MicGuard: A Comprehensive Detection System against Out-of-band Injection Attacks for Different Level Microphone-based Devices

Tiantian Liu, Feng Lin, Zhongjie Ba, Li Lu, Zhan Qin, and Kui Ren,
Zhejiang University and Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

https://www.usenix.org/conference/usenixsecurity24/presentation/liu-tiantian

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

MicGuard: A Comprehensive Detection System against Out-of-band Injection Attacks for Different Level Microphone-based Devices

Tiantian Liu , Feng Lin*, Zhongjie Ba , Li Lu , Zhan Qin , and Kui Ren

State Key Laboratory of Blockchain and Data Security, Zhejiang University
School of Cyber Science and Technology, Zhejiang University
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
*Corresponding author
Email: {tiantian, flin, zhongjieba, li.lu, qinzhan, kuiren}@zju.edu.cn

Abstract

The integration of microphones into sensors and systems, serving as input interfaces to intelligent applications and industrial manufacture, has raised public concerns regarding their input perception. Studies have uncovered the potential dangers posed by out-of-band injection attacks on microphones, encompassing ultrasound, laser, and electromagnetic attacks, injecting commands or interferences for malicious purposes. Despite existing efforts on defense against ultrasound injections, there is a critical gap in addressing the risks posed by other out-of-band injections. To bridge this gap, this paper proposes MicGuard, a comprehensive passive detection system against out-of-band attacks. Without relying on prior information from attacking and victim devices, MicGuard leverages carrier traces and spectral chaos observed by injection phenomena across different levels of devices. The carrier traces are used in a prejudgment to fast reject partial injected signals, and the following memory-based detection model to distinguish anomaly based on the quantified chaotic entropy extracted from publicly available audio datasets. MicGuard is evaluated on a wide range of microphone-based devices including sensors, recorders, smartphones, and tablets, achieving an average AUC of 98% with high robustness and universality.

1 Introduction

Microphones capture sound, one of the most important information mediums, to record auditory aspects of the physical world and open a door for intelligent machines to 'comprehend' human sayings. Not to mention telephoning and music recording, the integration of compact MEMS microphones into wireless and commercial devices have gained significant prominence in smart home, industrial applications, and live broadcasting. Microphones serve as vital input interfaces in various contexts, exemplified by their use in voice assistants-enabled smartphones for human-machine interaction [5], anomalous sound detection for fault diagnosis of

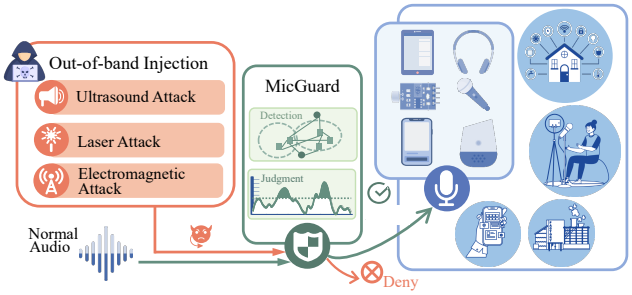


Figure 1: The designed MicGuard detects out-of-band injected signals and protects the microphone-based devices, promising the security of linked systems and applications

industrial processes [22], and television broadcasting for communication and recreation [12].

On the other side of the coin, the inherent open nature of microphones also unlocks a door for attackers to inject malicious *out-of-band* signals [13, 28, 30] for deception and sabotage. Surveying most injection attacks on microphones, out-of-band signals refer to signals that are outside the normal frequency range or transmission path of an audible pattern, such as ultrasound, laser, and electromagnetic (EM) signals. Given that these signals own high concealment and intrusiveness, particularly, attackers can perform ultrasound-based or laser-based attacks (i.e., dolphinattack [53], lightcommand [39]) to inject voice commands into voice-controlled devices. Additionally, they can also emit EM interference [23] to manipulate microphone recordings for further compromising linked broadcasting systems or sensor-fusion networks [38, 40].

However, prior works against injection attacks on microphones are limited to mitigating in-band injection (i.e., audible attacks [26, 47]) or further including ultrasound attacks [36], overlooking risks associated with the out-of-band injection. To fill this gap, we aim to design a comprehensive countermeasure against all mainstream out-of-band injection attacks. Compared with existing detection schemes, the desired detection facing out-of-band signals needs to meet the following requirements: 1) No extra hardware: the required sensors or hardware modification (e.g., speakers, IMU, mic arrays) in

some detection methods [18, 41, 52] bring about additional burdens for users and are unavailable for low-level mic-based devices [27]. 2) Independent of adversarial samples: most deep learning-based detecting mechanisms [15, 26] heavily rely on the quantity of training attacking samples. This is unfeasible to collect attacking samples that fully cover all possible attacking outcomes determined by victim conditions and attacking modes. 3) Low dataset cost: for ease of deployment, large amounts of positive data from the target microphone should not be required. 4) High accuracy and robustness: it should achieve high accuracy regardless of environment changes and microphone types.

To satisfy the requirements of usability and efficiency, we propose MicGuard shown in Figure 1, the first passive detection system against mainstream out-of-band injection attacks involving ultrasound, lasers, and electromagnetism. The key insight of MicGuard is to capture common traits caused by these out-of-band injection attacks without relying on prior knowledge from attackers or target microphones. By investigating the remaining attacking phenomena across ultrasound, laser, and EM modalities, we uncover two inherent characteristics: carrier traces and spectral chaos. It is found that carrier traces in recorded audio are possibly triggered by carrier waves serving as modulation and transmission of injection attacks, depending on victims' low-pass filters and attacking power. Based on it, we design a prejudgment stage for MicGuard to pre-reject some high-energy injection signals by identifying the carrier trace, minimizing computing cost and operation time. The spectral chaos refers to the disorder of acoustic spectra caused by out-of-band frequency response and non-speech radiation (e.g., thermal, electromagnetism). This is because the forced out-of-band injection definitely causes extra energy emissions in the transformation of non-speech analog-to-audio digital, given the internal circuits and parameters of microphones designed for in-band audio. To quantify these distortions, we propose the spectra-chaotic entropy, enabling MicGuard to memorize the normal entropy characteristic and then distinguish the abnormal. We design the memory-based detection network to store chaotic maps from normal audio in the feature memory bank. Notably, our designed detection model solely absorbs normal samples stemming from open-sourced audio datasets and distinguishes abnormal ones based on the distribution differences queried from the memory bank, all without requiring any information from target devices.

The MicGuard effectively bridges the gap between restricted defense and full-scale attacks to a considerable extent. It should be noted that the out-of-band attacks examined in this paper primarily focus on ultrasonic, laser, and EM attacks. These are chosen as they are among the most prevalent injection types and offer a representative cross-section of the injection attack landscape. In summary, our contributions are as follows:

- To the best of our knowledge, MicGuard stands as the most comprehensive detection system designed to combat out-of-band injection attacks against microphones.
- We explore the shared attacking patterns across three out-of-band injection attacks and uncover two key properties that can serve as benchmarks for detection. Building upon these findings, we develop MicGuard, a novel system capable of detecting anomaly-injected signals without requiring preliminary data from either attacks or targets.
- We evaluate MicGuard on 16 different levels of microphone-based devices, including smartphones, tablets, recording microphones, and sensors. Extensive experiments show that MicGuard can achieve exceptional detection ability with 98% AUC, and is resilient in ambient noise and attacking parameters.

2 Out-of-band Injection Attack Model

In this study, we introduce the out-of-band attacks on microphones, involving their vulnerabilities of microphones to ultrasound, laser, and EM modality, as shown in Figure 2, and consider the threat model that is also implicitly adopted in previous work.

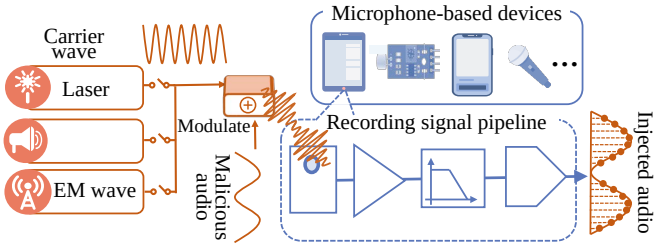


Figure 2: The attacking diagram of those three out-of-band injection attacks includes ultrasound, laser, and EM attacks, where the attacker modulates the malicious audio on the out-of-band carrier waves and transmits it into the microphones.

Attackers' Goal. The ultimate goal of attackers is to introduce malicious signals into microphones, compromising the associated systems, such as voice-controlled systems, television/radio broadcasting, and industrial automation.

Attackers' Capability. The victim microphones and microphones-based devices are commonly seen in daily life, multimedia, and industrial manufacturing, e.g., smartphones, recording microphones, and sound sensors. The attacker cannot gain physical or malware-based access to the target device. Aiming at successful injection, attackers can employ sophisticated hardware and energy resources to execute one of the following attacks:

- **Ultrasound attack** also called dolphinattack [53], is one of the mainstream out-of-band attacks on microphone sensors. Attackers modulate malicious audio on the amplitude of ultrasound carriers (e.g., $>$ 24kHz) and then transmit the ultrasound to interfere with microphones through

ultrasonic speaker arrays. Despite ultrasounds operating at high frequencies beyond the range of human hearing, the inherent non-linearity of amplifiers inside microphones inevitably generates multi-order multiplication of signals upon receiving the modulated ultrasound [35]. The second-order multiplication consequences render the microphone to record original audio modulated on carriers inadvertently. Detailedly, the well-designed ultrasound $S_{ultra}(t)$ is formulated as follows [18]:

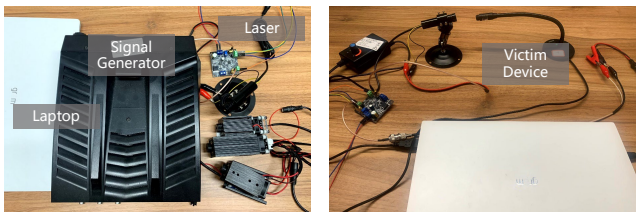
$$S_{ultra}(t) = (m(t)+1)cos(2\pi ft) \quad (1)$$

where $cos(2\pi ft)$ denotes an ultrasound carrier with the frequency of f and $m(t)$ is the modulated audio, e.g., a human voice command signal saying "Alexa, mute yourself." The corresponding non-linearity production is $m(t)+\frac{1}{2}m(t)^2$, where the second-order production $m(t)^2$ is so weak that it can be ignored. Figure 3(a) depicts the ultrasound attack setup, which consists of three frequency bands (25kHz, 32kHz, 40kHz), high-power ultrasonic speaker arrays, and two amplifiers. This configuration ensures that the modulated ultrasonic signal is properly fed into the target microphones and remains intact. If necessary, the attack can be amplified by a high-performance amplifier, and it is effective within line-of-sight range.

(a) The setup of ultrasound attack. (b) Victim devices.

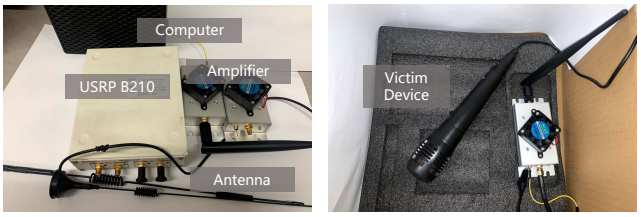
Figure 3: The experimental setup of ultrasound attacks is used and the targeted devices including smartphones, tablets, recording microphones, and sensors are tested in this paper.

- **Laser attack** also known as lightcommands [39], are a sophisticated means of compromising microphones using varying laser beams. By precisely modulating the intensity of lasers to encode audio, attackers can remotely manipulate the linked systems, e.g., voice-controlled systems and live broadcasting. The laser beam is supposed to be focused on the sound hole as hard as possible, whereby the laser directly shoots at the diaphragm to cause its vibration due to photoacoustic effects. The light-induced vibration is converted to the recorded audio of targeted microphones. The attacking setup of laser attacks is shown in Figure 4 and can be configured with 450nm laser transmitters at varying power levels of 100mW, 500mW, and 2500mW, respectively. The modulation driver of laser attacks are transistor-transistor-logic (TTL) modulation and amplitude modulation (AM), respectively.The attacker has remote line-of-sight access to the target microphone and emits lasers into its sound hole.
- **EM attack** targeting microphones in this paper are categorized as low-power intentional EM interference [23, 56], which manipulates the voltage changes of transducers by using specifically crafted EM signals. The stealthy injection of modulated signals on EM carriers into microphones can be attributed to the following reasons [41, 44]: 1) electronic components such as amplifiers and analog-to-digital converters (ADCs) in PCB circuits can couple with electromagnetic signals to absorb energy, which is also known as electromagnetic coupling; 2) the connecting wires between components act as antennas that can receive the invasive electromagnetic wave. To maximize the coupling effect, attackers must first determine the resonance frequency of target microphones using sweep frequency techniques. The manufactured signals that attackers intend to inject into microphones are AM-based modulated electromagnetic waves with a certain resonance frequency. Figure 5 shows the attacking setup for EM attacks, which includes an amplifier and a USRP. To ensure a high success rate for the EM attack, position the transmitting antenna as close as possible to the target or its USB cord, potentially removing the insulation from the charging wire. The attacking device's antenna is positioned close to the victim, potentially making direct contact with the wires to the charger of the target, and can also be connected to an amplifier to increase the intensity of the electromagnetic radiation.



(a) The setup of laser attack. (b) Attacking scenario.

Figure 4: The experimental setup of laser attacks and its attacking scenario are tested in this paper.



(a) The setup of EM attack. (b) Attacking scenario.

Figure 5: The experimental setup of EM attacks and its attacking scenario are tested in this paper.

3 Preliminary Analysis

Before we design a defense mechanism against out-of-band injections, it is necessary to analyze the consequence arising from these injection attacks and uncover their common characteristics. The characteristic is the key to detecting all injected traces across various microphones.

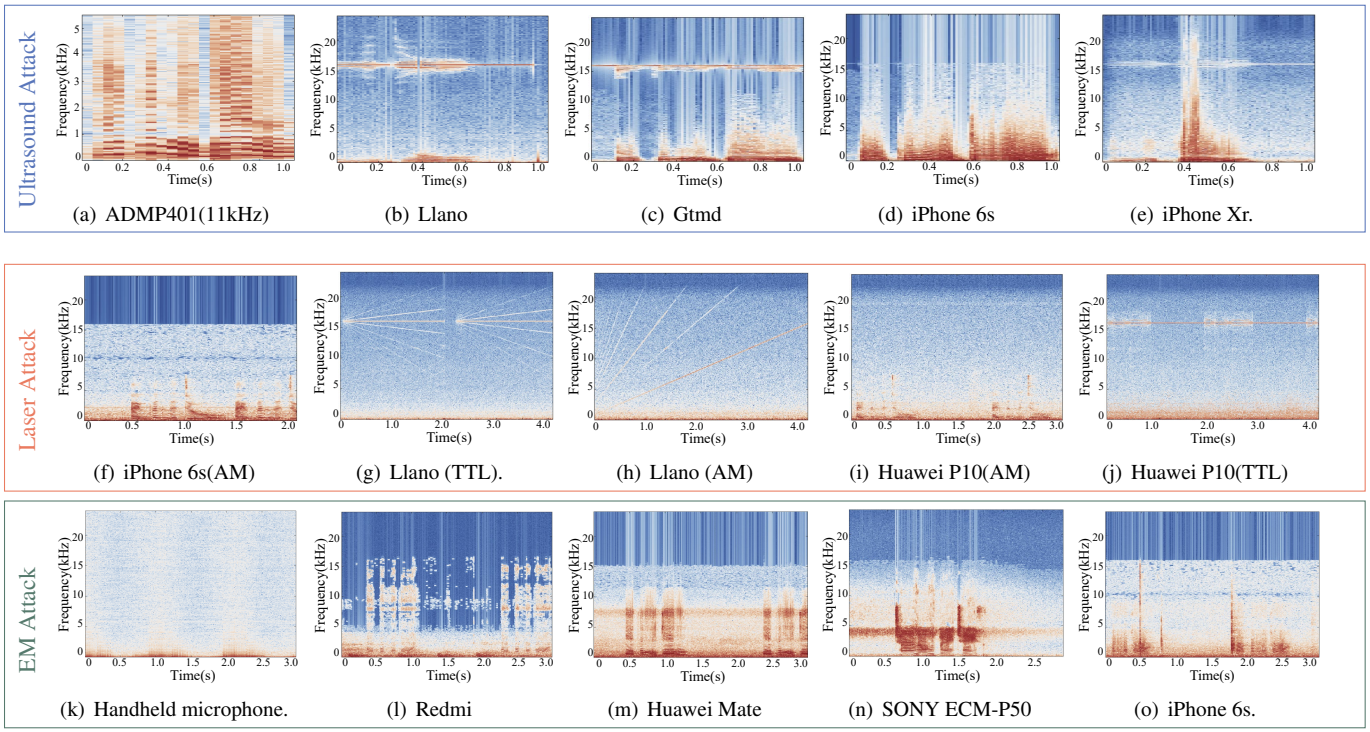


Figure 6: The measured spectrum of sampled sound from the microphone under ultrasound, laser, and EM attacks. The victim devices includes smartphone, recording microphones, and sensors, under ultrasound. TTL: laser attacks under TTL modulation. AM: laser attacks under amplitude modulation.

3.1 Trait 1 : Carrier trace

We execute the three out-of-band injection attacks on commonly-used microphones: microphone sensors, recording microphones, and smartphones. The attacking setups employed for these attacks are described in Section 2, and the victim microphones are shown in Figure 3(b). The injected results recorded by microphones are illustrated in Figure 6. The laser attacks are executed by TTL modulation and AM, respectively. The attacking results of TTL laser or AM laser are labeled by TTL or AM in the figure caption. The carrier frequency of ultrasound attacks and laser attacks under TTL is 32kHz and 16kHz, respectively. The 16kHz/32kHz in laser attacks is the frequency of emitted pulse of TTL laser, which stands as the carrier frequency of modulated audios.

Looking deeply at each spectrum of injected audio, we find that some frequency traces like straight lines span the entire timeline, as seen in Figure 6(b)(c)(e)(g)(j). Such frequency traces only happen in specific frequency points, whatever the types of victim microphones. Detailedly, a line of 16kHz frequency trace occurs in ultrasound-injected results of iPhone Xr and recoding microphones (i.e., Llano, Gtmd). Regarding the cause of this phenomenon, the observed frequency traces are the result of frequency shifting a single-frequency signal during the ADC sampling. Based on the following sampling formula:

$$|f_a| = |f - m f_s|, \quad |f_a| \leq \frac{f_s}{2}, \; m \in Z, \tag{2}$$

where a sinusoid of frequency f Hz is sampled at f_s samples/sec, and the sampled sinusoid of frequency f_a is folded in the band 0 to $f_s/2$. By substituting the 48kHz sampling rate f_s and 16kHz frequency traces f_a in Eq. 2, we can derive that the frequency value f of the signal being shifted is 32kHz. In a similar vein, look backing at the TTL laser injection results, it can be obtained that the frequency traces on the spectrum of laser injection results from a 32kHz or 16kHz.

Remarkably, *the frequency value is exactly equal to carrier frequency. 1)* Laser-injected trace attributed to the carrier: In laser attacks, the laser frequency range comprises the carrier wave (i.e., 16kHz) and modulated audio band, which does not extend beyond 16kHz significantly. Moreover, only TTL laser attacks can cause a straight trace in a spectrum. Whereby TTL-modulated lasers require a carrier while AM-modulated lasers do not, the trace effect is attributed to the carrier wave. *2)* Ultrasound-injected trace attributed to the carrier: However, it is generally seen that the carrier wave should be filtered out after the microphone low-pass filtering ($f >$ the cut-off frequency of the low-pass filter) and cannot be downsampled through ADC sampling. Practically, the gain-magnitude frequency response will gradually decay beyond the cut-off frequency of the low-pass filter (e.g., the cut-off frequency at 24kHz), rather than abruptly dropping to zero. Thus, the carrier with high power can bypass the low-pass filter. Particularly for professional recording microphones, the cut-off frequency of the low-pass filter is looser, which loosely allows carriers to pass through.

However, *why does the phenomenon of carrier traces not exist in the injection results of EM attacks?* To answer it, we should first understand the role of carrier signals among these

three injection attacks and analyze whether it can cause carrier traces in recordings of victim microphones. 1) Ultrasound attacks rely on ultrasonic carrier signals to modulate audio signals that adversaries want to inject and transmit into targets. The ultrasonic probe will inevitably emit ultrasonic waves that will be downsampled by the ADC, given that the internal low-pass filter does not entirely attenuate them. 2) Laser attacks choose a laser beam driven by TTL modulation wherein the carrier signal is a PWM optical signal. The carrier signal will vibrate the microphone's diaphragm consistent with its frequency so that the victim can record the carrier. Note that if the laser driver is amplitude modulation (at a much higher cost than TTL) and no carrier signal is required, then there will be no carrier trace in the laser injection results. 3) While EM attacks also need carrier signals for injection, the electromagnetic carrier is used to the electrostatic couple and is neither the actual sound nor sound source inside microphones. Thus, EM attacks cannot bring about carrier traces.

Insight 1: Considering the potential occurrence of carrier traces in the results of ultrasound-induced and laser-induced attacks, we can leverage the characteristic of carrier traces to quickly pre-determine whether an out-of-band attack exists.

3.2 Trait ② : Spectral chaos

While we have uncovered a trait to identify the attacking trace, its limited applicability drives us to explore additional detection trait that fulfills the following conditions: 1) Discriminability: this trait relies heavily on the distinctive characteristics of out-of-band injection attacks rather than normal sound; 2) Stability: This trait consistently exhibits strong reliability across all microphones.

Before we traverse all invented statistical or acoustic features, we first analyze the analog-to-digital process of microphones when facing out-of-band signals or sound, respectively. A MEMS microphone is an acoustic-electrical transducer that translates air pressure into an electrical quantity. The output voltage V_O of MEMS microphone can be modeled as a variable capacitor formed by a flexible membrane subject to incident pressure P_S [42]:

$$V_O = -\frac{\kappa C_0 V_B P_S}{\varepsilon_0 A}, \tag{3}$$

where A is the area of capacitor plate, ε_0 is the vacuum dielectric permittivity, V_B is the biased voltage, C_0 is the initial capacitance in the absence of sound and κ is the deformation sensitivity. For a linear microphone, the corresponding Fourier transformation is $V_O(\omega) = H_{mic}(\omega)P_S(\omega)$ and $H(\omega)$ is the frequency response function. When high-frequency signals, e.g., ultrasound and electromagnetic signals, flow through analog circuits of microphones, they will induce parasitic capacitance C or inductance L and thereby enhance the nonlinearity, simply formalized as: $\tilde{H}_{mic}(\omega) = H_{mic}(w)\frac{1}{j\omega C}$ or $\tilde{H}_{mic}(\omega) = H_{mic}(\omega)\, j\omega L$. Obviously, the out-of-band signals can cause deviations from the intended frequency response, resulting in distorted acoustic recordings. Moreover, since the hardware component of microphones is designed for in-band signals, the components cannot fully handle unwanted out-of-band signals, thus creating by-products, i.e., multiple harmonics and, in severe cases, crosstalk. According to the distortion measurement of the acknowledged formula [20]:

$$THD = \frac{\sum_{n=2}^{\infty} Power(\omega_n)}{Power(\omega_1)}, \tag{4}$$

the level of distortion is defined by the ratio of harmonic power to fundamental power. The harmonics originating from out-of-band sources substantially introduce distortions and chaos in sampled voltage.

Building upon the above analysis, we consider the distortion of recorded audio as an additional trait to identify the out-of-band injection occurrence. We leverage the concept of entropy, a well-established measure in information theory, to quantify the level of spectral chaos in audio and assess its resilience. Following the entropy principle introduced by Richman et al. [33], we utilize sample entropy to quantitatively assess the chaos of audio samples recorded from diverse victim devices exposed to various out-of-band injection scenarios. It is noted that audio samples contain distinct semantic content like 'open the door,' and each sample has a duration of 0.8-3 seconds with a 48kHz sampling rate. In Figure 8, the entropy evaluation demonstrates *1) high distinctiveness*: the original audio (avg: 0.0798) exhibits lower entropy compared to adversarial audio (avg: 0.3617, 0.6111, 0.3736), indicating that the injected audio has intense chaos; and *2) stability*: the high entropy of adversarial audios is consistent across different attacks, injected content and devices settings. These findings encourage us to design a detection method that leverages entropy information from the chaotic nature of microphone-recorded audios for out-of-band attack detection.

Insight 2: Out-of-band signal injection brings about distinct chaos across various types of victim devices, which can be employed as a key trait to detect attacking audio.

4 System Design

4.1 Prejudgment

As discussed in Section 4.1, the presence of carrier traces in recorded audio can serve as a prejudgment criterion to proactively reject signals arising from partial out-of-band attacks. This prejudgment module aims at fast detection with minimal computational overhead and operational time.

Notably, carrier traces in spectrums exhibit a horizontal line across the whole time domain. Based on it, we only need to detect the existence of horizontal lines in audio spectrum received by microphones to determine the abnormal out-of-band signals. In terms of straight-line detection, especially in image processing, there are several classical algorithms, such as Hough line detection [2] and line segment detector

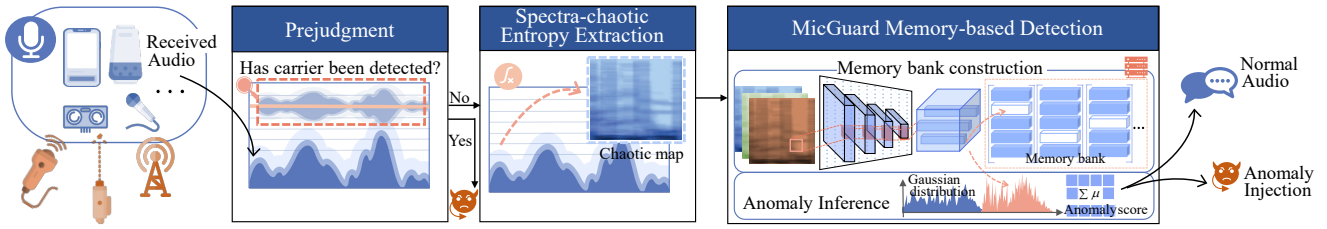


Figure 7: The system overview of MicGuard that firstly rejects partial injected out-of-band signals and identifies the remaining anomaly based on entropy extraction and detection model.

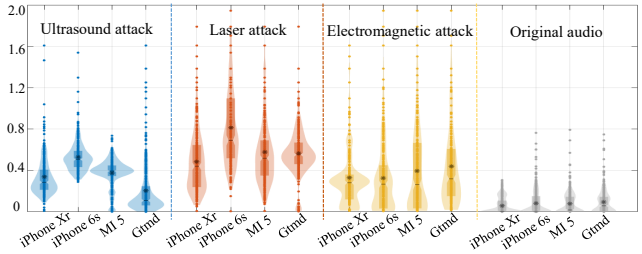


Figure 8: The calculated entropy results of received audio across different devices under different out-of-band attacks

(LSD) [14]. However, these methods require complex iterative computations and parameter adjustment, and they are more adept at handling high-quality images rather than chaotic images, making them unsuitable for the system prejudgment. In MicGuard, we design fast carrier trace detection, as detailed in Algorithm 1.

Algorithm 1 Fast Carrier Trace Detection

Require: A sample of recorded audio A
Ensure: The detected carrier trace C

$[S, f, t] = STFT(A)$;
$S = Normalize(S)$;
$CarrierTemplate = ones(1, length(t))$;
/*Calculate the correlation between each frequency line and carrier template*/
for each frequency point f_i in S **do**
  $Coeff(i) = corr(CarrierTemplate, S(f_i, :))$;
end for
$[Corr_{Max}, f_{MaxIndex}] = max(Coeff)$;
Calculate the slope L_S of $S(f_{MaxIndex}, :)$;
if $L_S \to 0$ and $Corr_{Max} > Threshold$ **then**
  Find the carrier trace $C = S(f_{MaxIndex}, :)$;
else
  No carrier trace in A;
end if

Once the microphone receives an audio sample A, the prejudgment module first performs short-time Fourier transform (STFT) on A to obtain the STFT matrix S where time frames span the columns and frequency frames span the rows. Considering that the carrier trace in the time-frequency diagram is the same as a straight line in shape, we construct a horizontal line template to match all frequency lines and locate the carrier trace by searching for the maximal correlation coefficient between each frequency line and template, as line (3)~(8) indicate. Higher correlation coefficients indicate that the energy distribution at the specific frequency point is closer to the horizontal line. After finding the frequency line with the maximal correlation coefficient, we calculate the slope of the selected frequency line to further judge whether the energy distribution is uniform. If the slope is approximately zero and its corresponding frequency line is highly relevant to templates, the system detects the carrier trace from out-of-band injection and rejects the audio sample at a fast response. Figure 9 shows the detection result of our algorithm and provides a comparison with other line detection algorithms (i.e., Hough line detection and LSD). While computer vision methods struggle to detect the horizontal carrier trace except for multiple thin broken line segments in a disordered spectrum, our approach can accurately locate the adversarial carrier required for certain out-of-band attacks. Considering the extreme recording scenario where users record single-tone signals, it is feasible for users to optionally store these pre-rejected signals after detection or turn off the prejudgment just this once.

(a) Hough. (b) LSD. (c) Our.

Figure 9: The comparison of our proposed algorithm with Hough line detection and LSD.

4.2 Spectra-chaotic entropy extraction

Although we have built up a prejudgment stage to protect against some of the high-energy malicious signals speedily, there is an urgent need to set up a subsequent detection mechanism to re-detect signals that pass through the prejudgment. The goal of this section is to extract the chaotic feature of captured sound that is discussed in Section 4.2, as the criterion of the following detection model.

The mismatch between out-of-band signals and microphone hardware settings will give rise to chaos in the final digitized audio. It is necessary for MicGuard to investigate how to extract intricate chaos comprehensively. The entropy measurement always functions on an entire time sequence,

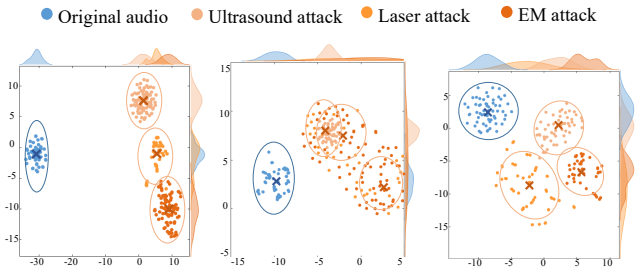
equivalent to downsampling the one-dimensional acoustic signal to a single value. Such high compression measurement overlooks multi-leveled disorders in injected audio and is unsuitable for out-of-band anomaly detection. To cope with it, MicGuard designs spectra-chaotic entropy extraction to avoid compression of informative entropy. Firstly, we use STFT on the audio sample to acquire a two-dimensional spectral matrix S where the frequency energy distribution changes over time. The STFT utilizes a Blackman window with a size of 512 points and an overlapping of 375 points. We adopt the local aggression on S by dividing S and composing each patch-level entropy representation without losing the time-frequency resolution and usability of the chaos feature. Mathematically, the patch window N with the patch size of p slides in spectrum S and locally aggregates information at position (x,y):

$$N_p(x,y) = \{S(i,j) | i \in [x-\frac{p}{2},...,x+\frac{p}{2}], j \in [y-\frac{p}{2},...,y+\frac{p}{2}]\}. \quad (5)$$

Then, we apply the entropy function on $N_p(x,y)$ to yield the spectra-chaotic map E_p as follows:

$$E_p(x,y) = f_{entropy}(N_p(x,y)). \quad (6)$$

Several optional entropy measurements, including power spectrum entropy [51], sample entropy, and local entropy [4], can substitute for the entropy function $f_{entropy}$. To determine the entropy function, we utilize t-distributed stochastic neighbor embedding (t-SNE) to assess the distinctiveness of the spectra-chaotic map across original audio and out-of-band injected audio. The visualization results of dimension-reduced entropy maps are shown in Figure 10. It is noted that the experimental data for analysis is sourced from Section 3.2. The corresponding inter-class distances are 65.50, 30.50, and 40.56, wherein the power spectrum entropy owns the largest distance between in-band and out-of-band classes. To sum up, the power spectrum entropy is chosen as $f_{entropy}$.



(a) Power spectrum entropy. (b) Sample entropy. (c) Local entropy.

Figure 10: The visualized comparison of different entropy function on original samples and injected samples.

4.3 MicGuard Memory-based Detection

Following the extraction of spectra-chaotic maps, this section aims to identify anomalous instances resulting from out-of-band attacks, which exhibit notable deviations from the normal audio samples. In realistic detection applications, the characteristics of anomalies can not be observed during the model training phase due to the unexpectedness of ultrasound, laser, and EM signals. Additionally, for practical ease of use, it is advisable to minimize the acquisition of normal data from the microphone to be protected. To address these issues, this paper designs the MicGuard memory-based detection model, as shown in Figure 11, which memorizes the distribution of the normal-only from open-source audio datasets and detects the outlying features across different attacking modalities.

4.3.1 Memory bank construction

Feature encoding. The first stage is the encoding of spectra-chaotic entropy maps that will be stored to construct a memory bank [9,34]. The memory bank is a substantial feature dataset that models the normal distributions at various hierarchies of chaotic maps. Upon receiving an input feature map, MicGuard memory bank will be queried to retrieve the most relevant items to identify whether it is similar to the memory storage.

After producing an input chaotic map x_i from a training dataset $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ consisting only of n normal samples from open-source audio online, the system utilizes the convolutional neural network (CNN) to extract the hierarchical embedded representations. Inspired by unsupervised classification tasks in ImageNet [10], the pre-trained CNN not only has sufficient downsampling and distinguishing in feature space but also incurs less training cost. We employ the frozen WideResNet50 [50] fully pretrained on ImageNet database as the feature encoder, composed of three bottlenecks $\{\mathcal{F}^1_{res}, \mathcal{F}^2_{res}, \mathcal{F}^3_{res}\}$ and the last average pooling layer $\mathcal{F}_{avg}$. Unlike classic residual neural networks (ResNets), the bottleneck of WideResNet50 has the core idea of widening the network by increasing the number of filters. Each bottleneck block has three convolution layers with twice the number of channels as ResNet and uses dropout for regularization. Augmenting the quantity of feature channels allows the network to refine a broader range of patterns, making it more adept at understanding intricate features and variations. Empirically, we make use of the intermediate feature maps from these four layers (i.e., three bottlenecks and an average pooling layers). These four feature maps are uniformly upsampled and then concatenated along the channel axis to form the multi-scale feature maps $z_i = \{\mathcal{F}^1_{res}(x_i), \mathcal{F}^2_{res}(x_i), \mathcal{F}^3_{res}(x_i), \mathcal{F}_{avg}(x_i)\}$. Notably, these feature maps are no longer original chaotic maps but rather their high-level compressed representation that contributes to the data distribution in the network memory.

Feature adaption. The following step is to fine-tune these feature maps on normal audio-oriented feature space rather than universally oriented domains, facilitating anomaly detection. This is because pre-trained encoder is not specifically designed for chaotic feature extraction. We propose using the deep support vector data description (DSVDD) [37] to learn a hypersphere that extracts common characteristics across normal data while eliminating non-informative interferences. The hypersphere encloses the input maps and clusters them densely within a feature space oriented toward normal audio. With the input features datasets $\mathcal{Z} = \{z_1, z_2, ..., z_n\}$, we

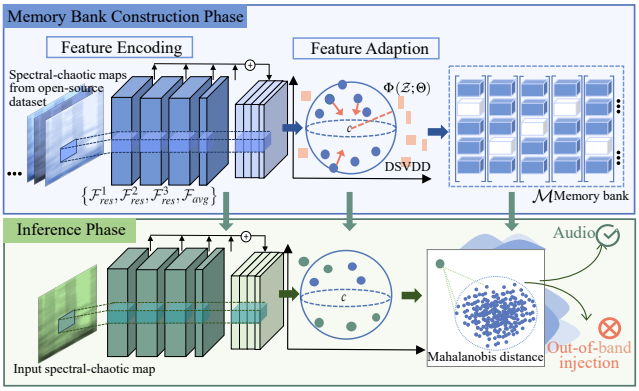


Figure 11: Illustration of MicGuard. Phase 1: The MicGuard encodes the spectral-chaotic maps extracted from open-source audio datasets to construct a memory-based feature bank. Phase 2: Once the input spectral-chaotic map is processed by encoding and adaption, MicGuard searches the memory bank to calculate the Mahalanobis distance between the input and stored normal characteristic distribution to reject the anomaly.

initialize the mapped results as $\Phi(\mathcal{Z};\Theta)$, where Φ is the mapping function with adaptable parameters Θ. The DSVDD employs deep neural networks as Φ to obtain high-dimensional representation and process massive data. The objective optimization is to learn the network parameters with minimizing the volume of a data-enclosing hypersphere, formulated as follows:

$$\min_{\Theta} \frac{1}{n}\sum_{i=1}^{n} \|\Phi(\mathcal{Z};\Theta)-c\|^2 + \frac{\lambda}{2}\|\Theta\|_F^2 . \quad (7)$$

The first item of optimizing objective Eq.7 is the quadratic loss of the distance between the every mapped representation $\Phi(z_i;\Theta)$ and the center of the hypersphere c. The second item is a network weight decay regularizer with hyperparameter λ to prevent over-fitting. The $\|\cdot\|_F$ denotes the Frobenius norm. The DSVDD contracts the hypersphere by progressively aggregating the normal inputs closer to the center of the sphere. As a result, we obtain the adapted feature bank $\mathcal{M} = \Phi(\mathcal{Z};\Theta)$ that can be used for subsequent anomaly inference.

4.3.2 Anomaly Inference

For each test input t_i from the above processing (i.e., encoding and adaption), the inference phase measures its similarity to the registered feature bank $\mathcal{M} = \{m_i = \Phi(z_i;\Theta), i \in [1,n]\}$ to identify whether it is out of the normal distribution. However, owing to environmental fluctuations and unstable acquisition equipment, noisy data is inevitably in the audio datasets or the received test input. Most unsupervised detection algorithms are susceptible to noise factors because they treat all information equally. To alleviate this noise impact, with the assumption that the amount and intensity of noise are less than the norm, we set multivariate Gaussian to model the joint probability density of multiple characteristic dimensions. The key advantage of using a multivariate Gaussian distribution is that it only fits the ordered principal feature distributions while disregarding out-of-order fluctuations such as noise. The multivariate Gaussian distribution model on memory bank $\mathcal{M}$ is defined as:

$$p(\mathcal{M}|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left\{-\frac{1}{2}(\mathcal{M}-\mu)^T\Sigma^{-1}(\mathcal{M}-\mu)\right\}, \quad (8)$$

where μ is the mean of $\mathcal{M}$ and Σ is the variance-covariance matrix by using maximum likelihood estimation. Under the Guassian distribution with mean μ and convariance Σ, a Mahalanobis distance $\mathcal{D}$ between a test input point t_i and the modeled distribution is measured as:

$$\mathcal{D} = \sqrt{(t_i-\mu)^T\Sigma^{-1}(t_i-\mu)}. \quad (9)$$

The Mahalanobis distance $\mathcal{D}$ is interpreted as the anomaly score, with a higher Mahalanobis distance indicating a greater outlier score. The t_i with an anomaly score beyond the threshold boundary is judged as the out-of-band injection.

5 Evaluation

To evaluate our designed system, we pre-train MicGuard only based on audio datasets available online and collect out-of-band spoofing data. We employ diverse microphones and microphone-equipped devices to record the injected sound for evaluation in different conditions.

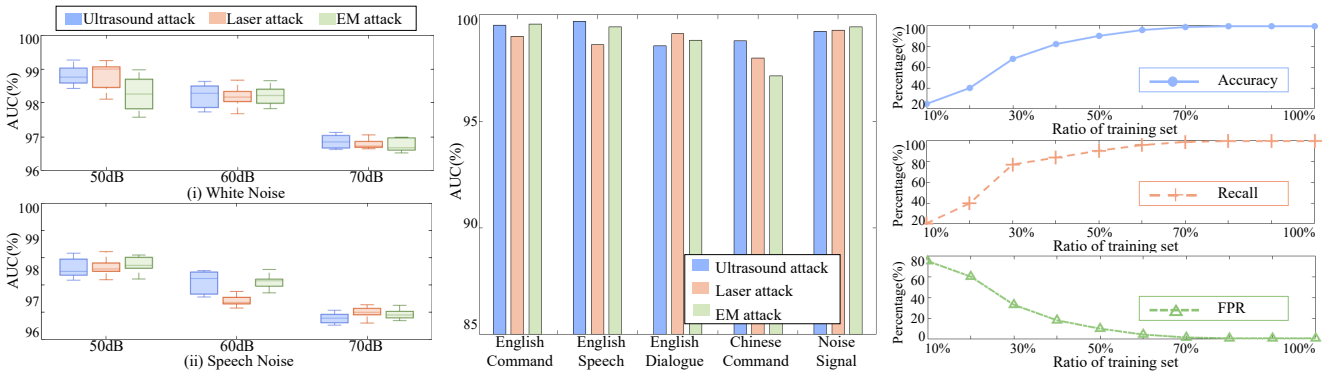
5.1 Evaluation Setup

Hardware Setup. The out-of-injected attacking setups have been described in Section 2. For victim devices, we employ smartphones, tablets, recording microphones, and microphone sensors. All subsequent experiments, including overall performance (Section 5.2) and robust analysis (Section 5.3), employ the identical attacking setup as delineated in our threat model (Section 2).

Dataset. Since MicGuard does not rely on any prior knowledge from the device to be protected, we only need to collect spoofing audio from diverse microphones and genuine audio for testing. *(1) Spoofing dataset:* In this phase, we utilize those three types of spoofing devices to inject audio samples into 16 devices, whose content includes voice commands, speeches, and artificial noises. After removing incorrectly recognized samples, we get 333,655 spoofing samples with 16kHz sampling rate. *(2) Genuine dataset:* In this phase, we recruit 20 participants to speak voice commands, which will be recorded by the same 16 microphone devices as in the spoofing audio collection. The experiments are under the approval of the IRB of our institutions. During the experiments, all participants are informed and approved of the purpose of our experiments. After splitting the audio and removing invalid samples, we collect 31,844 samples with 16kHz sampling rate. *(3) Training dataset:* In the training phase, we choose open-sourced LibriSpeech [31], a corpus of English speech, to fully pre-train the MicGuard memory-based detection model. We divide

Table 1: Experiment devices, category, and AUC results. We evaluate the detection ability of MicGuard among diverse devices and differnt powers of injection attacks in an office environment with a background noise with 30dB SPL.

Num.	Category	Devices	Ultra Attack(%)			Laser Attack(%)			EM Attack(%)	
			3W	6W	15W	0.1W	0.5W	2.5W	20dBm	30dBm
1	Smartphone	iPhone 6s	99.63	99.25	98.53	98.47	98.27	98.89	98.95	99.63
2	Smartphone	iPhone 14	98.10	99.21	99.25	98.91	99.57	97.95	98.30	97.81
3	Smartphone	Huawei P10	99.64	98.54	99.31	98.23	98.46	97.90	99.04	98.08
4	Smartphone	MI 5s Plus	99.62	98.54	99.31	98.23	98.17	98.81	99.11	99.42
5	Smartphone	Nubia	98.28	98.72	98.12	98.73	98.28	99.28	99.22	98.98
6	Smartphone	Redmi K50	99.32	99.14	98.65	98.76	98.97	99.57	98.66	98.47
7	Tablet	iPad	98.07	97.86	99.03	99.13	98.70	98.05	97.96	98.56
8	Tablet	iPad Pro	98.60	98.33	99.15	99.49	98.47	98.88	98.24	98.29
9	Recorder	Llano	99.54	97.89	99.23	99.62	99.38	98.69	99.54	99.32
10	Recorder	HP	99.31	97.98	98.32	98.84	98.91	97.82	98.09	98.62
11	Recorder	Philips	99.62	99.36	99.09	98.06	98.84	98.44	99.37	99.53
12	Recorder	SONY ECM-P50	99.05	99.12	99.04	98.08	99.54	98.11	98.82	98.15
13	Sensor	ADMP401	97.87	98.40	98.11	98.29	98.34	97.31	97.69	98.30
14	Sensor	MCP6022	99.41	99.61	98.03	99.40	99.24	98.39	97.95	98.08
15	Sensor	SPH0645	99.57	97.87	98.75	98.28	99.23	98.80	98.64	98.06
16	Sensor	INMP441	99.09	98.63	99.62	99.35	98.52	98.11	98.00	99.45



(a) Noise. (b) Injected Content. (c) Training Size.

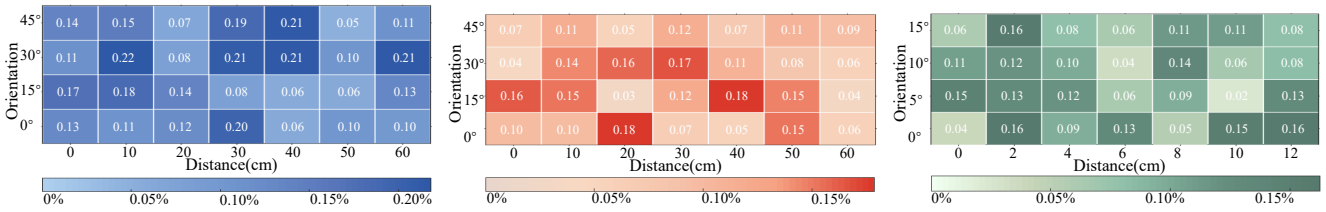
Figure 12: The Robustness analysis: the performance of MicGuard under ambient noise, injected content, and training size impact.

2,438 pieces of samples with a 16kHz sampling rate from LibriSpeech as our training dataset to construct a feature bank.

Metrics. We adopt the following metrics to evaluate the performance of MicGuard: False positive rate (FPR): it is the probability that the system fallaciously accepts the spoofer. False rejection rate (FRR): it is the probability where the system refuses access to the genuine. True positive rate (TPR): it is the probability that the system correctly accepts the normal audio, which is also called Recall. The receiver operating characteristic area under the curve (AUC): it is defined as the area under the ROC curve. When the number of positive and negative instance is imbalanced, AUC can provide a more robust evaluation of the model's performance Accuracy: it is the measurement of overall correctness in classification. Note that the unit of these metrics is percentage(%)

5.2 Overall Performance

We use the attacking setup as shown in Section 2 to perform those three out-of-band attacks and evaluate the overall performance of MicGuard in four categories of microphone-equipped products listed in Figure 3(b), including smartphones, tablets, recorders, and sensors. These products serve as voice interaction, social communication, industrial perception, and more. For each microphone device in evaluation, we set the attacking setup close to the target at a distance of 15cm. The default experimental environment is an office with background noise with a 30dB sound pressure level (SPL). We repeat the three out-of-band injection attacks on each device and also collect genuine audio from each device to measure the anomaly inference of MicGuard. In order to comprehensively assess the attacker's injection capability, we variably adjust the transmitted power of injection setups. The overall



(a) The FPR(%) under ultrasound attack. (b) The FPR(%) under laser attack. (c) The FPR(%) under EM attack.

Figure 13: The Robustness analysis: the detection distribution (FPR%) of MicGuard facing out-of-band injection attacks across various positions.

experimental results of MicGuard are shown in Table 1. MicGuard demonstrates remarkable effectiveness in countering out-of-band injected audio while accurately processing authentic samples across a wide range of devices. Specifically, MicGuard achieves an average AUC of 98.8%. We observe that there is a slight decrease in the AUC of the ADMP401 sensor. Presumably, unlike commercial microphones, such sensors have relatively poor audio sensing capabilities and occasionally produce loud noises or even garbled audio. Despite increasing the power of injection attacks, MicGuard maintains a high AUC above 97%. This is attributed to the fact that no matter how much power is injected, they invariably induce disruptions in the recorded audio.

5.3 Robustness Analysis

Noise. To investigate the impact of noise, we collect spoofing audio from three injection attacks and genuine audio under different types and strength levels of noise, respectively. Two commercial loudspeaker boxes play white noise and speech noise at the SPL of 50dB, 60dB, and 70dB. After collecting these audio, we feed them into the pre-trained model to calculate the AUC. Figure 12(a) shows the AUC of detection results, from which we can find that the averaging AUCs of all injection attacks are 98.61%, 98.20%, 96.79% under white noises and 97.66%, 96.88%, 95.90% under speech noises. The AUC decreases by up to 3%, and the higher the ambient noise, and the extent of the reduction is positively associated with the increase in ambient noise level, especially high-level noise. This is because the benign audio is falsely rejected since the added noise, similar to chaos, potentially disturbs the judgment of the model. The AUCs under speech noise are lower than those under white noise. It is speculated that the added speech noise may mask the original injected malicious signals, potentially causing a slight increase in the FPR.

Injected Content. Though we have obtained overall detection results on spoofing audio signals containing voice commands and speeches, it is unclear to what extent the semantic content modulated on out-of-band signals affects MicGuard detection outcomes. In this case, we divide and select the previously collected dataset into the commands, speeching, and dialogue datasets. Moreover, considering the limited language (i.e., English), we add two new spoofing datasets where the modulated content is Chinese commands and noise signals like pulse, Gaussian, and thermoelectric. The added noise data can furthermore confirm the anti-interference property of MicGuard. Figure 12(b) demonstrates the averaging AUC of 98.94%, indicating the robustness to the injected content. When the injected language is Chinese, there appears to be a slight decrease in AUC. This can be attributed to the fact that the pre-training dataset exclusively comprises English data, thereby leading to a more robust feature representation in the model for English inputs. Additionally, we attempt to inject noise to assess whether these artificial noises can bypass the decision boundary of models by accident. The high AUC surpassing 99% under noise injections demonstrates the robust performance of our system.

Training Size. It is well-known that the amount of training data is crucial in deep learning methods in terms of feature learning and generalizing. In this case, we split the training datasets into different proportions, namely, from 10% to 100%. Figure 12(c) shows the accuracy, recall, and FPR of our system with different training set sizes. As expected, the evaluation results substantiate that increased data for training achieves higher accuracy and recall. When the training proportion exceeds 70%, we observe a mere modest increase in system accuracy. This underscores the notion that merely augmenting the volume of training data may not enhance the generalization capacities of models.

Position. It is flexible for adversaries to launch out-of-band signals at varying distances and orientations from the victim devices. In this case, we investigate the attacking position on the effectiveness of MicGuard. Those three types of out-of-band injection signal launchers are positioned at distances ranging from 0cm to 60cm and off-axis angles spanning 0 to 45 degrees. In EM attacks, the attacking distance is restricted by the significant attenuation of electromagnetic radiation over distance. In realistic experiments, the EM attacking distance is 0cm to 12cm. For all attacking testing, we only collect efficient injected signals of victims. To ensure the success of injection attacks, we increase transmission power or the number of speaker arrays in ultrasound attacks. In laser attacks, the laser is aimed at microphones from different orientations. Note that 0cm means the transmitting probe is tightly attached to the target device. The visualized FPR of three injection attacks at different positions is illustrated in Figure 13. Overall,

our system can fully resist malicious signals from various angles and distances, achieving an impressively low FPR of just 0.113%. In fact, with the distance increasing, the energy of injected signals dissipates, even missing partial high-frequency spectrums, unlike typical audio. As a result, despite the reduction in causal-produced chaos, the excessively attenuated spectrum also exposes its out-of-band characteristics. Furthermore, there is no apparent correlation between orientation and FPR. It demonstrates that orientation variations have a minimal impact on the performance of MicGuard.

5.4 Ablation Study

Impact of Prejudgment. To assess the influence of the prejudgment module, we devise two models: one without the prejudgment module (i.e., W/O Pre) and the other retaining it intact. Both of them have the same testing data from three injection attacks and training data. Regarding overhead time, the comparison experiments run on an Nvidia GeForce RTX 2080Ti with a batch size of 24 and are repeated for 800 trials to measure the average overhead time. As shown in Table 2, W/ Pre has a small 0.4% improvement in accuracy over the spoofing injection datasets. Please note, that a reduction from an average time of 91.2ms (W/O Pre) to 55.0ms for W/ Pre means 1.7 times in running time efficiency. In personal use and industrial inspection, this is a relevant and significant runtime reduction. Compared to the W/O Pre method, the designed prejudgment fleetly detects the carrier if injected signals are AM modulated on ultrasound and TTL modulation on laser with no need for arbitrary follow-ups, which will largely reduce the time complexity. Furthermore, we test the latency of MicGuard on smartphones. We transfer the pre-trained MicGuard model into the mobile model for Android and use the standard Android Studio and Pytorch [32] to deploy it on the HUAWEI P40 smartphone. The average overhead on the smartphone is 231.4ms.

Table 2: Ablation studies of the prejudgment. W/O Pre means the MicGuard system is stripped of prejudgment. W/ Pre means the MicGuard system with the prejudgment

Method	Accuracy		Overhead Time	
	Ultra+Laser	Ultra+EM	Ultra+Laser	Ultra+EM
W/O Pre	98.87%	98.76%	93.22ms	89.26ms
W/ Pre	99.32%	99.11%	49.28ms	60.70ms

Impact of Spectral Chaotic Maps. To investigate the effectiveness of the proposed spectral chaotic maps, we conduct several ablation experiments where the input is changed into phase spectrum, magnitude spectrum, and raw signal without processing. Furthermore, we measure the combination benefits of spectral chaotic maps and each component of our method. The detailed experimental settings are as follows:

- Phase+W/ Adaption, where the input to the memory-based detection model is phase spectrums of received audio of microphones and the detection model still maintains the feature adaption module (W/ Adaption).
- Magnitude+W/ Adaption, where the input is magnitude spectrums.
- Raw+W/ Adaption, where the input is raw time-domain signals.
- Chaotic+W/O Adaption, where the input is spectral chaotic spectrums to the detection model that does not include feature adaption.
- Chaotic+Resnet50, where the input is chaotic maps but the feature encoding network is ResNet50 instead of WideResNet50 in detection designs.
- Chaotic+EfficientNet, where the input is chaotic maps and the feature encoding network is EfficientNet network.
- Ours, where the input is chaotic maps and the original system design retains integrity.

These above models are fully pretrained on the LibriSpeech database and equally evaluated on the same dataset. Table 3 presents the distinct results of model settings on out-of-band injection detection. Firstly, take a look at the case of changing the input feature to the detection model. Only the chaotic map has high performance, while conversely, all of the other remaining features exhibit unsatisfactory. This is because those conventional features cannot capture the inherent commonalities belonging to the consequences of out-of-band injection. At this time, when feature adaption is not used, the system experiences a reduction in accuracy and FRR, in comparison to the Chaotic+ResNet50 and Ours. By inducing normal features to be clustered more discriminatively, the memory feature library stores characteristic distributions that closely resemble acoustic disorders, which considerably enhances the ability to distinguish abnormalities. Lastly, there remains a question as to whether WideResNet50, when used as a feature encoding network, excels in transfer learning compared to other networks. According to the results of ResNet50, EfficientNet, and ours, it can be seen that the network with low model complexities performs worse in transfer learning of injection detection. The WideResNet50 has more convolution filters to produce wider feature maps that broaden feature spaces, allowing fine-tuning of the model on a new dataset with different characteristics.

Table 3: Ablation studies of the spectral chaotic maps, feature adaption and feature encoding.

Method	Accuracy(%)	FPR(%)	FRR(%)
Phase+W/ Adaption	74.71	24.70	31.47
Magnitude+W/ Adaption	76.22	23.03	31.60
Raw+W/ Adaption	73.58	25.89	31.92
Chaotic+W/O Adaption	92.91	7.11	6.87
Chaotic+Resnet50	93.98	5.89	7.43
Chaotic+EfficientNet	92.41	7.53	8.19
Ours	99.37	0.60	0.90

Table 4: Comparison of defense systems against out-of-band injection attacks on microphones.

System	Sensing Type	Defense Mechanism	Hardware Independent	No Attack Data Required	No Target Data Required	Defense Scope		
						Ultrasound	Laser	EM
TransShield [41]	passive	Extra similar circuits	No	Yes	No	✗	✗	✓
Zhang et al. [56]	passive	Bias voltage in sampling	No	Yes	Yes	✗	✗	✓
Audio-visual [15]	passive	Sensor fusion	No	No	No	✓	✗	✗
AIC [18]	active	Ultrasonic demodulation	No	Yes	Yes	✓	✗	✗
Li ei al. [26]	passive	Multichannel microphone	No	No	No	✓	✗	✗
EarArray [52]	passive	Microphone arrays	No	No	No	✓	✗	✗
NormDetect [25]	passive	Patterns in spectrum	Yes	Yes	No	✓	✗	✗
MicGuard	**passive**	Prejudge and detect chaos	**Yes**	**Yes**	**Yes**	✓	✓	✓

6 Discussion

6.1 Defense Scope

MicGuard focuses on the out-of-band injection attacks targeting microphones, expanding the existing defenses against more attacks and ensuring broader application coverage. Nonetheless, taking into account the various attack methodologies and target devices, certain limitations and potential improvements require further discussion.

Attacking methodology. There also exist some attacks aimed at microphones: *1) In-band attack:* The easy-to-operate replay attack and mimicry attack have the possibility to bypass speaker recognition. Whereas the defense scope of MicGuard excludes in-band audio, replaying audio and mimicking pronunciation can be effective in evading detection. In order to extend the defense scope, an alternative approach is to concatenate an existing in-band detection model [26] subsequent to MicGuard. Those existing detections primarily identify whether the sound source is from a human being or not. However, in some cases, the sound source is more than just the human voice, such as industrial production lines and live outdoor broadcasts. Thus, researchers are encouraged to design a more comprehensive detection mechanism for various application scenarios. *2) Adversarial attack:* We envision that the adversary has full knowledge of MicGuard system designs. Grounded on data-driven approaches, the adversary utilizes adversarial training to craft adversarial samples that will be modulated on ultrasound, laser, and EM waves. However, launching such an adversarial attack would still fail in practice due to the out-of-band effects that inevitably leave disorganized patterns in the injected audio.

Targeted devices. Inspired by other passive sensors similar to microphones, the design ideas of MicGuard can be transferred into temperature sensors, humidity sensors, and infrared sensors that also encounter out-of-band threats. The application scenario of MicGuard is not limited to anomaly detection on microphones and has the potential to be expanded into equipment failure prediction and production process monitoring.

6.2 Comparative Analysis of Existing Defenses

Compared with out-of-band defenses. In comparison to existing defenses, as listed in Table 4, MicGuard stands out as a remarkable solution in the field of out-of-band injection mitigation from the following perspectives. 1) Hardware independent: MicGuard operates independently of extra hardware components and microphone conditions, unlike hardware-based defenses and microphone-array methods. This versatility makes it suitable for various levels of microphone-equipped devices, ranging from low-end sensors and live recorders to smart home devices. 2) No prior attack data required: Our system needs no prior attacking data, in contrast to some detection methods that rely on pre-collected attacking data to train learning models. Due to the labor-intensive process of gathering malicious data and its unpredictability, our system is more adaptable to real-world scenarios. 3) No target data required: Our system also does not require prior acquisition data of target devices. Instead, it relies on a pre-trained feature dataset sourced from an open-access audio database. This approach not only reduces deployment costs but also enhances user-friendliness. 4) Defense scope: Taking into account that the majority of prior defense research is restrictively focused on a single injection attack type, MicGuard exhibits a significantly more comprehensive defense capacity, encompassing a wide array of out-of-band injection attacks.

Compared with liveness detection. Liveness detection is another well-known defense against in-band injection attacks on microphones and ultrasound attacks [54]. This technique resiliently counters voice replay attacks [8] and speech adversarial attempts [1] by leveraging the inherent vitality distinction between human voices and the audio speaker required to play spoofing voices. The key difference between liveness detection and MicGuard is application scope. 1) Limited to in-band ranges: Due to its reliance on "living traits," liveness

detection is tailored for human voice recording scenarios such as voice interaction [43, 55]. Though MicGuard is adept at safeguarding both human and non-human audio recording, it is not designed to defend against in-band spoofing signals. Moreover, most liveness detections overlook out-of-band injections like laser and EM waves or, at most, take into account ultrasonic injections. 2) Requirement of hardware setup: some liveness detection methods rely on the extra hardware setup, such as microphone arrays [29], radars [24], IMU [17], and magnetometer [7], rendering them unsuitable for low-level microphone-based devices due to either impracticality or high costs. MicGuard offers a versatile solution applicable across sensors to high-level tablet computers without requiring additional specialized hardware. 3) Need for prior data: MicGuard operates without the necessity of collecting preceding original and attacking data from victim devices. In contrast, traditional liveness detection methodologies [3, 6, 48] still require gathering such prior data to delve into biometric features through deep learning-based training and effectively distinguish anomalies. Delving into the strengths and weaknesses of both MicGuard and liveness detection, there is an anticipation of their combined application to mitigate both in-band and out-of-band attacks. The straightforward way is to add the liveness detection after MicGuard, but this incurs additional costs in terms of overhead, data, and hardware setup. A more comprehensive fusion of the two systems requires identifying the correlation features that underlie in-band and out-of-band detection, and we will leave it to future work.

7 Related work

Given the increasing risk of out-of-band signals compromising microphones, ongoing research is dedicated to uncovering the vulnerabilities of microphones susceptible to out-of-band injection attacks and introducing mitigation strategies to ensure the resilience of microphones.

Out-of-band injection attacks on microphones. Regarding the medium of injected signals, out-of-band injection attacks on microphones can be categorized as *1) Ultrasound attacks, 2) Laser attacks, and 3) EM attacks*. 1) Ultrasound attacks first arose from the nonlinearity of microphones [35, 53], whereby microphones can interpret voice commands modulated in the high-frequency band of ultrasound. By optimizing the arrangement of ultrasonic speakers to transmit disparate frequency-band spectra of signals, Roy et al. can achieve a remote attacking distance of up to 25ft [36]. 2) Laser attacks can directly cause microphones' diaphragms to vibrate in response to the amplitude change of the injected laser, first proposed by Sugawara et al [39]. Some attackers replicate the laser attack through windows to target in-vehicle VCSs [46] by virtue of the long-range propagation and penetration of laser nature, like 110m. 3) EM attacks can induce electrical currents or voltages inside circuits of microphones due to coupling effects, leading to unwanted sampled data in the pipeline of the analog-to-digital converter [21, 23]. Previous EM attacks need to be in close proximity to microphones owing to their rapid attenuation. One straightforward way to extend attacking ranges is to employ power amplifiers [45]. Rather than resorting to direct injection techniques on microphones, attackers exploit EM interference to infiltrate televisions [49] or power chargers to inject commands [11] into VCSs without physical access.

Mitigating out-of-band injection attacks on microphones is crucial for maintaining the integrity and security of sensing systems. One line of defense strategies relies on microphone arrays [26, 52] to differentiate the forged ultrasonic signals by analyzing the variations in multichannel acoustic signals. Spectrum features are also regarded as a vital criterion for distinguishing between sound and abnormal ultrasonic signals [25]. Another avenue for mitigating ultrasonic interference is sensor fusion techniques [15, 16]. By assembling different information from multiple diverse sensors, the intelligent integrated system rectifies conflicts to avoid erroneous judgment. Moreover, He et al. turn to utilize ultrasonic speakers to actively nullify the pernicious effects of invading aggressive ultrasounds [18]. In the domain of defense against EM interferences, researchers tend to modify the original hardware [19, 41, 56] to detect aberrant analog signals caused by out-of-band EM interference. Most systems are only concerned with mitigating one type of injection attack, i.e., either ultrasonic or electromagnetic attack. Regrettably, the defense against laser attacks has been largely overlooked despite their potential risks and emerging prevalence. It highlights a critical gap in security measures and prompts the need for heightened attention and proactive countermeasures to deal with all possible out-of-band injection attacks. This paper designs MicGuard as an attempt to fill the gap, first detecting the aberration on microphones induced by all out-of-band injection attacks.

8 Conclusion

This study tries its best to fill the blank of defense against out-of-band injection attacks, including ultrasound, laser, and EM interferences. We unveil the common characteristics underlying these out-of-band injection phenomena and design a prejudgment and detection system to distinguish the anomaly. Without the dependency on prior data from targets and attacks, the carrier trace is set as the first criterion, and the quantified chaos derived from open-source datasets is incorporated into the memory-based model for ultimate detection. MicGuard is comprehensively evaluated on low-level to high-level microphone-based devices such as sensors, recording microphones, tablets, and smartphones. It demonstrates exceptional detection capabilities, achieving over 98% accuracy against out-of-band attacks even under aggressive conditions.

Acknowledgement

The authors would like to thank our Shepherd and all the anonymous reviewers for their insightful comments. This paper is partially supported by the National Natural Science Foundation of China (62032021, 62372406, 62072395, 62172359, 62102354, and U20A20178) and the Zhejiang Provincial Natural Science Foundation of China under Grant (LD24F020014).

References

[1] Hadi Abdullah, Muhammad Sajidur Rahman, Washington Garcia, Kevin Warren, Anurag Swarnim Yadav, Tom Shrimpton, and Patrick Traynor. Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 712–729. IEEE, 2021.

[2] N. Aggarwal and W.C. Karl. Line detection in images through regularized hough transform. *IEEE Transactions on Image Processing*, 15(3):582–591, 2006.

[3] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. Void: A fast and light voice liveness detection system. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2685–2702, 2020.

[4] Mark LG Althouse and Chein-I Chang. Image segmentation by local entropy methods. In *Proceedings., International Conference on Image Processing*, volume 3, pages 61–64. IEEE, 1995.

[5] Amazon. Amazon alexa, 2022. https://developer.amazon.com/en-US/alexa.

[6] Hangcheng Cao, Hongbo Jiang, Daibo Liu, Ruize Wang, Geyong Min, Jiangchuan Liu, Schahram Dustdar, and John C. S. Lui. Liveprobe: Exploring continuous voice liveness detection via phonemic energy response patterns. *IEEE Internet of Things Journal*, 10(8):7215–7228, 2023.

[7] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pages 183–195. IEEE, 2017.

[8] Peng Cheng and Utz Roedig. Personal voice assistant security and privacy—a survey. *Proceedings of the IEEE*, 110(4):476–507, 2022.

[9] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[11] J Lopes Esteves and Chaouki Kasmi. Remote and silent voice command injection on a smartphone through conducted iemi: Threats of smart iemi for information security. *Wireless Security Lab, French Network and Information Security Agency (ANSSI), Tech. Rep*, 2018.

[12] David Geerts. Comparing voice chat and text chat in a communication tool for interactive television. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 461–464, 2006.

[13] Ilias Giechaskiel and Kasper Rasmussen. Taxonomy and challenges of out-of-band signal injection attacks and defenses. *IEEE Communications Surveys & Tutorials*, 22(1):645–670, 2019.

[14] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. LSD: a Line Segment Detector. *Image Processing On Line*, 2:35–55, 2012.

[15] Jiwei Guan, Lei Pan, Chen Wang, Shui Yu, Longxiang Gao, and Xi Zheng. Trustworthy sensor fusion against inaudible command attacks in advanced driver-assistance systems. *IEEE Internet of Things Journal*, 2023.

[16] Jiwei Guan, Xi Zheng, Chen Wang, Yipeng Zhou, and Alireza Jolfaei. Robust sensor fusion algorithms against voice command attacks in autonomous vehicles. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 895–902, 2021.

[17] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. Accuth: Anti-spoofing voice authentication via accelerometer. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 637–650, 2022.

[18] Yitao He, Junyu Bian, Xinyu Tong, Zihui Qian, Wei Zhu, Xiaohua Tian, and Xinbing Wang. Canceling inaudible voice commands against voice control systems. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2019.

[19] Ryo Iijima, Tatsuya Takehisa, and Tatsuya Mori. Cyber-physical firewall: monitoring and controlling the threats caused by malicious analog signals. In *Proceedings of the 19th ACM International Conference on Computing Frontiers*, pages 296–304, 2022.

[20] Yiqi Jia. Far-field mems microphone array beamforming–measurements, simulations, and design. 2020.

[21] Chaouki Kasmi and Jose Lopes Esteves. Iemi threats for information security: Remote command injection on modern smartphones. *IEEE Transactions on Electromagnetic Compatibility*, 57(6):1752–1755, 2015.

[22] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 313–317, 2019.

[23] Denis Foo Kune, John Backes, Shane S Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. Ghost talk: Mitigating emi signal injection attacks against analog sensors. In *2013 IEEE Symposium on Security and Privacy*, pages 145–159. IEEE, 2013.

[24] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 312–325, 2020.

[25] Xinfeng Li, Xiaoyu Ji, Chen Yan, Chaohao Li, Yichen Li, Zhenning Zhang, and Weyuan Xu. Learning normality is enough: A software-based mitigation against inaudible voice attacks. In *32th USENIX Security Symposium (USENIX Security 20)*, 2023.

[26] Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1884–1899, 2021.

[27] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. Wavoid: Robust and secure multimodal user identification via mmwave-voice mechanism. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15, 2023.

[28] Tiantian Liu, Feng Lin, Zhangsen Wang, Chao Wang, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. Magbackdoor: Beware of your loudspeaker as a backdoor for magnetic injection attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 3416–3431. IEEE, 2023.

[29] Yan Meng, Jiachun Li, Matthew Pillari, Arjun Deopujari, Liam Brennan, Hafsah Shamsie, Haojin Zhu, and Yuan Tian. Your microphone array retains your identity: A robust voice liveness detection system for smart speakers. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1077–1094, 2022.

[30] Tao Ni, Yongliang Chen, Weitao Xu, Lei Xue, and Qingchuan Zhao. Xporter: A study of the multi-port charger security on privacy leakage and voice injection. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2023.

[31] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210, 2015.

[32] Pytorch. Pytorch mobile: End-to-end workflow from training to deployment for ios and android mobile devices, 2024. https://pytorch.org/mobile/android/#quickstart-with-a-helloworld-example.

[33] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology-heart and circulatory physiology*, 278(6):H2039–H2049, 2000.

[34] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[35] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017.

[36] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The {Long-Range} attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560, 2018.

[37] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[38] Yasser Shoukry, Paul Martin, Yair Yona, Suhas Diggavi, and Mani Srivastava. Pycra: Physical challenge-response authentication for active sensors under spoofing attacks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1004–1015, 2015.

[39] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2631–2648, 2020.

[40] Yazhou Tu, Sara Rampazzi, Bin Hao, Angel Rodriguez, Kevin Fu, and Xiali Hei. Trick or heat? manipulating critical temperature-based control systems using rectification attacks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2301–2315, 2019.

[41] Yazhou Tu, Vijay Srinivas Tida, Zhongqi Pan, and Xiali Hei. Transduction shield: A low-complexity method to detect and correct the effects of emi injection attacks on sensors. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 901–915, 2021.

[42] Arthur HM Van Roermund, Andrea Baschirotto, and Michiel Steyaert. *Nyquist AD converters, sensor interfaces, and robustness: advances in analog circuit design, 2012*. Springer Science & Business Media, 2012.

[43] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2062–2070. IEEE, 2019.

[44] Zhifei Xu, Runbing Hua, Jack Juang, Shengxuan Xia, Jun Fan, and Chulsoon Hwang. Inaudible attack on smart speakers with intentional electromagnetic interference. *IEEE Transactions on Microwave Theory and Techniques*, 69(5):2642–2650, 2021.

[45] Zhifei Xu, Runbing Hua, Jack Juang, Shengxuan Xia, Jun Fan, and Chulsoon Hwang. Inaudible attack on smart speakers with intentional electromagnetic interference. *IEEE Transactions on Microwave Theory and Techniques*, 69(5):2642–2650, 2021.

[46] Zhijian Xu, Guoming Zhang, Xiaoyu Ji, and Wenyuan Xu. Evaluation and defense of light commands attacks against voice controllable systems in smart cars. *Noise & Vibration Worldwide*, 52(4-5):113–123, 2021.

[47] Chen Yan, Xiaoyu Ji, Kai Wang, Qinhong Jiang, Zizhi Jin, and Wenyuan Xu. A survey on voice assistant security: Attacks and countermeasures. *ACM Computing Surveys*, 55(4):1–36, 2022.

[48] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1215–1229, 2019.

[49] Xuejing Yuan, Yuxuan Chen, Aohui Wang, Kai Chen, Shengzhi Zhang, Heqing Huang, and Ian M Molloy. All your alexa are belong to us: A remote voice control attack against echo. In *2018 IEEE global communications conference (GLOBECOM)*, pages 1–6, 2018.

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, 2016.

[51] Aihua Zhang, Bin Yang, and Ling Huang. Feature extraction of eeg signals using power spectral entropy. In *2008 international conference on BioMedical engineering and informatics*, volume 2, pages 435–439. IEEE, 2008.

[52] Guoming Zhang, Xiaoyu Ji, Xinfeng Li, Gang Qu, and Wenyuan Xu. Eararray: Defending against dolphinattack via acoustic attenuation. In *NDSS*, 2021.

[53] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 103–117, 2017.

[54] Linghan Zhang, Sheng Tan, Yingying Chen, and Jie Yang. A continuous articulatory-gesture-based liveness detection for voice authentication on smart devices. *IEEE Internet of Things Journal*, 9(23):23320–23331, 2022.

[55] Linghan Zhang, Sheng Tan, Zi Wang, Yili Ren, Zhi Wang, and Jie Yang. Viblive: A continuous liveness detection for secure voice user interface in iot environment. In *Annual Computer Security Applications Conference*, pages 884–896, 2020.

[56] Youqian Zhang and Kasper Rasmussen. Detection of electromagnetic interference attacks on sensor systems. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 203–216, 2020.