot Attack was feasible in three real-world scenarios. We can succeed in the attack in two environments with reverberation, i.e., the hallway and inside the room. We also showed that the experiment was successful outside the room. In addition, the attack success distances achieved were much longer than the state-of-the-art inaudible voice

2. At the time of the experiment, Alexa did not support the 'next' voice command for the calendar.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

TABLE 2
The longest distance the attack was effective at a hallway, a seminar room, and outside. In the hallway experiment, the attack was effective at a distance of 10+ m. In the case of the hallway and room, the longest distance is 4+ m. We show the picture of each place in Appendix.

	Hallway [m]		Room [m]		Outside [m]	
Devices	Acti.	Recog.	Acti.	Recog.	Acti.	Recog.
Google Home	15.0	11.7	4.2	4.0	4.2	4.2
Amazon Echo	19.9	12.1	4.8	4.0	5.8	4.2

command attack that uses ultrasound [6], which indicated that the maximum distance for Amazon Echo averaged 1.65 m with a background noise of 55 dB SPL.

5.2 Noise tolerance

We studied how the noise affects the attack success rate. For this study, we used the experiment room, as described in Section 4.1.1. Because we were examining the effects of noise, the sound generated by the parametric loudspeaker was fixed at 60 dB(A) and the distance between the parametric loudspeaker and the target device was 1.5 m.

5.2.1 Stationary noise

Using the dynamic speaker, we generated $1/f$ noise with an SPL ranging from 45 dB(A) to 78 dB(A) (the maximum SPL for the dynamic speaker). the common environmental noise levels are shown in [27]. To calculate the signal-to-noise ratio (SNR), we use the following formula Eq. 7 [28]

$$\text{SNR [dB]} = \text{SPL of sound [dB]} - \text{SPL of noise [dB]} \quad (7)$$

We use the sound level meter to measure the SPL of voice command and noise. Figure 8 shows the results. For both devices, the attack was most successful when the noise SNR was over than 0 [dB], i.e., when the input command and noise have the same volumes. Activation voice commands were more tolerant of noise. This observation agrees with those previously-described in Section 5.1.

5.2.2 Nonstationary noise

We evaluate noise tolerance in an environment that has nonstationary noise. As nonstationary noise, we adopt babble noise. We used the room dedicated for acoustic experiments. We chose three types of noise settings: Default, Speech Blocker, and Chic dinner, which are taken from Ref. [29]. These noise types contain conversations in English. We summarize the results in Figure 9. We attempt to input the voice command 10 times in each setup. For both devices, the attack was successful when SNR was -5 [dB] and over. When the SNR is 0 [dB], i.e., when the volumes of input command and noise are same, attacks sometimes failed. In other cases, these results follow the observation of Fig 8.

5.3 Impact of voice commands

To study the impact of voice commands, various commands are inputted into the target devices. In this experiment, the distance between the parametric loudspeaker and the target devices was fixed at 1.5 m. Again, the output audible SPL

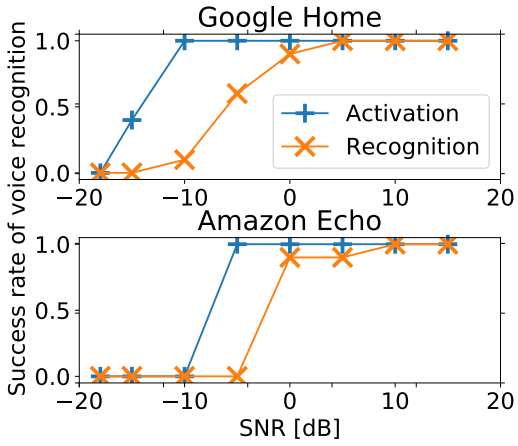


Fig. 8. Stationary noise versus attack success rate. The audible sound from the parametric loudspeaker was fixed to 60 dB(A). The attack was most successful when the SNR was larger than 0 [dB].

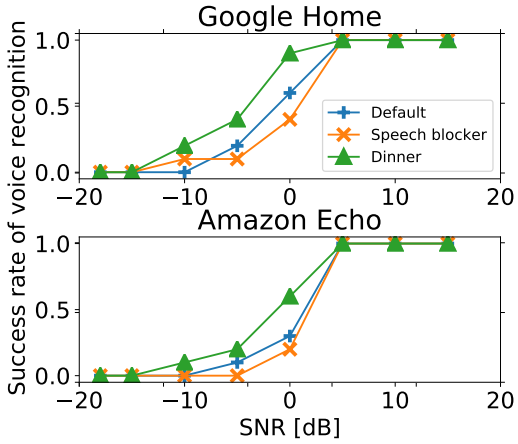


Fig. 9. Non stationary noise versus attack success rate. We used the recognition command for each device. These results follow the observation of Fig 8.

of the parametric loudspeaker was set to 60 dB(A). Each command was tested 10 times.

Table 3 shows the results. As indicates by the results, the attack success rate was high for commands of short lengths. We note that although the lengths of these commands were short, they can be used for malicious purposes; for example, by starting with the recognition command "Set volume 0," an attacker can improve the probability of success for the next attacks as a voice response from the device will not be heard by a nearby person. The attacker can also turn IoT devices on/off. If this device is a piece of heating equipment, considerable physical damage is possible. In contrast, for longer commands, the attack success rate was low.

We conjecture that there are several reasons behind this observation, e.g., the occurrences of infrequent words or the accumulation of recognition errors. These results agree with [6], who showed that longer commands, emitted as ultrasounds, were prone to failure.

TABLE 3
Attack success rates for various voice commands. The attack success rate was high for commands of short length (2–5 words.) The commands "turn on / off [device name]" are used for many smart home devices. The commands "turn in to 0" or "Set volume 0" change the volume minimum, which can make the output of device stealthy.

Device	Voice commands	Success rate
Google	OK Google	10/10
	Max volume	10/10
	Turn in to 0	10/10
	What's on my next schedule	10/10
	Turn on the light	10/10
	Turn off the light	10/10
	Play some music	10/10
	Tell everyone my password is abc	5/10
	Broadcast my credit card number is 1234567890	3/10
Amazon	Alexa	10/10
	Pair devices	10/10
	play some music	10/10
	What's on my next schedule	9/10
	Set volume 0	9/10
	Turn on the light	9/10
	Turn off the light	10/10
	Tell everyone my password is abc	2/10
	Broadcast my credit card number is 1234567890	1/10

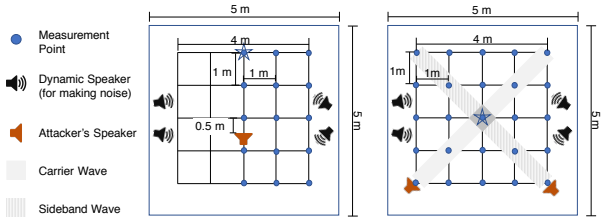


Fig. 10. Overview of the experimental setup. Left: user study of the linear attack in the acoustic room. Right: user study of the cross attack in the acoustic room. We use four dynamic speakers to adjust the noise level.

5.4 Evaluation of the cross attack

To perform the cross attack, the AM sound wave was separated into the carrier wave and the lower sideband wave using MATLAB [30]. The two sound waves were amplified and emitted through the two parametric loudspeakers. The amplifiers were adjusted so that the SPL of the audible sound was at its maximum at the target area (center of the room). The average SPL of audible sound was 42.7 dB(A). The cross attack was tested by changing the position of the target device, as shown in Figure 10 (Right). In the figure, the blue circles indicate measurement points, where a sound level meter was set. Two parametric loudspeakers were set so that they would cross at the center point. Unlike the linear attack setup, this setup was not symmetrical and each parametric loudspeaker transmitted a different signal (i.e., a carrier wave and a sideband wave, respectively). We established $5 \times 5 = 25$ measurement points. As shown in the figure, we installed four dynamic speakers to fine-tune the SPL of ambient room noise. We configured the directions of the dynamic speakers such that noises were equally distributed throughout the room. We fixed the distance between the target device and two parametric loudspeakers

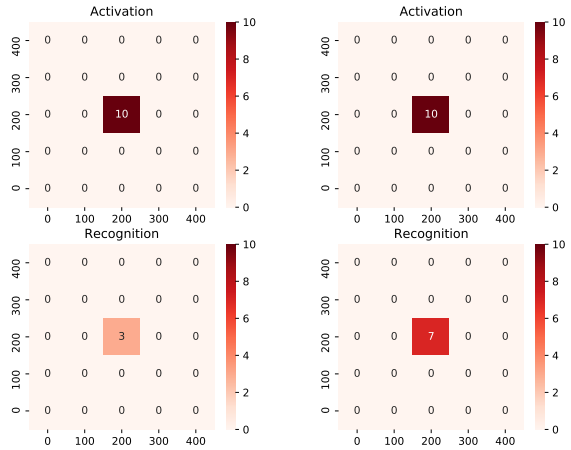


Fig. 11. Number of successful cross attacks at each position (max is 10). Top: Activation and Bottom: Recognition. Left: Google Home, and Right: Amazon Echo. The demodulation point was adjusted to the center, point (200, 200).

to $2\sqrt{2}$ m, and the SPL of noise was set to 43 dB(A).

At each position, the attack was repeated 10 times, with the number of successes counted. Figure 11 shows the results. The first finding was that the attack was successful only in the area targeted by the cross attack. Second, for the activation voice command, the attack success was 100% for both devices. Finally, although the success rate was low for voice recognition ("what's on my next schedule?"), it remains a realistic threat, given the fact that an adversary can repeat the attack until it succeeds.

5.5 Summary

Throughout this section, we evaluated attack feasibility. First, the experiments demonstrated that the attacks were successful over long distances. In the experiment room (500 cm $\times$ 500 cm), Google Home attacks were 100% successful at 350 cm and Amazon Echo attacks were more than 90% successful at 150 cm. The hallway experiments demonstrated that, for both devices, attacks were successful at distances greater than 10 m. Second, we discovered that the attacks were tolerant of environmental noise. For both devices, the attack success rate remained high at a noise SPL of 60 dB(A). This SPL corresponded to the SPL used for the experiments described in Section 6. Finally, the attacks were successful with various types and lengths of voice commands.

6 Human Study Experiments

In psychoacoustics, hearing is different from objective SPL measurements [31]. We tested to confirm whether the directional sound generated from parametric loudspeakers could be perceived by humans around the targeted device. To this end, we conducted extensive user study experiments to answer the RQ3: "Is the attack stealthy for nearby people and unrecognizable for them?" To complement the results of our human studies (subjective evaluation), SPL measurements were taken with the sound level meter (objective evaluation).

6.1 Experimental setups

Figure 10 presents an overview of the experimental setup. For the linear attack mode, both a parametric loudspeaker and a dynamic speaker were used to observe their differences. In the figure, the blue circles indicate measurement points, where a participant was seated. As the setup was symmetric in nature, $3 \times 5 = 15$ measurement points were set only in the right half. We omitted the left half to reduce the workload of the participants without sacrificing the generality of the results. The distance between the measurement points was set to 1 m. For the cross attack, two parametric loudspeakers were set so that they would cross at the center point. We established $5 \times 5 = 25$ measurement points, with a chair at each measurement point (See Appendix B, Figure 2).

The output power of the adversary's parametric / dynamic loudspeakers was adjusted so that the SPL of the audible sound (not the ultrasound) measured 3 m away from the parametric loudspeaker was 60 dB(A). Accounting for the inaudible sound wave, the total SPL was 120—130 dB(A) for all these settings. Finally, for the four dynamic speakers that generate $1/f$ noise, we adjusted the output power such that the audible SPL was 60 dB(A) at a distance of 3 m. For reference, the SPLs of common environmental noises are summarized in [27].

6.2 Human study overview

6.2.1 Participants

For the user study, we recruited 20 normal-hearing participants. Of these, 12 were female and eight were male, with ages ranging from 19 to 27. Thus, the participants were younger on average. Because younger people tend to have better hearing, we selected a severe condition to evaluate recognizability.

The participants consist of students at our university. We let the participants choose the preferred language from the two choices, Japanese and English. While 16 participants who selected Japanese are all native speakers of Japanese, other three participants who selected English were fluent in English but not necessary were the native speakers of English. Two of them are from Indonesia and the other is from China. For each participant, consent was obtained before enrolment. All participants were informed that they could quit the experiment whenever they desired. Other ethical considerations are discussed in section 7.

6.2.2 Procedure

For each setup, each participant was first directed to sit in a chair set at the position marked with the star symbol in Figure 10. Then, the height of loudspeaker(s) was adjusted so that the participant's sitting height matched the position of the loudspeaker(s). For each participant, the heights and angles of the speakers were fixed throughout the experiments. After the beginning of a session, a random word is emitted twice from the speaker at a random moment in time. A participant reports whether they recognize the word. If they recognize it, they write down the word that they recognized.

From the set of random words, those containing between 3 and 6 phonemes were selected. It was also ensured that the words would be difficult to predict beforehand, e.g., wake-up words typically used for voice assistance systems were avoided. Each participant repeated the sessions after moving to another chair.

To ensure the quality of the subjective evaluations, we used a silent task with each participant. During the silent task, no voice sounds were emitted. If a participant reported that they heard something during the silent task, the other results reported by the participant were considered unreliable and removed. Consequently, two participants' results were removed from the final analyses.

6.2.3 Evaluation of recognizability

To quantify the recognizability reported by the participants, we used a Jaccard index for the sets of letters in two words t and r, which are a test word and a reported word, respectively. For instance, if a test speech word is 'fest' and the reported word is 'test', the Jaccard index is computed as $J(\text{'fest'}, \text{'test'}) = 3/5 = 0.6$. For reference, a randomly sampled answer sheet reported by one of the participants is shown in Appendix B.

In total, for each measurement point, we collected 18 scores reported by the 18 participants. At least one score for each measurement point was In total, for each measurement point, we collected 18 scores , reported by the 18 participants. At least one score for each measurement point was omitted, as there was one silent task for each participant. To quantify the recognizability, the average of the reported scores was taken for each measurement point.

6.3 Results of the human study

Figure 12 shows the linear attack results. The heat maps represent the average Jaccard index scores. Notably, for the dynamic loudspeaker experiment, most participants successfully recognized the test speech words across a wide range. In fact, the test words were audible even behind the speaker. On the other hand, for the parametric loudspeaker experiment, the audible space was limited to a narrow area (i.e., the direction of directional sound propagation). The generated sound wave was somewhat inaudible over a short range owing to the fact that the generated ultrasonic beam moved forward before it was demodulated in the air.

Figure 13 shows the cross attack results. It is important to note that there seem to be no audible spaces in the room. However, as shown in the previous subsection 5.4, the cross attack was successful in emitting malicious voice commands to the voice assistance systems. This contradiction can be explained as follows: as the cross point was limited to a very narrow area, it did not "hear" the areas close to the participant's ears. Even if a participant was able to catch either a carrier wave or a lower sideband wave, they would not recognize them unless they caught both sound waves at a cross point. To complement the results of the human study, the results of the objective sound level meter evaluations are presented in Figure 14.

6.4 Summary

In this section, we examined the recognizability of sounds generated from parametric loudspeakers. For comparison, we also examined the characteristics of the sound generated by a dynamic speaker. Both the subjective and objective

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

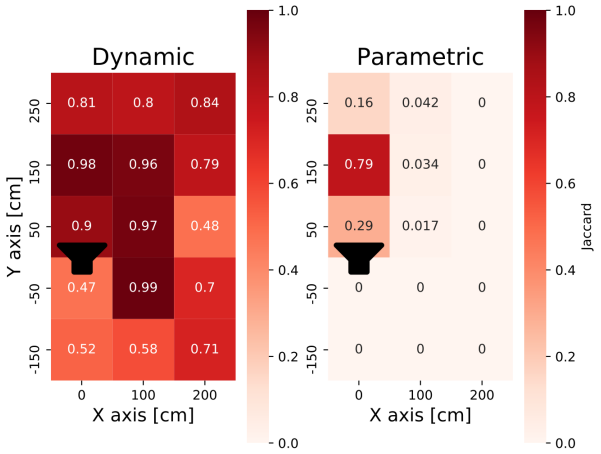


Fig. 12. Average Jaccard index scores of the linear attack measured in a 200 cm × 400 cm area. Left: dynamic speaker and Right: parametric loudspeaker. The point (0, 0) is defined as the location of the loudspeaker. User cannot hear the on space except in front of the parametric loudspeaker.

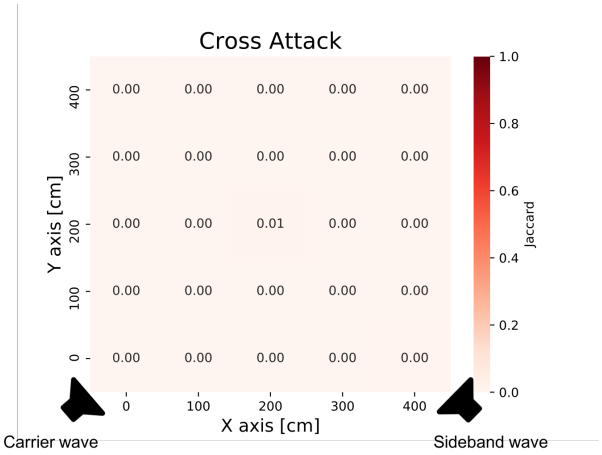


Fig. 13. Average Jaccard index scores for the cross attack measured in a of 400 cm × 400 cm area. The point (200, 200) is defined as the demodulation point. We found that the users cannot hear sound waves everywhere except in the center.

evaluations revealed that the directional sound generated from the parametric loudspeakers achieved sufficient unrecognizability to perform the Audio Hotspot Attack. Specifically, the sound generated with the cross attack was difficult for a human near the target device to perceive.

7 Discussion

In this section, we discuss the limitations and extensions of Audio Hotspot Attack, possible countermeasures against it, and ethical issues considered during the experiments.

7.1 Limitations and possible extensions

Because the Audio Hotspot Attack uses sound wave(s) to inject malicious voice commands, it will not succeed if there is an obstacle between the target device and the parametric loudspeaker(s) (e.g., a wall or a window). This limitation also applies to other inaudible voice command attacks [32], [33], [34]. One possible method of overcoming this limitation would be to install parametric loudspeaker(s) on a ceiling, thus creating a "sound shower." In fact, parametric loudspeakers are often mounted on ceilings to make sounds audible only at one point in the room, without the risk of interruption form an obstacle. Even when it is unrealistic to mount a parametric loudspeaker on the ceiling, it would still be effective to place it at a raised or a side position to ensure that the sound wave emitted avoids obstacles.

We used two smart speakers, Google Home and Amazon Echo, as examples of popular devices with voice assistance systems. Other types of voice assistance systems include smartphones, in-car navigation systems, and commercially available medical devices. Studying the effectiveness of the Audio Hotspot Attack on most of these other devices will be conducted in future studies; however, we did verify that the attack worked on several smartphones. Although the evaluation of the latter is not as thorough as that presented in section 5, some results have been given in the Appendix for reference.

Finally, although we sought to make these studies scientifically reproducible, the target devices are updated regularly. Furthermore, as the majority of the off-the-shelf voice assistant devices today run the speech recognition on the server side, it is prone to change over time. Therefore, once changes are made to the hardware or software in the voice assistance devices, other results may differ from the ones we obtained. As off-the-shelf products are "black box" in nature, it is difficult to fully understand how input sound waves are processed by the device's hardware and/or software. Therefore, to make the results of the experiments to be invariable and reproducible, it would be desirable to develop open-source hardware and software platforms, which would allow researchers to share and compare results using similar tools. At present, we are developing such a platform so that interested researchers can conduct further work on security and privacy issues related to voice assistance systems.

7.2 Countermeasures

Audio Hotspot Attack leverages the natural phenomenon of ultrasound self-demodulation in the air; therefore, it is not practical to try to block voice commands before they reach the target device. One possible solution is to detect the voice commands and differentiate them from others that are legitimate. There are two ways to achieve this goal. An easy and effective approach is to employ speaker recognition; in fact, smart speakers such as Google Home or Amazon Echo have already adopted this functionality. However, as discussed in Section 3, such approaches are still vulnerable to advanced replay or voice-morphing attacks. Therefore, we require methods that can detect voice commands being emitted from parametric loudspeakers. In the following section, we discuss three potential approaches to achieve this goal.

7.2.1 *Detecting ultrasonic sounds*

Although the ultrasounds emitted from a parametric loudspeaker are demodulated in air, there are un-demodulated ultrasonic components in the observed sound wave. Figure 15 shows the spectrogram of a speech signal emitted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2019.2953041, IEEE Transactions on Emerging Topics in Computing

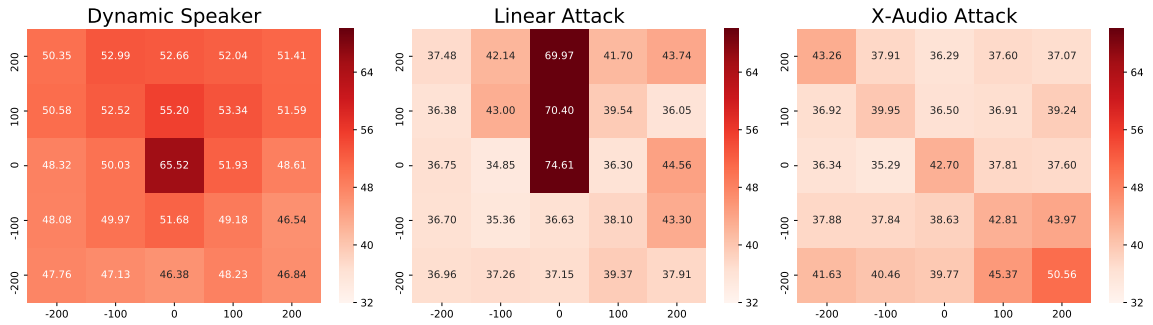


Fig. 14. SPL measured for the three attack modes. The unit for the numerical values is dB(A). The setup is same as in the human study. We have the speaker on the point (0,0) in the case of the dynamic speaker and linear attack. In the case of the X-Audio attack, (0, 0) is the demodulation point for voice commands.

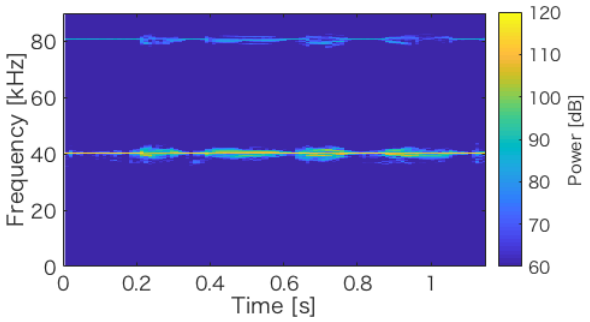


Fig. 15. Spectrogram of a speech signal emitted from a parametric loudspeaker. The signal was recorded with an ultrasonic microphone. The frequency range was set above 20 kHz (inaudible frequency). The content is “OK Google”.

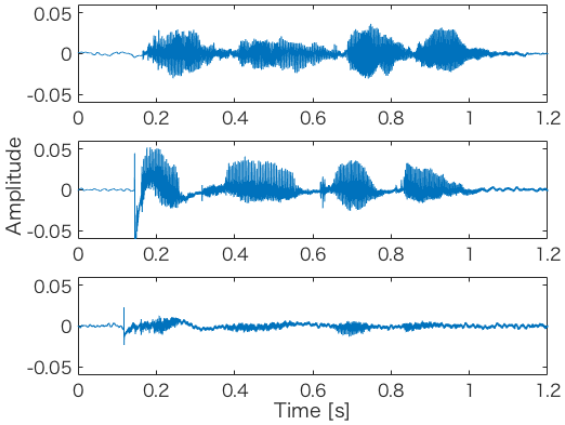


Fig. 17. Speech signals generated from a dynamic loudspeaker (top), a parametric loudspeaker (middle, linear attack), and Bottom: two parametric loudspeakers (bottom, cross attack). The content is “OK Google”.

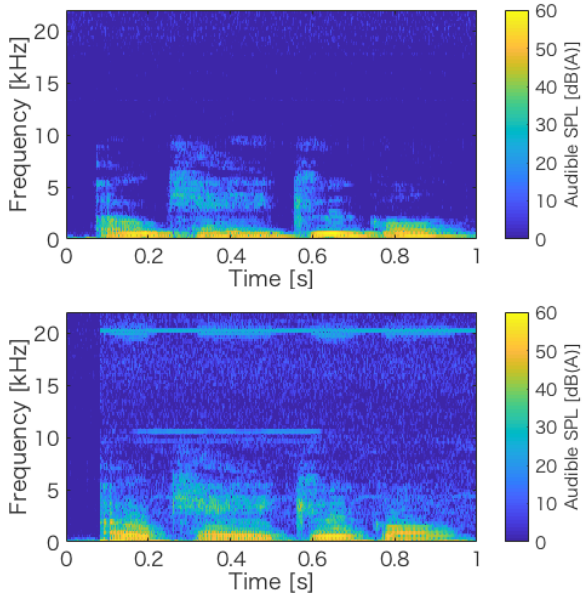


Fig. 16. Spectrogram of a speech signal emitted from a dynamic loudspeaker (top) and a parametric loudspeaker (bottom). The signals were recorded with a normal microphone. The frequency range was set below 20 kHz (audible frequency). We can see the folding noise at 10 kHz and 20 kHz in the bottom spectrogram. The content is “OK Google”.

from a parametric loudspeaker. The original speech data was “Ok Google,” which was generated using Amazon Polly (Ivy). In the spectrogram, the power of the ultrasonic component is around 40 kHz, which corresponds to the carrier frequency of the AM-modulated sound. A harmonic overtone around 80 kHz was also observed. Thus, even ultrasound is self-demodulated in the air, and it is possible to observe ultrasonic components of sound waves.

A straightforward approach to detecting such ultrasonic components is to apply an ultrasonic sensor. Although ultrasonic microphones are expensive, ultrasonic sensors are cheap and readily available. As Zhang et al. suggested [33], using MEMS microphones on mobile devices could be an alternative solution, as these microphones can sense acoustic sounds with frequencies higher than 20 kHz. Once a device detects the non-negligible amounts of ultrasonic components of a received sound wave, it may suspend the operation and require interaction with the device owner to resume the operation.

7.2.2 *Analyzing the frequency patterns of audible sounds*

Figure 16 presents the spectrograms of a voice signal (“OK Google” as spoken by Amazon Polly) emitted from a dynamic loudspeaker and a parametric loudspeaker. Although the original voice data was the same, there are different

characteristics in the frequency patterns of the observed sound waves. As can be seen in Eq. 4 (Section 2), the SPL of the sound wave generated from a parametric loudspeaker is proportional to the frequency of the original sound signal. This indicates that if the sound is emitted from a parametric loudspeaker, higher or lower frequency components are more or less likely to be observed, respectively, at the target. The horizontal lines shown in the lower spectrogram correspond to the *folding noise*, which is also known as *aliasing*. We can detect attacks if we observe the folding noise in spectrograms. To validate the effectiveness of this approach, we performed a brief experiment. From a given sound wave, we extracted components that had the frequencies above 10 kHz, which is over the audible frequency of 8 kHz. We then computed the power of the extracted sound wave. While the normal sound wave had almost zero power, the sound wave of the directional sound beam had non-zero power. By simply applying a threshold-based detection, we were able to distinguish the sound emitted by a loudspeaker from the one emitted by a parametric speaker with 100% accuracy.

Figure 17 shows speech signals emitted from a dynamic loudspeaker and parametric loudspeakers. Again, these speech signals were generated by the same original voice signal ("OK Google"), via Amazon Polly (Ivy). For the speech signals emitted from parametric loudspeakers (middle and lower panels in the figure), there is an intrinsic spike at the beginning of the speech signal. These spikes can be used as a fingerprint for detecting speech generated from a parametric loudspeaker. These spikes and other intrinsic characteristics can be used to differentiate speech generated from a parametric loudspeaker compared to speech generated from a regular voice using heuristics or machine learning-based approaches.

7.2.3 Voice Presentation Attack Detection (PAD) method

As inaudible voice command attacks will be combined with the presentation attacks, we can apply the presentation attack detection (PAD) method, which we assumed our target voice assistant systems had not implemented, to detect an Audio Hotspot Attack [9]. The ultimate countermeasure against such an attack is to be able to distinguish a synthesized voice from an authentic human voice. Liveness detection [33], [35], [36], which judges whether an input voice has come from a human or a dynamic speaker, is an example of the PAD method that could achieve this goal. In real environments, attacks on speech recognition devices are by means of the latter. Therefore, it would be sufficient for a voice assistance system to be able to judge whether a sound comes from a human or a dynamic speaker, even if it is unable to identify a specific individual. Voice Gesture [33], as proposed by Zhang et al., attempts to detect the movement of a person's mouth, by using changes in ultrasonic waves that occur as a consequence of the mouth movements and the position of the tongue when an approximately 20 kHz ultrasonic wave is emitted from a smart device (e.g., a smartphone or tablet) to the mouth of the user. This method detects differences in movement between a mouth and a dynamic speaker. The mouth movement changes for each pronunciation variation, whereas the surface of a dynamic speaker exhibits very little movement. The liveness detection method could be used to detect an Audio Hotspot Attack because ultrasonic transducers use fewer movements than the human mouth.

In our experiments, we have shown that simple rule-based or threshold-based detection work as countermeasures against the Audio Hotspot Attack. However, more robust countermeasures will be required in realistic environments. In [9], some typical countermeasure methods using the machine learning model are proposed. On the contrary, in [37], the authors pointed out that the machine-learning model does not work well for the datasets obtained in different setups. Overcoming the problem of overfitting to the specific datasets and/or environments is left for future work.

7.3 Ethical Considerations

7.3.1 Human study research

We performed a human study to test the unrecognizability of the Audio Hotspot Attack using parametric loudspeakers. The experiments were carefully designed such that they did not impose a burden on either the hearing or psychological states of the participants. The procedure for the human study was approved by the ethical review board at Waseda University. Prior to the experiments, we performed a pilot study to ensure the validity of our measures. Then, Participants were provided with all information required to make a meaningful decision as to whether or not they were willing to participate in the experiment (informed consent). We explained the reasons for conducting the study, what the experimental procedures, potential risks and benefits were, and the ways in which participants could get more information on the study. The SPL of the sound waves was sufficiently low such that it did not cause the participants any discomfort. Participants were also given two-minute breaks every ten minutes and were able to stop participating at any time without incurring any penalty.

7.3.2 Offensive security research

The objective of this work was to explore the feasibility of the threats caused by inaudible voice command attacks. It was demonstrated that inaudible voice command attacks are viable through methods such as an Audio Hotspot Attack. Although this attack was proof of concept, we have also provided potential countermeasures by which they can be counteracted. Furthermore, with the aid of the national CERT, we have initiated communication regarding this with several manufacturers of voice assistance systems. Feedback, including plans for implementing the countermeasures within the products concerned, has been received. By the time of publication, vendor reaction will have been received and will also be reportable.

8 Related Works

Voice command attacks

DolphinAttack [6], [34] is an attack that inputs inaudible commands on a target microphone by AM modulating the sound, with the ultrasound as the carrier wave. The basic idea is based on the fact that the output of the MEMS and ECM microphones that are mounted on smartphones has nonlinearity [32], [38]. A nonlinear term is obtained

by squaring the input signal in the output signal when an AM ultrasonic signal by the prepared voice is inputted to the microphone. That is, the output of the microphone receiving the AM-modulated ultrasound includes the frequency component of the original speech signal, and the speech recognition algorithm of the system that received the low-pass filtered signal is applied as recognized speech, even though the input signal only generates high-frequency waves. The output generated by the nonlinear term has a smaller voltage value than the normal output and therefore it is easy to detect.

On the other hand, in an Audio Hotspot Attack, there is a marked difference in that *audible* sounds, which have been self-demodulated from the ultrasound waves, are received by a target device. This phenomenon is established because air is nonlinear and demodulates the AM-modulated ultrasonic signal, as shown in Section 2. Indeed, we cannot eliminate nonlinearity from the air because it is a natural phenomenon. In other words, even if microphone nonlinearity is completely removed, Audio Hotspot Attacks are still feasible even though inaudible voice commands are infeasible. In addition, Audio Hotspot Attacks can be employed from greater distances than DolphinAttacks because ultrasound has higher-than-audible frequencies, and therefore, it decays faster.

Audio adversarial examples

Audio Adversarial Examples [39] apply Image Adversarial Example [40], [41] techniques to voice waves. Adversarial examples are input to machine learning models that an attacker has intentionally designed to cause the model to make a mistake. The recognition results of the machine learning model are easily affected by a small amount of perturbation (small noise). Adding a small amount of noise to the original sound intentionally results in erroneous recognition. Therefore, Audio Adversarial Examples can be misidentified as arbitrary commands. The user cannot notice the subtle additional noise and targeted malicious commands are therefore executed on the voice assistant.

Existing attacks assume that software or hardware vulnerabilities are related to attack successes. Hidden voice commands and Audio adversarial examples use the vulnerabilities inherent to machine learning, and DolphinAttack uses vulnerabilities of MEMS microphones. On the other hand, the Audio Hotspot Attack uses a physical phenomenon i.e.,non-linearity in the air. Audio Hotspot Attack countermeasures are therefore more difficult to create given they do not rely on any existing vulnerabilities.

9 Conclusion

In this work, we proposed a new inaudible voice command attack named "Audio Hotspot Attack." Its feasibility was evaluated through extensive user studies and reproducible experiments. We demonstrated that when directional sounds are emitted from parametric loudspeakers and not perceived by a nearby person, attacks can succeed over relatively long distances (2–4 m in a small room and up to 10+ m in a hallway); further, these attacks are tolerant against environmental noises. Although the Audio Hotspot Attack is currently a proof-of-concept, possible countermeasures to render the threats unsuccessful have been provided. The proposed attack uses ultrasound self-demodulation, which is a parametric phenomenon. We believe that this concept sheds new light onto ongoing security research focused on mobile and IoT devices, from the viewpoint of acoustic inputs.

Acknowledgments

A part of this work was supported by JSPS Grant-in-Aid for Challenging Research (Exploratory), Grant Number 18K19789.

References

[1] R. Iijima, S. Minami, Z. Yunao, T. Takehisa, T. Takahashi, Y. Oikawa, and T. Mori, "POSTER: Audio hotspot attack: An attack on voice assistance systems using directional sound beams, (poster presentation)," in *Proc. of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 2222–2224. [Online]. Available: http://doi.acm.org/10.1145/3243734.3278497

[2] Apple. (2018) ios - siri. [Online]. Available: https://www.apple.com/ios/siri/

[3] Google. (2018) google-assistant. [Online]. Available: https://assistant.google.com

[4] Amazon. (2018) Amazon alexa. [Online]. Available: https://alexa.amazon.com/spa/index.html

[5] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. A. Wagner, and W. Zhou, "Hidden voice commands," in *Proceedings of 25th USENIX Security Symposium*, 2016, pp. 513–530.

[6] G. Zhang *et al.*, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC, CCS*, 2017, pp. 103–117.

[7] M. Yoneyama *et al.*, "The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design," *The Journal of the Acoustical Society of America*, vol. 73, no. 5, pp. 1532–1536, 1983. [Online]. Available: https://doi.org/10.1121/1.389414

[8] P. J. Westervelt, "Parametric acoustic array," *The Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 535–537, 1963.

[9] M. Sahidullah *et al.*, "Introduction to voice presentation attack detection and recent advances," in *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection, Second Edition*, 2019, pp. 321–361. [Online]. Available: https://doi.org/10.1007/978-3-319-92627-8_15

[10] W.-S. Gan *et al.*, "A review of parametric acoustic array in air," *Applied Acoustics*, vol. 73, no. 12, pp. 1211 – 1219, 2012, parametric Acoustic Array: Theory, Advancement and Applications. [Online]. Available: www.sciencedirect.com/science/article/pii/S0003682X12000904

[11] S. N. Gurbatov, O. V. Rudenko, and A. I. Saichev, *Waves and Structures in Nonlinear Nondispersive Media [electronic resource] : General Theory and Applications to Nonlinear Acoustics*, 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[12] I. O. for Standardization, *ISO/IEC 30107. Information technology – biometric presentation attack detection.* International Organization for Standardization, 2016.

[13] T. Kinnunen *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2–6. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1111.html

[14] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proceedings of INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 930–934. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_0930.html

[15] N. S. Dibya Mukhopadhyay, Maliheh Shirvanian, "All your voices are belong to us: Stealing voices to fool humans and machines." in *In Proceedings of the European Symposium on Research in Computer Security.*, ser. Springer, 2015, pp. 599–621.

[16] Z. Wu *et al.*, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA*, 2014, pp. 1–5.

[17] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop*, 2016, p. 125. [Online]. Available: http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html

[18] eMarketer. (2018) Amazon echo losing share as speaker rivalry heats up. [Online]. Available: https://www.emarketer.com/content/amazon-echo-losing-share-as-speaker-rivalry-heats-up

[19] switchscience. (2017) Super directional speaker kit. [Online]. Available: https://international.switch-science.com/catalog/1842/

[20] Accuphase. (2001) Acouspade. [Online]. Available: https://www.accuphase.co.jp/cat/pro-power.pdf

[21] U. A. Technologies. (2018) Acouspade. [Online]. Available: http://ultrasonic-audio.com/acouspade-technical-specification/

[22] YAMAHA. (2018) Monitor speaker ms101 iii owner's manual. [Online]. Available: https://www.manualslib.com/manual/267200/Yamaha-Ms101-Iii.html

[23] RION. (2018) The nl series line up. [Online]. Available: http://scantekinc.com/files/PDFs/Rion/Rion_NL-21_22_31_32_Series.pdf

[24] B&K. (2018) Product data: Teds microphones (bp2225). [Online]. Available: https://www.bksv.com/-/media/literature/Product-Data/bp2225.ashx

[25] motu. (2018) Ultralitemk4 overview. [Online]. Available: http://motu.com/products/proaudio/ultralite-mk4

[26] Amazon. (2018) Amazon polly. [Online]. Available: https://console.aws.amazon.com/polly/home/SynthesizeSpeech

[27] Center for Hearing and Communication. (2018) Common environmental noise levels. [Online]. Available: http://chchearing.org/noise/common-environmental-noise-levels/

[28] T. D. Rossing, *Springer Handbook of Acoustics*, 2nd ed. Springer, 2014.

[29] S. Pigeon. (2019) Babble noise -frequency-shaped babble noise generator. [Online]. Available: https://mynoise.net/NoiseMachines/babbleNoiseGenerator.php

[30] MathWorks. (1994–2019) Matlab. [Online]. Available: https://www.mathworks.com/products/matlab.html

[31] S. Grondin, *Psychology of Perception*. Springer, 2016.

[32] T. Vaidya *et al.*, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *9th USENIX Workshop on Offensive Technologies, WOOT*, 2015. [Online]. Available: https://www.usenix.org/conference/woot15/workshop-program/presentation/vaidya

[33] L. Zhang *et al.*, "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*. ACM, 2017, pp. 57–71. [Online]. Available: https://doi.org/10.1145/3133956.3133962

[34] N. Roy *et al.*, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys'17*, 2017, pp. 2–14. [Online]. Available: http://doi.acm.org/10.1145/3081333.3081366

[35] L. Zhang *et al.*, "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*, 2016, pp. 1080–1091. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978296

[36] S. Shiota *et al.*, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, 2016, pp. 259–263. [Online]. Available: https://doi.org/10.21437/Odyssey.2016-37

[37] P. Korshunov and S. Marcel, "A cross-database study of voice presentation attack detection," in *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection, Second Edition*, 2019, pp. 363–389. [Online]. Available: https://doi.org/10.1007/978-3-319-92627-8_16

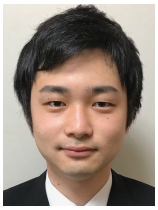
[38] D. F. Kune *et al.*, "Ghost talk: Mitigating EMI signal injection attacks against analog sensors," in *Proceedings of 2013 IEEE Symposium on Security and Privacy, SP*. IEEE Computer Society, 2013, pp. 145–159. [Online]. Available: https://doi.org/10.1109/SP.2013.20

[39] N. Carlini and D. A. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*. IEEE, 2018, pp. 1–7. [Online]. Available: https://doi.org/10.1109/SPW.2018.00009

[40] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013. [Online]. Available: http://arxiv.org/abs/1312.6199

[41] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

AUTHOR BIBLIOGRAPHY



Ryo Iijima received his B.E. and M.E. degrees from Waseda University in 2018 and 2019, respectively. He has been pursuing his Ph.D. at Waseda University since 2019. His research interests include acoustic security and machine learning. He received the best paper award at the Computer Security Symposium, and best student research award at Cyber Security Symposium Dougo in 2018. He is a member of IPSJ.

Shota Minami received his B.E. and M.E. degrees in Intermedia Art and Science from Waseda University in 2017 and 2019, respectively. His research interest is directivity control of parametric loudspeaker using ultrasonic transducer array. He is a member of ASJ.

Yunao Zhou received his B.E. degree in the Information Security from Xidian University in 2016. He has been pursuing his M.E. degree at Waseda University since 2018. His research interest is cyber security.

Tatsuya Takehisa has been an invited advisor with the National Institute of Information and Communications Technology since 2013. He has been as an advisor with the Nissin Inc since 1999. He has over 20 years of experience as an embedded system engineer. His main research interests are cybersecurity and embedded system security. He is a member of IEEE and IEICE.

Takeshi Takahashi received his Ph.D. degree in telecommunication from Waseda University in 2005. He was with the Tampere University of Technology as a Researcher from 2002 to 2004, and Roland Berger Ltd., as a Business Consultant, from 2005 to 2009. Since 2009, he has been with the National Institute of Information and Communications Technology, where he is currently a Research Manager. His research interests include cyber security and machine learning. He is a member of ACM, IEEE, and IEICE.

Yasuhiro Oikawa received his Ph.D. degrees in Electrical Engineering from Waseda University in 1995, 1997, and 2001, respectively. He is a professor of the Department of Intermedia Art and Science, Waseda University. His main research interests are communication acoustics and digital signal processing of acoustic signals. He is a member of IEEE, ACM, ASA, ASJ, IEICE, IPSJ, VRSJ, and AIJ.

Tatsuya Mori is currently a professor at Waseda University, Tokyo, Japan. He received his B.E. and M.E. degrees in applied physics, and Ph.D. degree in information science from Waseda University, in 1997, 1999 and 2005, respectively. He joined the NTT lab in 1999. Since then, he has been engaged in the research of measurement and analysis of networks and cyber security. From Mar 2007 to Mar 2008, he was a visiting researcher at the University of Wisconsin-Madison. He received the Telecom System Technology Award from TAF in 2010 and the Best Paper Awards from IEICE and IEEE/ACM COMSNETS in 2009 and 2010, respectively. He is a member of ACM, IEEE, IEICE, IPSJ, and USENIX.