EarArray: Defending against DolphinAttack via Acoustic Attenuation

Guoming Zhang
Zhejiang University
realzgm@zju.edu.cn

Xiaoyu Ji*
Zhejiang University
xji@zju.edu.cn

Xinfeng Li
Zhejiang University
xinfengli@zju.edu.cn

Gang Qu
University of Maryland
gangqu@umd.edu

Wenyuan Xu*
Zhejiang University
xuwenyuan@zju.edu.cn

***Abstract*—DolphinAttacks (i.e., inaudible voice commands) modulate audible voices over ultrasounds to inject malicious commands silently into voice assistants and manipulate controlled systems (e.g., doors or smart speakers). Eliminating DolphinAttacks is challenging if ever possible since it requires to modify the microphone hardware. In this paper, we design EarArray, a lightweight method that can not only detect such attacks but also identify the direction of attackers without requiring any extra hardware or hardware modification. Essentially, inaudible voice commands are modulated on ultrasounds that inherently attenuate faster than the one of audible sounds. By inspecting the command sound signals via the built-in multiple microphones on smart devices, EarArray is able to estimate the attenuation rate and thus detect the attacks. We propose a model of the propagation of audible sounds and ultrasounds from the sound source to a voice assistant, e.g., a smart speaker, and illustrate the underlying principle and its feasibility. We implemented EarArray using two specially-designed microphone arrays and our experiments show that EarArray can detect inaudible voice commands with an accuracy of 99% and recognize the direction of the attackers with an accuracy of 97.89%.**

I. Introduction

More than 3.25 billion voice assistants (e.g., Siri, Alexa) have been installed around the world, and it is anticipated that by 2023 the number will reach up to eight billion [1]. Researchers have identified various attacks against such systems, and one of the most devastating attacks is DolphinAttack [14], whereby attackers can inject inaudible voice commands and performs various malicious attacks, such as open a door, make a phone call, or place an order. DolphinAttacks modulate malicious voice commands onto ultrasounds and thus create inaudible voice commands. As the ultrasound is received by a microphone, its non-linearity vulnerability will demodulate the voice command from the ultrasound carrier into the baseband and the injected command exhibits almost no difference from the audible command.

To defend against DolphinAttacks, researchers have proposed two types of strategies. The first class detects the attacks by analyzing the subtle yet distinct characteristics embedded in the received sound signals, e.g, the unique high-frequency components caused by the demodulation process in the microphones [14] or the nonlinearity distortion created when the malicious voice commands pass through microphone circuits [35]. However, a sophisticated attacker [36] can remove such distinct characteristics and bypass the detection. The second class is to actively cancel the malicious inaudible voice commands by emitting an inverted ultrasound [36]. Such methods not only rely on extra ultrasound devices but also require to constantly emit ultrasound, which has shown to induce health issues, e.g., hearing loss, nausea, headache [2], [3], and to repel pets, such as dog and cat.

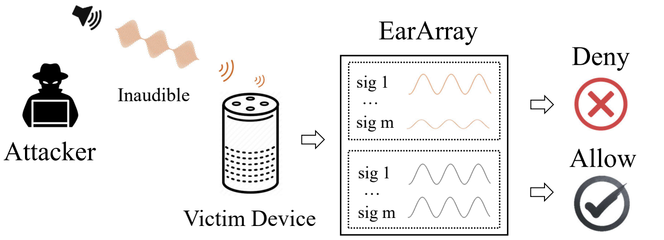


Fig. 1. When an inaudible voice command (i.e., DolphinAttack) is played to a smart speaker, the inherent frequency difference will result in measurable discrepancy in terms of propagation attenuation. By measuring the attenuation properties of the incoming sound, EarArray can recognize whether it is an inaudible voice command and even tell the direction of the attacker. The sig 1 represents the voice signal captured by the first microphone.

In this paper, we proposed a lightweight detection method, EarArray, which requires no extra hardware or hardware modification. Instead of utilizing the signal distortion caused by the microphones, EarArray looks into the propagation difference between inaudible voice commands (i.e., ultrasound) and audible ones. As depicted in Fig. 1, when a voice command is played by a speaker, the voice propagates to the smart device (e.g., smartphone and smart speaker) and reaches to its microphones at various times depending on the distances between the speaker and the microphones. On the smart device side, each microphone will receive a sound with the amplitude inversely proportional to the square of distance as well as the attenuation rate. Notably, the attenuation rate is, in terms, proportional to the square of sound frequency. Therefore, EarArray shall be able to estimate whether the command is an audible one or an inaudible one with the measurements obtained by multiple microphones (microphone array, for example) on a smart device because the decaying rate of ultrasounds ($> 20kHz$) is larger than audible sounds (typically below $5kHz$ for human voice). The advantage of EarArray is that it relies on the physics of sound propaga-

*Corresponding author

Network and Distributed Systems Security (NDSS) Symposium 2021
21-25 February 2021, Virtual
ISBN 1-891562-66-5
https://dx.doi.org/10.14722/ndss.2021.24551
www.ndss-symposium.org

tion, and will not be affected by the microphones' hardware characteristics.

EarArray utilizes an interesting yet challenging intuition, and the effectiveness of EarArray depends on answering the following questions. (1) Is the attenuation rate difference between the audible sounds and ultrasound sound large enough to be measured and utilized? (2) How to estimate the propagation attenuation efficiently? (3) Since the attacker may hide at any direction to the smart devices, will it always be possible to detect inaudible commands regardless of where she is? To answer the aforementioned questions, we first theoretically model the propagation of sound in terms of attenuation to quantify the measurable difference between ultrasound and sound. Then we specially designed two microphone arrays with three (mimicking the case of a smartphone) and five microphones (mimicking the case of a smart speaker like Amazon Echo) respectively. The specially-designed microphone arrays are placed on a cuboid and a cylinder and a multiplex data acquisition card is used to collect audio data from the channels of the microphone array simultaneously. By doing this, EarArray is able to estimate the attenuation rate by measuring the amplitude of the recordings from each of the microphone channels, calculates the power spectral density of the measured signal, and extracts three representative features. To be lightweight, EarArray is a software-based solution and can be integrated into existing commercial products such as smartphones and Amazon Echo without involving any extra hardware. EarArray makes use of the three key features and can utilize a simple machine learning algorithm, i.e., a support vector machine (SVM) to identify inaudible commands. In practice, EarArray can be installed on smart speakers and smartphones, as long as the device has three or more microphones [*]. To better defend against inaudible voice commands with EarArray, the smart speaker manufacturer can further optimize the distribution of microphones, e.g., keeping the angles of microphones in different plane, which is already a solution for most smart speakers such as Amazon Echo, as shown in Fig. 4. We extensively evaluated the performance of EarArray by varying the attack location, angles, ambient noises, and the carrier frequency of inaudible voice commands, etc. Our experiments show that EarArray can be effective and robust in various conditions.

In summary, our contributions are summarized as follows:

- We discovered that the propagation attenuation of audible commands and inaudible ones is different and thus can be used to detect DolphinAttacks. We theoretically analyze the attenuation difference by simulating the sound propagation over a microphone array.
- We designed EarArray that detects DolphinAttacks by estimating the propagation attenuation of voice commands.
- We implemented two prototypes of EarArray and evaluated the performance of EarArray with two specially-designed microphone arrays. EarArray can detect inaudible voice commands with an accuracy of 99% and recognize the direction of the attackers with an accuracy of 97.89%.

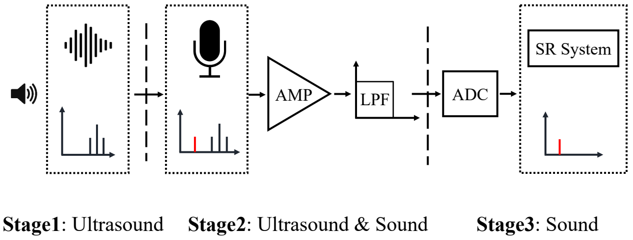


Fig. 2. The transformation progress of inaudible voice commands. By modulating audible voice commands on ultrasound (e.g., Amplitude Modulation), the voice commands can be inaudible. By utilizing the nonlinearity of microphones, the voice commands can be demodulated from the high-frequency carrier and then recognized by a speech recognition system.

II. Rationale of Defending DolphinAttack by Acoustic Attenuation

In this section, we first review the inaudible voice command attack known as DolphinAttack [14], then we present the basics of the attenuation of acoustic waves and give the rationale why it can be used to detect DolphinAttack. Finally, we analyze popular smart devices and show the feasibility of detecting DolphinAttacks from the support of multiple microphones.

A. DolphinAttack: An Inaudible Voice Commands Attack

The key idea of inaudible voice commands attack [14] is to modulate voice commands on ultrasonic carriers such that these inaudible commands can be captured by the microphone and demodulated back to the original voice commands. Since the frequencies at which the modulated voice commands propagate in the air are above 20 kHz, this kind of attack is completely inaudible to human ears and hard to be detected by human.

Fig. 2 shows the three stages of how the inaudible voice commands are transformed. In Stage 1, the voice commands are Amplitude-modulated (AM) on ultrasound carrier (e.g., 25 kHz) and therefore there are only high-frequency ultrasound components shown in the frequency domain. In Stage 2, both ultrasound and the low-frequency voice commands are recovered by using the nonlinearity of microphone hardware. In the final Stage 3, high-frequency ultrasound component has been filtered by the low-pass filter and only the voice commands remain. Thus the demodulated voice command will be the same as normal voice commands, making it very difficult to directly detect the inaudible voice commands attack. Especially, the attack voice commands appear just after the microphone module.

Our defense method against such attacks is inspired by the physical phenomenon that the incident wave of different frequencies traveling around a geometrical object (such as smart devices) will have different attenuation properties. This is because that the attenuation of the acoustic wave is directly related to the frequency, distance, and obstacle, etc. As we will elaborate next in Section II-B. Therefore, we can use the attenuation distinction of ultrasound and sound perceived from the smart devices to detect inaudible voice command attacks.

B. Attenuation of Acoustic Sounds

Acoustic attenuation describes that the intensity of an acoustic wave decreases as the wave propagates in the medium.

[*] Nowadays almost all smartphones have at least three microphones for noise reduction, as shown in Tab. I

Here we consider three main sources for the acoustic attenuation: 1) the inverse-square law; 2) sound absorption; and 3) diffraction.

1) The inverse-square law: The inverse-square law states that the intensity of waves is inversely proportional to the square of the distance (d) from the source of the wave. As the acoustic wave propagates in the air, the circumference of the acoustic circle expands larger with the increase of the traveling distance. So the energy per unit length decreases. Theoretically, the attenuation of wave (L_{inv}) can be expressed as:

$$L_{inv} \propto d^2 \tag{1}$$

According to Eq.1, the sound pressure level (SPL) of sound received by microphones located at places with different distances from the sound source will vary. This type of attenuation is related to the traveling distance.

2) Absorption Attenuation: This type of attenuation is caused by thermal or viscous effects and is related to the acoustic frequency. The thermal effect is that the coherent molecular motion of the sound waves is transformed into the incoherent molecular motion in the air, which directly transmits the vibration to the medium as heat. Another cause is the energy consumption caused by the viscosity of the air and the attenuation of sound in air also varies with temperature and humidity [15]. This type of attenuation is a power law frequency-dependent acoustic attenuation, and can be expressed as:

$$E(d + \Delta d) = E(d)e^{-a(w)\Delta d} \tag{2}$$

where E denotes the amplitude of an acoustic field variable, Δd represents wave traveling distance, w denotes the angular frequency of wave, $a(w)$ is the attenuation coefficient, and

$$a(w) = a_0 w^n, n \in [0, 2] \tag{3}$$

and a_0 and n are tissue-dependent attenuation parameters [16].

From Eq. (2) and (3), we can conclude that attenuation increases with acoustic frequency and distance. Therefore, high-frequency signals received by different microphones placed at different positions will differ more than low-frequency signals. This phenomenon inspires us to detect the inaudible voice commands by using multiple microphones.

3) Diffraction Attenuation: Diffraction [15] occurs when the spreading of the wave bends around obstacles, corners, and through openings. For example, when an obstacle is in the path of a spreading wave, the wave will bend around the obstacle and spread into the shadow regions behind it. The amount of diffraction will be inversely proportional to the acoustic frequency (f):

$$L_{dif} \propto f \tag{4}$$

where L_{dif} represents the attenuation caused by diffraction.

When the wavelength of a low-frequency sound is equal to or larger than the size of a smart device, the effect of diffraction of low-frequency sound will be obvious. Large wavelength waves will diffract around the edge of the smart device and not be decayed. So only small portion of the wave will be scattered from smart device. The intensity of the sound received by microphones at different positions will not vary significantly. But for high-frequency ultrasound, the propagation process is more sensitive to distance and obstacles as its wavelength is extremely small, ultrasound traveling toward multiple microphones will reach each microphone with different attenuation.

TABLE I. A LIST OF POPULAR SMART DEVICES AND THEIR MICROPHONE ARRAYS.

Type	Manuf.	Model	MICs	Distribution
Smartphone & Tablet & PC & Wearable device	Apple	iPhone 11	3	3-D
	Apple	iPhone XR	4	3-D
	Apple	iPhone X	4	3-D
	Apple	iPhone 8	4	3-D
	Apple	iPhone 7	4	3-D
	HUAWEI	Mate 10	3	3-D
	HUAWEI	Mate 9	4	3-D
	HUAWEI	P40 Pro	2	3-D ‡
	SAMSUNG	Note 10 Plus	3	3-D
	SAMSUNG	S20 Ultra	3	3-D
	SAMSUNG	Note8	2	3-D
	Apple	ipad pro	5	3-D
	Apple	Macbook pro	3	3-D
	Apple	Airpods pro∗	2	3-D‡
Smart speaker	Amazon	Amazon Echo	7	2-D
	Google	Google Home	2	2-D
	Alibaba	Tmall Genie	6	2-D
	Mi	Xiaomi Speaker	6	2-D
	Jingdong	DingDong mini2	6	2-D

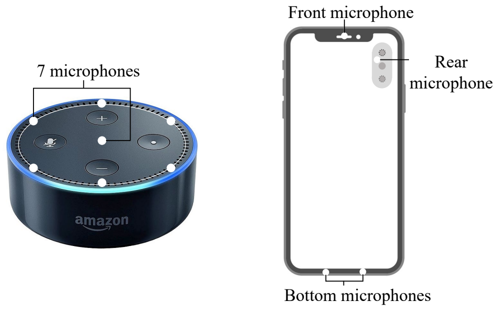
* Each side has one microphone.
‡ Not on the same surface.

Remark: Based on the above analysis, we conclude that the incident wave with different frequencies will produce totally different attenuation property after encountering smart devices. The frequency of inaudible voice commands is in range of 20 kHz to 50 kHz while normal voice commands have frequency 50 Hz to 2 kHz. This suggests that it might be possible to detect the inaudible voice commands attack by measuring and analyzing the attenuation property from the microphones of smart devices. We will demonstrate this later in the paper.

C. Microphone Arrays on Commercial Smart Devices

As the real sound fields are three-dimensional, to fully measure and analyze the sound attenuation property, the microphones on smart devices should be located at different positions and facing different directions. Table I summarizes representative smart devices. We can see that most of the smartphones have multiple microphones which form a spatial 3–D microphone array (the microphones are located on different sides). Fig 3 shows the distribution of microphones on Amazon Echo and iPhone XR. For Amazon Echo, each microphone is located on the top surface which can be classified as 2-D distribution. But iPhone XR has four microphones that are located on different sides, which belongs to 3-D distribution. All of the popular smart speakers, however, adopt the planar microphone arrays with 2-7 microphones placed on the top surface [13].

Although some of the smart devices have multiple microphones in 3-D arrays, users don't have the permission to record multi-channel sound. To the best of our knowledge, all the Apple mobile devices or iOS applications even didn't support stereo sound recording until iPhone XS was released [17], but iPhone XS can only record stereo sound in videos. For most



(a) Amazon Echo (b) iphone XR

Fig. 3. Modern smart devices support multiple microphone to obtain high-quality audio. The microphone array of Amazon Echo (7 microphones) [21] and iPhone XR (4 microphones) [18].

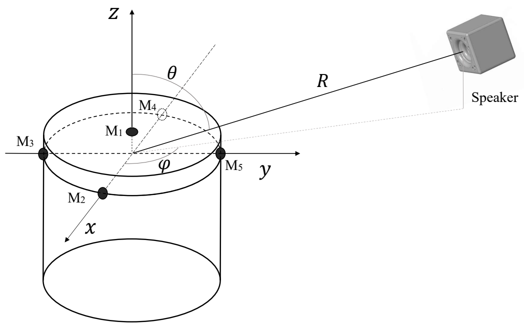


Fig. 4. The propagation model for sound transmission from a speaker to a voice assistant (e.g., Amazon Echo). The sound source can be around the smart speaker, by varying the distance parameter and the angle parameters including θ and φ.

Android devices, we can use a stereo recording application to capture sound in stereo with built-in microphones, nevertheless, the multi-channel recording isn't supported either. It is noteworthy that Siri and other voice assistants can pick up sounds using multiple built-in microphones, thus, the smart devices with multiple microphones have the ability to capture multi-channel signal. On these devices, our proposed defending method will not need to install any new microphones.

III. Acoustic Modeling

In this section, we first model the attenuation of audible and inaudible sound in the process of propagation. Then, the sound field distribution around a cylinder was simulated using COMSOL [20], and measured by five microphones located at different positions of the cylinder.

A. Theoretical Analysis

Fig. 4 depicts the scenario of the sound propagation when the sound meets a cylinder. Assume that the signal emitted from speaker is $x(t)$ and $y_i(t)$ is the signal received by i_{th} microphone. Without loss of generality, let $x(t)$ be a single frequency signal with frequency f when the sound is audible; in the case of an inaudible voice attack, $x(t)$ will be an amplitude modulated signal, the frequency of baseband signal will be the same as the audible sound, and the carrier frequency is f_c, which is a high-frequency signal.

Audible signal. After the audible sound propagates distance d_i, without considering the frequency response of microphone, the received signal at the ith microphone can be modeled as:

$$y_i(t) = h(d_i, f, \gamma_i) cos(2\pi ft + \frac{d_i}{c}) \tag{5}$$

where h represents the transfer function after transmitting from the speaker to the microphone which is affected by distance d_i, frequency f, and γ_i which is the obstacle-dependent attenuation parameter, c is the acoustic velocity. The transfer function will affect the received signal's phase and strength.

Inaudible signal. In this case, the inaudible signal is the amplitude modulated signal. After propagating close to the i_{th} microphone, without considering the phase change, the signal can be expressed as:

$$y_i(t) = h(d_i, f, f_c, \gamma_i) cos(2\pi f_c t)(1 + cos(2\pi ft)) \tag{6}$$

When the inaudible modulated signal is captured by the microphone, the modulated low frequency sound can be successfully demodulated and recovered from the nonlinearity of microphone circuits [14]. Without considering the attenuation of the microphone, the received signal can be expressed as:

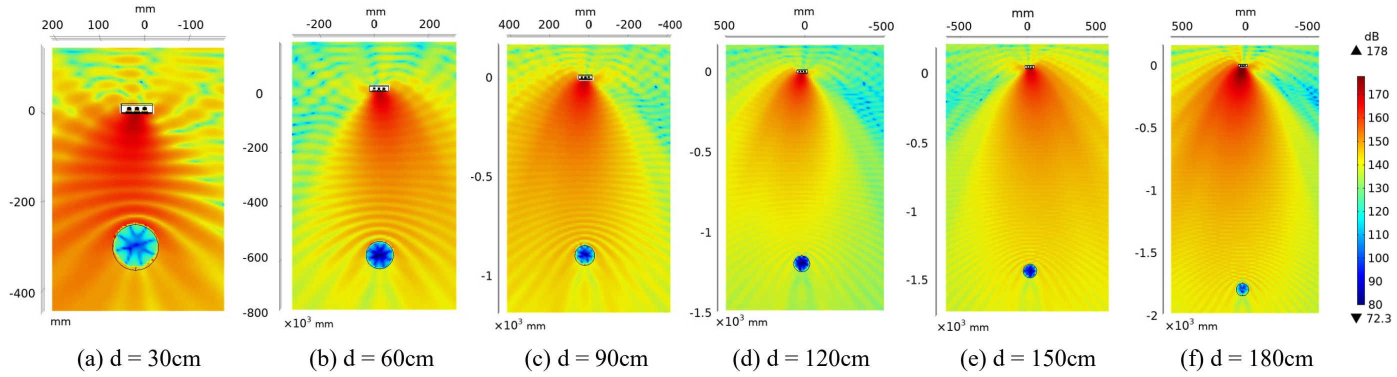
$$y_i(t) = h(d_i, f, f_c, \gamma_i) cos(2\pi f(t + \frac{d_i}{c})) \tag{7}$$

Comparing Eq. 5 and Eq. 7, we observe that the attenuation transfer function of attack signal is also related to the carrier frequency f_c, the attenuation will increase with frequency, thus, the attenuation difference of different paths is more significant.

B. Simulation

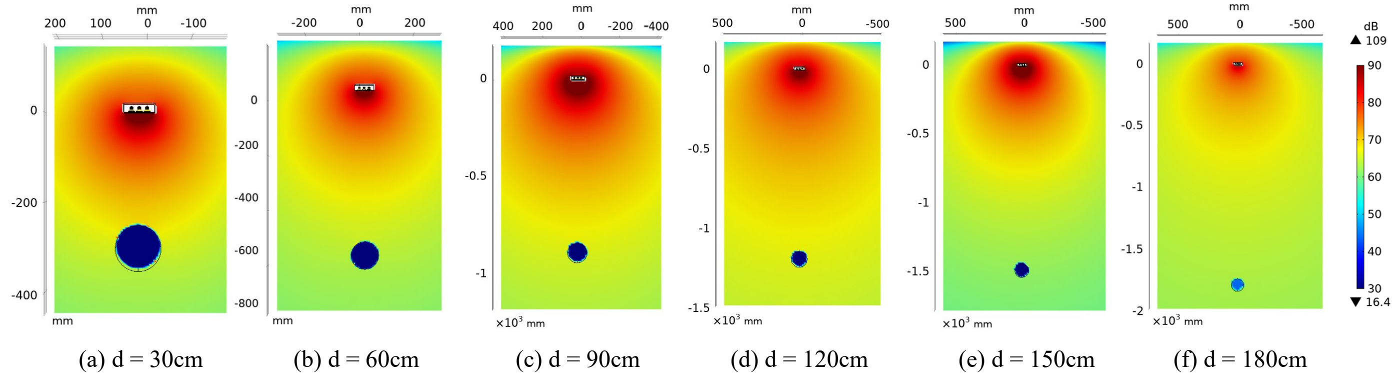
As mentioned above, due to the impact of frequency and shelter of surface, the sound field distribution surrounding the smart speaker is spatial-dependent. To evaluate and demonstrate the spatial properties of the sound field, we design a simulation experiment with different distances between the cylinder and the speaker array using COMSOL [20]. Fig. 5 and Fig. 6 show the settings and the blue cylinder represents the smart speaker while the flat box represents the speaker array. In the simulation, the distance between the sound source and the speaker ranges from 30 cm to 180 cm, the audible frequency is set to 500 Hz, and the inaudible frequency is set to 25 kHz.

As shown in Fig. 5, the sound field distribution of the 25 kHz signal generated by the speaker array is concentrated in the propagation direction of sound waves. There is almost no energy in the opposite direction, which represents the direction of high-frequency waves. And the microphone in the direction of the high-frequency signal receives the strongest energy. The microphones on the left and right also receive a certain amount of energy due to the diffraction of the sound waves. However, the backside of cylinder has the most attenuation of the acoustic energy due to the presence of solid cylinders. The simulation results show that the phenomenon remains valid even with longer distances.



(a) d = 30cm (b) d = 60cm (c) d = 90cm (d) d = 120cm (e) d = 150cm (f) d = 180cm

Fig. 5. **Ultrasound (25 kHz) field simulation of acoustic attenuation**. The top views show sound field distribution when incident waves hit a cylinder. The distance between the simulated sound source and the microphone ranges from 30 cm to 180 cm. The size of speaker is 35 cm × 35 cm, the height and diameter of cylinder are 16 and 10 cm.



(a) d = 30cm (b) d = 60cm (c) d = 90cm (d) d = 120cm (e) d = 150cm (f) d = 180cm

Fig. 6. **Audible sound (500 Hz) field simulation of acoustic attenuation**. Compared with Fig. 5, we can see that the acoustic field distribution are more uniform than that of ultrasound.

Fig. 6 shows the results when the speaker emits 500 Hz sound waves. Unlike the high-frequency (25 kHz) simulation results, now the energy is all around the speaker, not only in the propagation direction. In addition, the simulation results suggest that the field distribution around the microphone array is close to uniform with the attenuation of propagation.

In summary, the simulation results verify that the energy received by microphones in each channel are significantly different due to the attenuation and diffraction of sound wave when the frequency of incident wave is 25 kHz, but it is uniform in the case of low-frequency source.

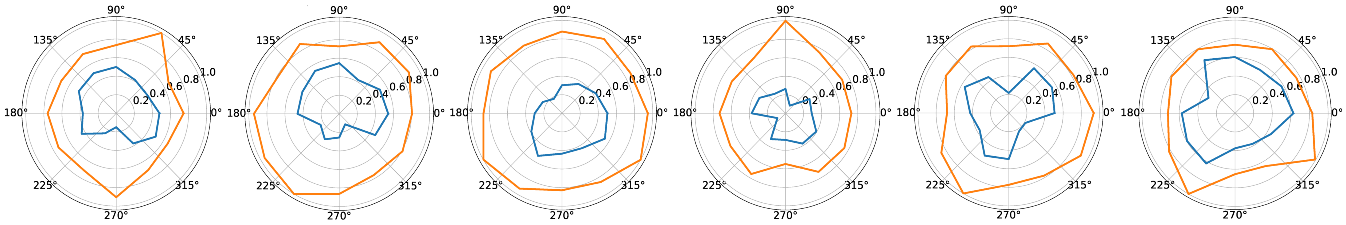
C. Verification of Acoustic Attenuation

To further verify the sound field distribution around the microphone array, we use the microphone array to pick up acoustic signals after playing the sound and inaudible sound (Modulated sound), and then show the field differences using the variance of band power of the five received signals. We now report the experiment setup and the results at different distances.

1) Experiment Setup: During the experiments, we use an iPhone x smartphone to generate a 500 Hz single tone and played by a portable mini Bluetooth speaker JBL GO [24]. The inaudible sound is an AM modulated signal and generated by the signal generator and will be played by a transmitter array. The transmitter array is designed with 40 ultrasonic transducers in parallel. The frequency of the baseband signal is also 500 Hz, the carrier frequency f_c is set to 25 kHz. The distance between the center of the microphone array and the sound source range between 30 cm and 180 cm, the step is 30 cm. To fully evaluate the influence of the position of sound source, the sound source will be rotated around the z-axis, as Fig. 4 shows, θ is set to 90 degree, φ changes from 0 to 360 degree, the step is 30 degree. A photo of the experimental setup is shown in Fig. 10.

2) Results: The results are shown in Fig. 7, where we use variance of band power to represent the uniformity of sound field. That is, to calculate the variance of the signals received by the five microphones at each degree, as described above in the experiment setup. If there is an acoustic signal and its sound field is uniform, then the energy of the five channels would be very close. If the signal is modulated by a high-frequency carrier wave, because the carrier has an effect of directionality and poor diffraction, some microphones have strong energy, while other channels have very weak energy. As a result, we see a large variance.

As shown in Fig. 7, these curves correspond to a speaker changing from 0° to 360° with a step of 30° on the x-y plane and launching baseband signals directly or modulated the baseband on carrier signals. The variance curve of audible signals ranges from 0.17 to 0.5. However, as for AM modulated signals, we can see the energy variance ranges from 0.65 to 1 when the distance is 30 cm. With the distance between the speaker and the microphone increases, the variation range of variance curve is between 0.15 and 0.62, for high-frequency



(a) d = 30cm (b) d = 60cm (c) d = 90cm (d) d = 120cm (e) d = 150cm (f) d = 180cm

Fig. 7. Verification of acoustic attenuation. Under the setting of Fig. 4, acoustic attenuation is calculated as variance of the band power of sound signals from five microphones under various settings, e.g., distance between sound source and the voice assistant, angles, and carrier frequency of ultrasound. The attack is launched at different distances range from 30 cm to 180 cm. Clear difference of ultrasound and audible sound has been shown as contrast.

sound, the variation ranges from 0.57 to 1. The difference between acoustic signal and high-frequency modulated signal is significant.

In summary, we observe that the diffraction attenuation increases with acoustic frequency. As the acoustic frequency increases, the sound field becomes significantly spatial uniform. Thus, the difference of high-frequency ultrasound received by the five microphones is higher than low-frequency sound.

IV. Design of EarArray

In this section, we introduce the design of EarArray to detect inaudible voice commands based on acoustic attenuation.

A. Overview

Fig. 8 shows the overview of the system architecture. The voice commands are first captured by the built-in microphone array on a device, e.g., the Echo dot and the audio signals are then fed into our EarArray system. Finally, the EarArray system will output the detection result i.e., whether the command is a DolphinAttack signal or from a human user. To achieve the above purpose, we have designed EarArray and it mainly consists of three major components, which are 1) Audio signal preprocessing, 2) Feature extraction and 3) Attack detection & localization.

B. Audio Signal Preprocessing

The audio signal preprocessing module is used to filter the noise in the input signals from multiple microphones and then prepare audio samples with a specific length for the feature extraction module.

Signal denoising. Due to sound interference from the environment, the signals from microphones are noisy. To improve the SNR of the signal, we exploit a band-pass filter to get rid of interference from unwanted frequencies. Considering the fact that the typical frequency of human sound is from 50 Hz to 2 kHz, we set the cut-off frequencies of the band-pass filter as 50 Hz and 2 kHz respectively in our design.

Voice activity detection and segmentation. The sound signal after microphones are a sequence of speech signals interleaved with non-speech signals. To further improve the quality of the sound signal, we choose to abandon the non-speech signal intervals. To do this, we first detect non-speech signal intervals by exploiting the voice activity detection (VAD) algorithm [19]. VAD is a common method of detecting the presence or absence of speech in sound signals. To detect and delete non-speech intervals, a band-power-based detection algorithm is used. To calculate band power of each channel, we first compute power spectral density (PSD) based on Welch's method which reduces noise in the estimated power spectra and then compute the band power in the given frequency range. The equation of VAD can be expressed as:

$$y(t) = \begin{cases} y_n(t), p_n > thres \\ Non - speech, p_n \leq thres \end{cases} \tag{8}$$

Where, $y(t)$ denotes the voice signal after VAD, $y_n(t)$ denotes the nth segment, p_n denotes the band power of the nth segment. $thre$ can be expressed as:

$$thres = \lambda_1 * max(p) + \lambda_2 * min(p), p = p_1, p_2, ...p_n \tag{9}$$

Where, λ_1 and λ_2 can be set to 0.04 and 3.

Specifically, we divide a whole sound signal into several segmentations with a step of 400 ms and the overlap of each frame is set to 200 ms. For each signal segmentation, we calculate its power of specific frequency band (50–2000 Hz) and discard those whose band power lower than a threshold.

By doing this, the non-speech signal segmentations can be removed and only speech-related signal segmentations are kept, the process is shown in Fig. 9. And note that, the VAD algorithm is applied on the channel with the highest band power since all channels are almost synchronized, the non-speech signal of the other channels can be abandoned according to the highest band power channel.

C. Feature Extraction

For the segmentations from the audio signals, we investigate the features representing the spatial inhomogeneity of sound. The preliminary analysis using band power variance computed by five channel signals indicate that the sound characteristics generated by pure-tone signal can be clearly distinguished from pure-tone AM signals. In the next, we calculate three representative features, the feature extraction process depicted in Fig. 9.

Range and standard deviation of band power. As the speech signal is a narrow frequency bandwidth signal, we use the frequency band power to indicate the sound intensity on five channels. The range of band power can be expressed as:

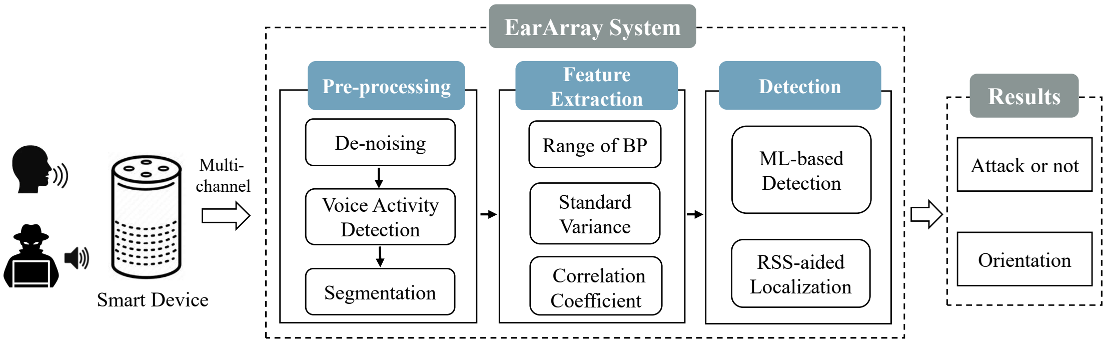


Fig. 8. The workflow of EarArray. The audio signals are first captured from multiple microphones on a device and then fed into the detection component which includes pre-processing, feature extraction, and attack detection. Attack detection results including attack source orientation will be output.

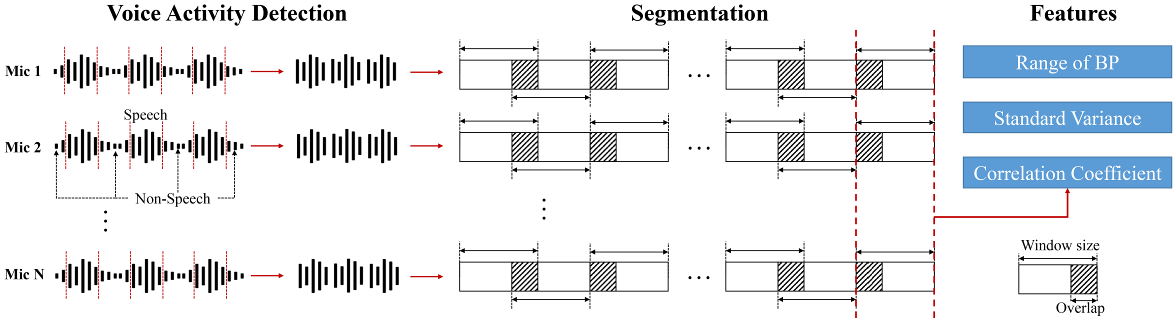


Fig. 9. Pre-processing workflow. VAD algorithm is used to detect the speech and non-speech segments, and then the speech segments are concatenated. Our window size is 0.4 s, and the overlap between each window is set to 0.13 s. The features can be obtained by calculating the energy of each speech window in the same column as the Figure depicts.

$$range = max(P) - min(P), P = \{p_1, p_2...p_m\} \tag{10}$$

Where, p_m denotes the band power of the mth channel. In the same way, we can also get the standard deviation of band power std.

Pearson Correlation Coefficient. Besides $range, std$, we use the Pearson correlation coefficient $corr$ between two spectra to estimate the uniform of the sound signals instead of using time-domain signals, that is because the phase difference between any two channels' signal will affect the Pearson correlation coefficient. In the frequency domain, the phase difference can be eliminated. As we have 5 channel signals, we choose the pair of signals with the biggest difference in energy to get $corr$.

Finally, we obtain 3 different features $range, std, corr$ to represent the uniform of measured sound field. To show the feasibility of using the 3 features to detect inaudible voice commands, we calculate the features of inaudible voice commands and normal voice commands and show the results in Fig. 11, from which we can find that the two different types of features are distinguishable and the gap between them is obvious. Thus, these features can represent the difference between sound signal distribution.

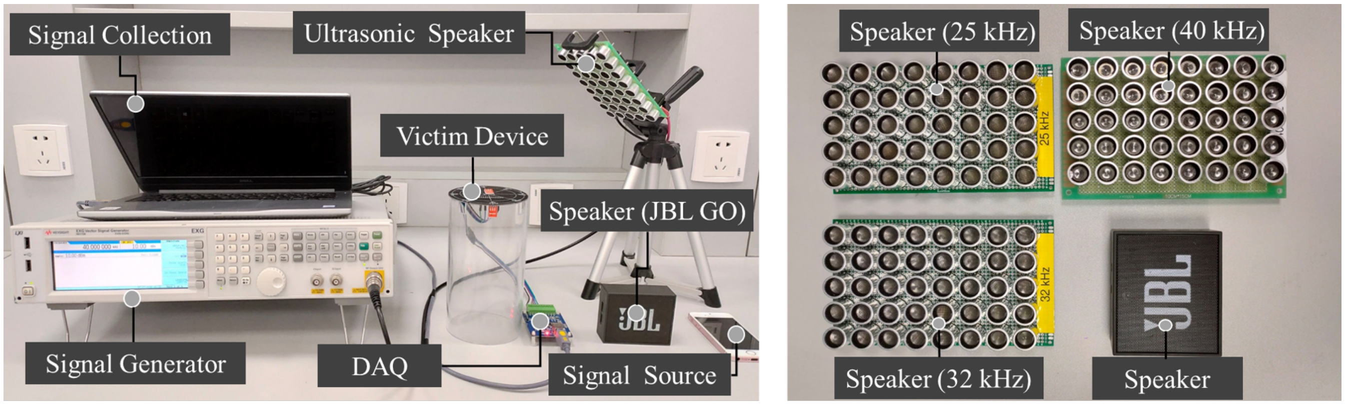
D. Attack Detection and Localization

1) Attack Detection: EarArray utilizes a machine learning (ML) based method to detect the inaudible voice commands with the above features. We choose Support Vector Machine (SVM) as the classifier in our design considering its simplicity and low-cost in terms of computation. We collect multi-channel voice segmentations in the off-line training phase as training samples, the voice samples include two types of voice commands: 1) Inaudible voice commands with different carrier frequencies, e.g., 25 kHz, 40 kHz, etc.,; 2) Audible voice commands. Both of the samples are collected by the specially-designed microphone-array device, as shown in Fig. 10. The traces are collected at different locations around the sound source. After training with these samples, the characteristics of the two sound signals will be registered in the trained model.

In the process of the detection phase, features of the unknown label voice samples will be calculated and finally classified according to the trained model.

2) Orientation Localization: After detecting there is an inaudible voice command attack, EarArray can also report the orientation of the attacker. Almost all of the popular smart speakers support sound source localization based on the TDoA algorithm, this kind algorithm can work well in sound source localization when the signal-to-ratio (SNR) is high. However, the signal in some channels is too weak to apply the TDoA algorithm effectively as the serious attenuation of high-frequency inaudible voice commands. Thus, the performance of the method will dramatically decrease under attack.

To overcome the above localization challenge facing an inaudible voice command attack, we propose a band-power-based localization method for each microphone channel to



(a) Experimental setup

(b) Transducer Array & JBL GO

Fig. 10. The experimental setup. Three self-made ultrasonic speaker arrays with center frequencies of 25, 32, 40 kHz. A hardware signal modulator is used.

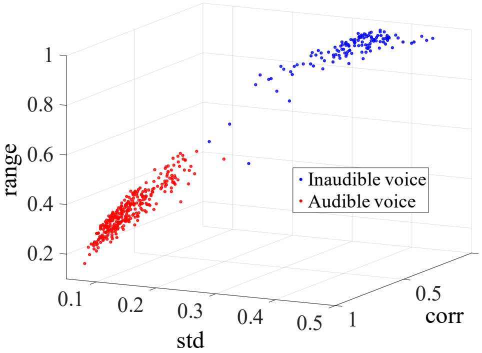
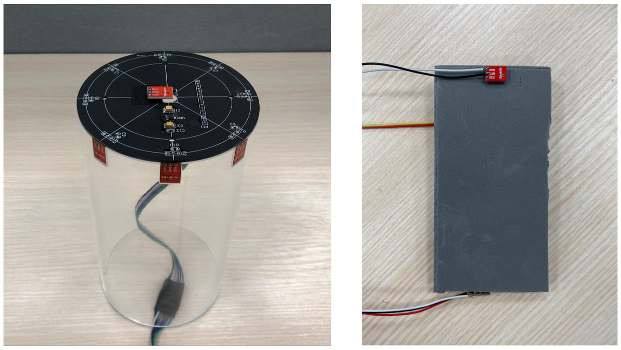


Fig. 11. The feasibility to detect inaudible attacks with the three features which including range, standard deviation (std), and Pearson Correlation Coefficient (corr) of band power



(a)

(b)

Fig. 12. (a) The specially-designed prototype of a microphone array which supports 9 channels, and we use 5 microphones in this paper; (b) The prototype of a smartphone with three microphones.

infer the direction of the attacker. As the band power in each channel is sensitive to distance and obstacles, the microphone facing toward sound source has the strongest signal strength and vice versa. Thus we regard the attacker is from the direction where the microphone has the strongest power. As a fact, the attacker orientation can be located in a quite large angle because we cannot have a very fine-grained signal power estimation. Thus, in `EarArray` the localization module only outputs a coarse direction, namely, “North”, “South”, “West” or “East” and the user can further identify the location of the attacker by looking at the structure of the house like windows and walls. We evaluate the performance of such a coarse localization in the Sec. V.

V. Implementation and Evaluation

In this section, we start with the introduction of our specially-designed prototype of a microphone array to better evaluate the performance of `EarArray` and then elaborate on the evaluation.

A. Implementation

To verify the performance of `EarArray`, we design a special 3-D microphone array as we don’t have permission to record multi-channel sound on popular smart devices. Imagine that as the acoustic wave encounters the smart speaker, the backside will produce a shadow region, and with increasing sound frequency, the shadow will become more significant, and this kind of spatial property of the sound field will be completely captured by the 3-D microphone array. A data acquisition card is used to collect five channel signals for subsequent analysis. To play the inaudible voice commands, we design 3 ultrasound transducer arrays with center frequencies of 25, 32, 40 kHz which are shown in Fig. 10(b).

As we can see from Fig. 12(a), the height of the cylinder body of the microphone array is 15 cm and its diameter is 10 cm, the distance between the sound inlet of each microphone and the top edge is set to 1.2 cm. Five ADMP401 MEMS microphones [23] were used in our prototype. The microphone array can be applied to fully measure the sound field instead of all microphones are located on the top plane. In particular, we use one microphone located on the center of the top surface of the cylinder, which is like the Echo microphone placement in Fig. 3. The other four microphones are uniformly located around the curve of the top surface. To evaluate the performance of `EarArray` on commercial smartphones, we also design and implement a prototype with only 3 microphones, whose size is 15 × 7.5 × 1cm. The three microphones are located on the top, back surface, and bottom side respectively,

TABLE II. THE LIST OF VOICE COMMANDS USED IN THE EXPERIMENT.

Speaker	Voice command
TTS & Volunteers *	How is the weather today Turn on airplane mode Call 1 2 3 4 5 Facetime 6 7 8 9 0 Read my new messages

* In the experiment, we totally recruited 25 volunteers (5 females and 20 males, aged between 20 and 29)

as shown in Fig. 12(b).

Using the specially designed smart speaker, we conducted experiments to evaluate the effectiveness of EarArray in terms of various factors, including carrier frequencies, attack distances, angles, background noise, and voice commands types. We also use the smartphone prototype to evaluate the performance of EarArray on existing hardware with three microphones. According to these experiments, we demonstrate that EarArray can detect the inaudible voice commands with accuracy of above 99%, meanwhile, the accuracy of localization is up to 97.89%. We summarize the main results as follows:

- EarArray shows the detection accuracy can be up to 99% in various conditions and positions.
- EarArray can achieve localization accuracy as high as 97.89%.
- EarArray is robust in terms of attack parameters, i.e., attack distance, ambient noise, the angle between a smart speaker and the attacker, etc.

B. Experimental Settings

Hardware setup. The experimental setup is shown in Fig. 4 and Fig. 10. The benchtop transmitter is used for modulating the voice commands played by a smartphone and then emitting the inaudible voice commands with 3 narrow bandwidth ultrasonic transducer arrays (The center frequencies are 25, 32, 40 kHz respectively). The low-frequency audible voice commands will be played by Bluetooth speaker JBL GO controlled by an iPhone X. In our experiments, we use the designed microphone array as the victim device. And the positions of ultrasonic speaker and JBL GO are controlled by φ, θ, and R as shown in Fig. 4.

Voice commands. We recruited 25 volunteers including 5 females and 20 males whose ages range between 20 to 29. The volunteers were required to speak the 5 voice commands shown in Tab. II. The whole process will be recorded by iPhone 6Plus, Galaxy S6, OPPO Reno2, OPPO Reno3 and then we collect 500 voice samples. We also use Google Text-to-Speech (TTS) engine to generate the five voice commands, which will be used in the following experiments. We process the sound samples offline in Python 3.7.

Environments. All the experiments are conducted in an office with ambient noise of about 55 dB SPL except the experiments which explore the impact of background noise. The transmitting power of speaker is limited to 1 Watt.

Metrics. To evaluate the performance of the detection system proposed in this paper, the following experiments are conducted using 5 metrics. Accuracy: The rate that correctly identifying legitimate and illegitimate voice commands, true negative rate (TNR), true positive rate (TPR), precision, and recall.

C. Overall Performance

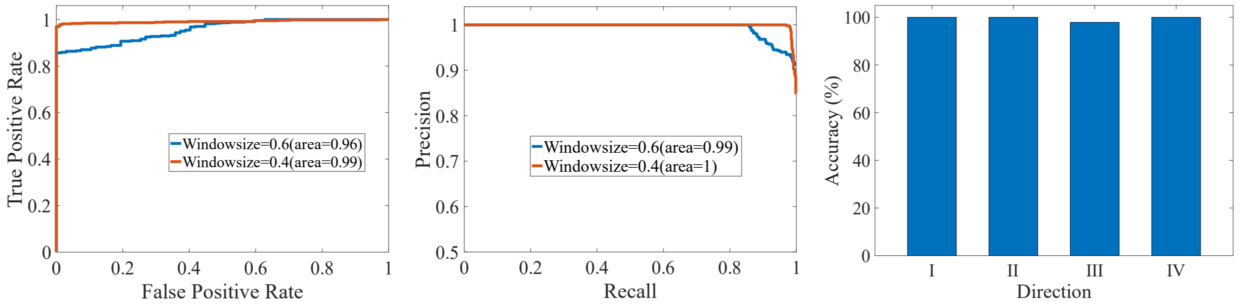
1) Detection Accuracy: As we can see from Fig. 4, the attacker could launch an attack at anywhere around the smart speaker, to explore the detection performance from any spatial location and given carrier frequencies (25, 30, 40 kHz), we play the five audible voice commands and five inaudible voice commands with ultrasonic transducer arrays and JBL GO at positions controlled by φ, θ, and R. We first rotate speakers around the diameter of the top surface (x-axis), that is, φ is set to 0 degree, θ changes from -120 to 120 degree with a step of 30 degree. The distance R is set to 30 cm and 60 cm respectively. Secondly, the sound source will be rotated around the z-axis, θ is set to 0 degree, φ changes from 0 to 360 degree, the step is 30 degree.

To show the overall performance of EarArray, we calculate the TPR, FPR, precision, and recall using all the recorded samples, and plot the ROC curves and Precision-Recall (R-P) curves as shown in Fig. 13(a)(b), from which we can observe that EarArray successfully detects inaudible voice commands with high reliability. The area under ROC curve (AUC) can be up to 100% when the window size is fixed to 0.4 s, when the window size is 0.6 s, the AUC is also up to 96%. And the areas under P-R curves are above 99%.

2) Localization Accuracy: To investigate the accuracy of localization for the inaudible voice commands, we choose the audio samples recorded from microphones of different directions, that is φ ranges from 0 to 360 degree with step size of 30 degree, the attack distance is set to 60 cm. For each direction, the experiment repeats three times, and we calculate and average the accuracy of localization, the direction I represent when φ ranges from 0 to 90 degree. The results of localization are depicted in Fig. 13(c), from which we can observe that the accuracy of localization is 100, 100, 97.89, 100%. In the process of experiments, we find that the frequency response of microphone M3 is lower than the others which will influence the received signal strength.

D. Impact of Distances

With the increase of distance, the sound field scattered from the speaker will change accordingly, especially when the to explore whether the variation of distance affects the detection performance, the distance between sound source and smart speaker is set to 30, 60, 100, 200, 300 cm. As the effective attacking distance is hardware-dependent, when the distance is 3 m, the SPL of received inaudible voice command is weak which can't achieve a successful attack, thus, the maximum distance is set to 300 cm. The results are shown in Fig. 14 (a) confirm that with increase of distance, the TPR and TNR don't have obvious change. On the whole, even the SNR is getting worse as the increase of distance, the performance of EarArray doesn't distance effect.

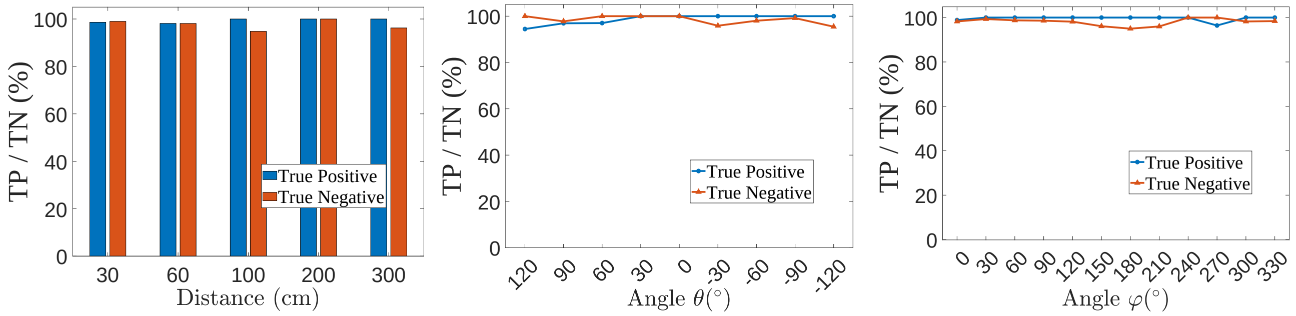


(a) Overall ROC curves.

(b) Overall Precision-Recall curves.

(c) The impact of the angle on the performance of localization.

Fig. 13. The overall performance of `EarArray` with attack detection and localization. The window size is the length of each voice sample.



(a) The impact of distance.

(b) The impact of angle, $\varphi = 0$.

(c) The impact of angle, $\theta = 0$.

Fig. 14. (a) The impact of distance which ranges from 0.3 to 3 m; (b) The impact of angle φ which ranges from 0 to 120 degree; (c) The impact of angle θ which ranges from -90 to 90 degree.

E. Impact of Angle.

The incident angle determines the position relationship between sound source and five microphones, which directly influence the SPL distribution at each microphone. And the attacker can launch an attack in any concealed or most effective position. To explore the impact of incident angle on the performance of `EarArray`, we do the following experiments at a distance of 60 cm with different angles θ, φ.

Impact of θ. To evaluate the influence of θ on `EarArray`, φ is set to 0, and θ is changing from -120 to 120 degree (θ is 90 degree when the sound source is paralleling to x-axis positive direction), the step size is 30 degree. With increasing of the absolute value of θ, the sound source will close to the ground, thus, the maximum absolute value of θ is set to 120 degree. The results are shown in Fig. 14(b), as the value of θ gradually increases from -120 to 120 degree, the values of TPR and TNR are not obvious change and remain close to 100%, which indicates the performance of `EarArray` doesn't affect by angle.

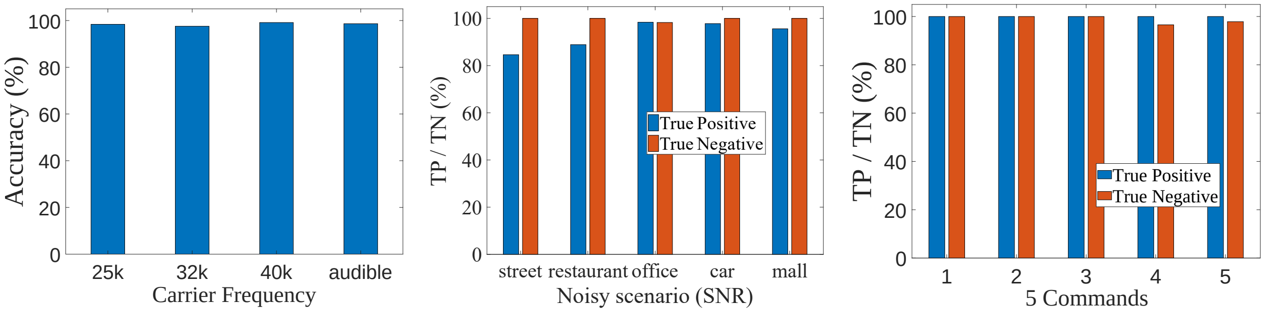
Impact of φ. In this experiment, we fixed θ at 0 degree and change φ from 0 to 330 degree (φ is 90 degree when the sound source is paralleling to y-axis positive direction), the step size is 30 degree. The experimental results plot in Fig. 14(c), from the figure, we can observe that the performance of the `EarArray` does not change significantly with the changing of φ, when φ is 120 and 210 degree, the values of TNR are slightly lower than the average value. That is because the frequency response of microphone (M3) is different from the others which will influence the reality of measured sound field distribution. In future work, dynamic gain control (DGC) strategy should be applied to eliminate the interference of differences in microphones.

F. Influence of Carrier Frequencies

Carrier frequency is a dominant factor that affects the attack success rate, and it also shows great variance across devices [14]. For `EarArray`, the carrier frequency is also an important parameter that will directly influence the distribution of the sound field and then affects the detection performance. To investigate the effect of the carrier frequency, we conduct the following experiments. We modulate the five voice commands on 25, 32, 40 kHz carries respectively, and launch the attack at a distance of 60 cm. We repeated the experiment 3 times and finally calculate the detection results as shown in Fig. 15(a). From which we can observe that the maximum accuracy is 99.14% at the carrier frequency of 40 kHz, which reflects the sound field distribution of high-frequency sound is more uneven, thus, it's easier to detect.

G. Impact of Ambient Noise

Background noise will not only affect the recognition rate of the speech recognition system but also change the distribution of the spatial sound field. To evaluate the impact of ambient noises, we simulate five scenarios by playing accordingly audio with a given range of SPL in our office, note that, the SPL of noise should be measured near the victim device. The SPL of street, restaurant, office, car, and shopping mall is 75–85, 65–75, 55–65, 60–70, 60–75 dB respectively. The detection results are shown in Fig. 15(b), from which we can find that the TPR will slightly decrease with the increase of ambient noise, but the TNR does not affect by noise. That is because the SPL of ambient noise is higher

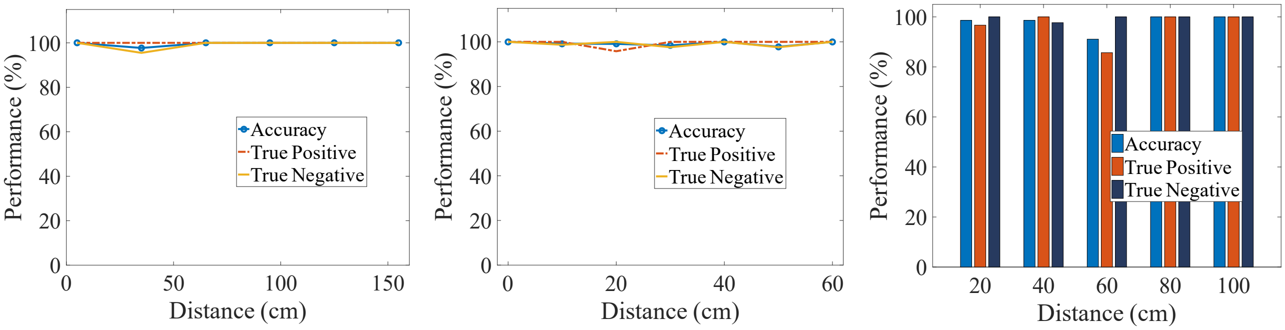


(a) The impact of carrier frequency.

(b) The impact of ambient noise.

(c) The impact of voice commands.

Fig. 15. (a) The carrier frequencies are 25, 32, 40 kHz respectively; (b) The impact of background noises. We simulate the five scenarios by playing background sounds at chosen SPLs; (c) We use five different commonly used voice commands which are listed in Tab. II.



(a) The impact of obstacle between the attaker and victim device.

(b) The impact of wall reflection.

(c) The impact of multiple attack speakers.

Fig. 16. The performance of EarArray on three possible evasion techniques.

than audible voice commands, and the ambient noise includes some non-speech high-frequency components which lead to an uneven sound field and misclassification. On the whole, the experimental results show that EarArray is robust to different noise sources.

H. Impact of Different Voice Commands

To examine the effectiveness of EarArray with regard to voice commands, we select five types of voice commands which are list in Tab. II. We then play the 500 voice samples using JBL Go and collect the audible signal, to collect the inaudible signal, we modulate the 500 voice samples on 40 kHz carries, and launch the attack at a distance of 60 cm. Fig. 15(c) shows the impact of voice commands on TPR and TNR. As we can see that the TPR and TNR of various voice commands are almost the same and range from 98.4% to 100%. The results illustrated that different types of voice commands would not have an obvious impact on EarArray.

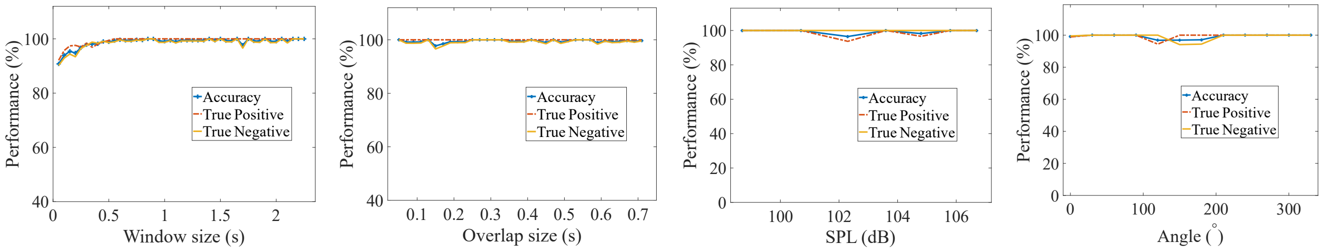
I. Evasion Techniques

To show the effectiveness of EarArray when the adversary knows of the defense method and tries to circumvent the protection, we consider three evasion strategies: 1) the attacker is hiding behind an obstacle; 2) the victim device is close to walls; 3) multiple speaker injection.

1) Hiding Behind Obstacles: In this experiment, we choose a cylinder as the obstacle whose size is the same as that of the smart speaker, and the distance between the obstacle and the smart speaker is fixed to 40 cm while the distance between the transducer array and obstacle is changing from 5 cm to 155 cm with a 30-cm step. Note that if the obstacle is much bigger than the transducer array (9×15 cm), it will block the attack signal when the transducer array is close to the obstacle. From Fig. 16(a), we can see that only when the distance between the obstacle and transducer array is 35 cm, the detection accuracy is decreased to 97.7%, the accuracy of EarArray at other distances can be up to 100%. Thus, the obstacle would not reduce the effectiveness of EarArray.

2) Influence of Surrounding Walls: When the smart speaker is close to walls, acoustic reflection from walls will influence the distribution of the sound field, which might be leveraged by adversaries. To evaluate the effectiveness with regard to surroundings, we conduct the following experiments, the smart speaker placed equidistant from two walls that are at a right angle to each other. The distance between the walls and the smart speaker ranges from 0 cm to 60 cm, the distance between the transducer array and the smart speaker is fixed at 60 cm. The results are depicted in Fig. 16(b), from which we can observe that the accuracy changes slightly with distances, which remains within a certain range from 97.8% to 100%. Thus, EarArray is resistant to the influence of surrounding object and the defense effectively works.

3) Multiple Speaker Injection: To disrupt the uneven distribution of the sound field and escape detection, the adversary might use two transducer arrays around the smart speaker. In this experiment, we use two same transducer arrays to evaluate the effectiveness of EarArray. The two transducer arrays (speaker 1 and speaker 2) face toward the smart speaker, simultaneously play the same inaudible voice commands, and the three devices are in the same line. The distance between speaker 1 and smart speaker ($d1$) is set to 60 cm, the distance



(a) The influence of window size. (b) The influence of overlap time. (c) The influence of SPLs. (d) The detection performance on a smartphone with three microphones.

Fig. 17. The impact of window size, overlap size, SPLs, and smartphone on the performance of `EarArray`.

between speaker 2 and smart speaker ($d2$) is changing from 20 cm to 100 cm, the step is 20 cm. Experimental results as shown in Fig. 16(c) demonstrate that when the value of $d2$ is equal to $d1$, the accuracy EarArray is 91.1% and the TPR is 85.7%, this result means that the two speakers could make the sound field more uniform when the SPL of reached signals from two speakers are the same, and will, in a manner, reduce TPR of EarArray. To improve the performance of detecting multi-speaker injection, a more sophisticated structure of the 3-D microphone array should be studied, for example, the diaphragm of microphone doesn't face toward the sound inlet, thus, the inaudible signal will greatly attenuate before propagating into the diaphragm. We leave this to future work.

J. *Influnce of Windowsize and Overlap Size*

To investigate the impact of various window sizes on EarArray, we make the window size changes from 0.05 to 2.25 s, the overlap size t_o is set to $t_w/2$, and calculate the accuracy, TPR, TNR respectively. Fig. 17(a) depicts how the window size influences the performance of EarArray. As the window size increases, the accuracy/TPR/TNR gradually increases, and when the window size is above 0.4 s, the values are maintained above 99% and slightly fluctuate.

Usually, a longer overlap size t_o brings more correlation between adjacent voice samples, however, with given total voice signal for training or testing, the processed data will be repeated and decrease the detection efficiency. To find a balance between detection performance and efficiency, we fixed the window size to 1 s and change the overlap size from 0.05 s to 0.7 s. Fig. 17(b) shows the performance of detection under different overlap sizes. We can observe that as the overlap size bigger than 0.17 s, the accuracy keeps at a higher range from 98.9% to 100%. In general, to obtains a good balance between efficiency and performance, the value of t_o in the range from $t_w/4$ to $t_w/2$ is suitable.

K. *Impact of Sound Pressure Level*

To explore the impact of SPLs on EarArray, we lower the SPL of inaudible voice commands to the minimum value of 98.7 dB which still can be recognized by a smart device [5]. The SPLs were measured by a measurement microphone [7] which is placed next to the smart speaker. Fig. 17(c) shows the impact of the SPLs on the accuracy, TPR, and TNR. Although a lower SPL always means a smaller signal-to-noise ratio (SNR) for given noise levels, the accuracy of EarArray still can be up to 100% when the SPL is 98.7 dB. With the increases of SPLs, the accuracy decreases to 96.5% and then increases to 100%, which suggests that EarArray is effective even with a lower pressure level of the inaudible signal.

L. *Smartphone with Three Microphones*

In this section, we use the smartphone prototype to evaluate the effectiveness of EarArray on smart devices whose microphone array design is similar to actual smartphones. We play audible and inaudible voice commands 90 cm away from the prototype which was placed on a table, the φ ranges from 0 to 330 degree, the step is 30 degree. The experimental results plot in Fig. 17(d), and indicate that the accuracy was above 96.8% and was slightly affected by angles especially when the angle is between 120 and 150 degree. We think that is because the difference between the acoustic attenuation of the three channels is the least when the speaker at the three positions, and in a manner, reduce the accuracy. In summary, EarArray also has a good performance on the smartphone when the number of microphones decreases to three. In fact, as the number of microphones increases and those microphones were placed around the surfaces of smart device and facing different directions, the acoustic attenuation will be measured more sufficiently, thus, the performance of EarArray will be better.

VI. Related Work

As Voice Control Systems (VCSs) are playing a more and more important role in our daily life, cyber attacks against VCSs start to draw people's attention. Regarding the concealment feature of the attacks, current VCS attacks can be divided into two types: audible and inaudible attacks. Inaudible attacks, the attacker tries to approach the victim and plays recorded audio to the victim device. The audio is usually generated in a tricky way such that it is incomprehensible to humans while comprehensible to the device. The audible attacks are theoretically feasible but have limitations in practice due to the fact that the audio used for attack is usually distinguishable from white noise. In inaudible attacks, the attacker modulates malicious voice commands on ultrasonic carriers, so the commands become inaudible to human ears, but still receivable to microphones on VCSs duce to nonlinear effects of microphone circuits. Inaudible attacks are a totally imperceptible attack.

A. *Audible Attack on VCSs*

Generative adversarial networks (GANs) are originally proposed for image recognition missions, due to their great performance researchers are inspired to explore the feasibility of using GANs in speech generation. Existing works have proposed to use GANs for generating incomprehensible audio commands to attack speech recognition systems [25], [28], these malicious audio is disguised as some random noise so the victim won't be able to notice the attack. This work [26] proposed a approach to deceive human's ears by hiding an executable voice command in adversarial audio which appears to be simple music to human. Nicholas et al. [10], [11] demonstrated that the targeted audio adversarial examples by modifying existing audio can be generated, which introduce a new domain to study adversarial examples. Based on the above work, Metamorph achieves the over-the-air attack.

As speech biometrics recognition gradually takes place of traditional identification technologies [29], the newly developed automatic speaker verification (ASV) systems can defend traditional VCS attacks to some extent. However, the replay attack still poses a great threat to these ASV systems, it can bypass the verification by simply replaying a pre-recorded audio. Villalba et al. [30] presented many vulnerabilities of ASV systems towards far-field replayed attacks. In the anti-spoofing competition ASV2017 [31], Witkowski, et al. [32] pointed out that replay attacks can be detected by analyzing the high-frequency band of the replayed recordings. Zeyan et al. [33] improved the discriminating ability of the relative phase (RP) features by proposing two new auditory filter-based RP features for replay attack detection. To detect the remote attaker, Lee et al. [12] proposed a sonar-based liveness detection system. Speaker-Sonar emits an inaudible sound and tracks the user's direction and to compare it with the direction of the received voice command. If the inaudible attack is launched by a nearby and moving attacker, the sonar-based method will fail.

B. Inaudible Attack on VCSs

Kasmi et al. [34] introduced a kind of new voice command injection into modern smartphones using intentional electromagnetic interference with headphone cables. The limitation is that the attack devices must be plugged into a smartphone. DolphinAttack [14], [5], [8] translated typical audible voice commands into ultrasonic frequencies making it inaudible to the human ear, but still decipherable by the microphones and the always-on voice assistants. [14], [5] also proposed a defense strategy from the software level based on audio feature extraction. However, the signal features vary significantly along with the different types of microphones. The defense method in [35]depends on the nonlinearity of microphone circuits, which is different from what we proposed, i.e., we utilize the prorogation characteristic of sound in the air. We have validated experimentally that nonlinearity is device-dependent. For instance, amplitude skewness, one of the main features used for nonlinearity, are valid with Sumsung Galaxy S6 Edge+ but are not valid for iPhone 4S and iPhoneSE. As mentioned in [36], it is possible to conceal all three proposed features related to nonlinearity. He et al. [36] designed a "guard" signal transmitter to detect and capture the attack signal, it is also capable to neutralize the attack signal. Light Commands [6] use light to inject commands into voice-controllable systems by aiming an amplitude-modulated light at the microphone's aperture. To detect light-based command injection, they attempt to detect the attack by comparing signals from multiple microphones or add a barrier film before the microphone's diaphragm to blocks straight light beams. As only one microphone receives a signal while the others receive nothing, `EarArray` also can be applied to detect the Light Commands. SurfingAttack [9] injects the inaudible signal using ultrasound propagation in solid media, by utilizing the nonlinearity, the ultrasound signal can be demodulated and recognized by the speech recognition system. By monitoring the frequency component in high frequency range, SurfingAttack can be detected. Different from their method, EarArray utilizes the difference in propagation attenuation between ultrasound and sound to detect attacks. UltraComm [4] proposes an approach for acoustic communication which different from `EarArray`. It leverages nonlinearity effect of microphones and transmit modulated data on frequency above 20 kHz and recovers it in the audible frequency band.

Our work in this paper shows the distinguishing field features between the acoustic signals and high-frequency modulated signals. We analyzed different attack scenarios of smart voice assistant and smartphones and implemented the defense mechanism based on the analysis of sound field features. In contrast to the defense method mentioned in [35], [36], our defense algorithm is more efficient and it is instructive for the design of microphone array in the future.

VII. Conclusion and Future Work

Voice assistants brought convenience to our daily life, however, they have also exposed our privacy to a certain type of various malicious attacks using inaudible voice commands. In this paper, we proposed a light-weight mechanism named `EarArray` to defend voice assistants against these inaudible voice command attacks. We theoretically analyzed the attenuation property of audible voice command and inaudible voice command and proposed to use sound field distribution as features to tell apart normal commands initiated by human beings and inaudible command generated by machines. We have conducted plenty of experiments to prove the feasibility of `EarArray`, results show that `EarArray` can achieve 99% accuracy for attack detection, and 97.89% localization accuracy for inaudible voice commands.

Our future work includes overcoming the problem that mobile phones cannot read multi-channel data at the same time, extracting field patterns from more different types of voice assistant devices, and hopefully generalizing our defense algorithm for all voice assistant devices.

Acknowledgment

We would like to thank our reviewers for their valuable comments, which significantly improved our paper. Thanks Bo Yang, Weizhen He, Dingkui Liu, and Ruihao Mao for their help in the simulation and experiment. This work is supported by China NSFC Grant 61925109, 61941120, 62071428, and ZJNSF Grant LGG19F020020. Gang Qu's work was done when he visited Zhejiang University as a Qiushi Chair Professor.

References

[1] statista. Number of digital voice assistants in use worldwide from 2019 to 2023.https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/?tdsourcetag=s_pcqq_aiomsg.20120, 2020.

[2] Maccà I, Scapellato ML, Carrieri M, Maso S, Trevisan A, Bartolucci GB. High-frequency hearing thresholds: effects of age, occupational ultrasound and noise exposure. Int Arch Occup Environ Health 88(2), 197–211. 2015.

[3] World Health Organization Ultrasound. Environmental Health Criteria 22. 1982.

[4] Zhang G, Ji X, et al. UltraComm: High-Speed and Inaudible Acoustic Communication[C]. Quality, Reliability, Security and Robustness in Heterogeneous Systems. 15th EAI International Conference. 2019.

[5] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang and W. Xu, "The Feasibility of Injecting Inaudible Voice Commands to Voice Assistants," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2019.2906165.

[6] Sugawara T, Cyr B, Rampazzi S, et al. Light commands: laser-based audio injection attacks on voice-controllable systems[C]//29th USENIX Security Symposium (USENIX Security 20). 2020: 2631-2648.

[7] GRAS, "Gras 46be 1/4" ccp free-field standard microphone set," https://www.grasacoustics.com/products/measurement-microphone-sets/product/143-46be

[8] Song L, Mittal P. POSTER: Inaudible voice commands[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 2583-2585.

[9] Yan Q, Liu K, Zhou Q, et al. SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves[C]//Network and Distributed Systems Security (NDSS) Symposium. 2020.

[10] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 1-7.

[11] Chen T, Shangguan L, Li Z, et al. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems[C]//Proceedings of the Network and Distributed Systems Security (NDSS) Symposium. 2020.

[12] Lee Y, Zhao Y, Zeng J, et al. Using Sonar for Liveness Detection to Protect Smart Speakers against Remote Attackers[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(1): 1-28.

[13] Jarrett D P, Habets E A P, Naylor P A. Theory and applications of spherical microphone array processing[M]. New York: Springer, 2017.

[14] Zhang G, Yan C, Ji X, et al. DolphinAttack: Inaudible voice commands[C]. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 103-117.

[15] Richard E. Berg. Sound Physics. https://www.britannica.com/science/sound-physics, 2020.

[16] He P. Simulation of ultrasound pulse propagation in lossy media obeying a frequency power law[J]. IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 1998, 45(1): 114-125.

[17] Apple. iPhone XS - Technical Specifications. https://support.apple.com/kb/SP779?locale=en_US, 2020.

[18] Apple. iPhone XR - Technical Specifications. https://www.apple.com, 2020.

[19] M. Grimm, K. Kroschel, Voice Activity Detection. Fundamentals and Speech Recognition System Robustness in Robust Speech Recognition and Understanding, Vienna, Austria: I-Tech., 2007, ch. 5, pp. 460.

[20] COMSOL. COMSOL. http://cn.comsol.com/. 2020.

[21] Amazon. Amazon Echo. https://www.amazon.com/all-new-Echo/dp/B07NFTVP7P?th=1. 2020.

[22] Cobos M, Lopez J J, Spors S. A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing[J]. EURASIP Journal on Advances in Signal Processing, 2010, 2010(1): 415840.

[23] Analog Devices. ADMP401: Omnidirectional microphone with bottom port and analog output obsolete data sheet. https://www.analog.com/media/en/technical-documentation/obsolete-data-sheets/ADMP401.pdf. 2011.

[24] Harman International Industries, "Jbl go," . https://www:jbl:com/JBL+GO:html. 2018.

[25] Tavish Vaidya. Cocaine noodles: exploiting the gap between human and machine speech recognition. Presented at WOOT 15 (2015), 10–11.

[26] Yuan X, Chen Y, Zhao Y, et al. Commandersong: A systematic approach for practical adversarial voice recognition[C]//27th USENIX Security Symposium (USENIX Security 18). 2018: 49-64.

[27] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 1-7.

[28] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands[C]//25th USENIX Security Symposium (USENIX Security 16). 2016: 513-530.

[29] Patil H A, Kamble M R. A survey on replay attack detection for automatic speaker verification (ASV) system[C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 1047-1053.

[30] Villalba J, Lleida E. Speaker verification performance degradation against spoofing and tampering attacks[C]//FALA workshop. 2010: 131-134.

[31] Kinnunen T, Sahidullah M, Delgado H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[J]. 2017.

[32] Witkowski M, Kacprzak S, Zelasko P, et al. Audio Replay Attack Detection Using High-Frequency Features[C]//Interspeech. 2017: 27-31.

[33] Oo Z, Wang L, Phapatanaburi K, et al. Replay attack detection with auditory filter-based relative phase features[J]. EURASIP journal on audio, speech, and music processing, 2019, 2019(1): 8.

[34] Kasmi C, Esteves J L. IEMI threats for information security: Remote command injection on modern smartphones[J]. IEEE Transactions on Electromagnetic Compatibility, 2015, 57(6): 1752-1755.

[35] Roy N, Shen S, Hassanieh H, et al. Inaudible voice commands: The long-range attack and defense[C]//15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). 2018: 547-560.

[36] He Y, Bian J, Tong X, et al. Canceling Inaudible Voice Commands Against Voice Control Systems[C]//The 25th Annual International Conference on Mobile Computing and Networking. 2019: 1-15.