EveGuard: Defeating Vibration-based Side-Channel Eavesdropping with Audio Adversarial Perturbations

Jung-Woo Chang*§, Ke Sun*†§, David Xia‡, Xinyu Zhang*, Farinaz Koushanfar*

*University of California San Diego   †University of Michigan   ‡University of Illinois Urbana-Champaign

Abstract—Vibrometry-based side channels pose a significant privacy risk, exploiting sensors like mmWave radars, light sensors, and accelerometers to detect vibrations from sound sources or proximate objects, enabling speech eavesdropping. Despite various proposed defenses, these involve costly hardware solutions with inherent physical limitations. This paper presents EveGuard, a software-driven defense framework that creates adversarial audio, protecting voice privacy from side channels without compromising human perception. We leverage the distinct sensing capabilities of side channels and traditional microphones—where side channels capture vibrations and microphones record changes in air pressure, resulting in different frequency responses. EveGuard first proposes a perturbation generator model (PGM) that effectively suppresses sensor-based eavesdropping while maintaining high audio quality. Second, to enable end-to-end training of PGM, we introduce a new domain translation task called Eve-GAN for inferring an eavesdropped signal from a given audio. We further apply few-shot learning to mitigate the data collection overhead for Eve-GAN training. Our extensive experiments show that EveGuard achieves a protection rate of more than 97% from audio classifiers and significantly hinders eavesdropped audio reconstruction. We further validate the performance of EveGuard across three adaptive attack mechanisms. We have conducted a user study to verify the perceptual quality of our perturbed audio.

1. Introduction

Loudspeakers are omnipresent in today's technology-based society. Their use extends beyond facilitating phone calls and video conferencing for the exchange of private information. They have been widely integrated into intelligent mobile and IoT devices, enhancing human-machine interaction through speech recognition. The associated use cases are anticipated to reach a market size of $150.68 billion by 2032 [52], [62]. As people increasingly rely on loudspeaker-equipped devices, voice privacy is becoming increasingly important.

Unfortunately, the diverse sensors in intelligent devices are imposing an alarming risk to voice privacy. Although

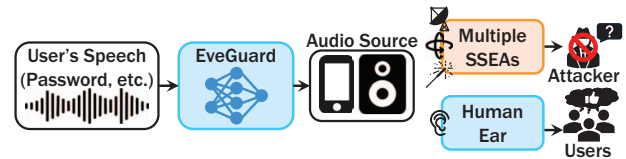§. The first two authors contributed equally to this work.

Figure 1: Overview of EveGuard, inserting imperceptible adversarial perturbations to the target speech to protect users' voice communication from multi-sensor eavesdropping attacks.

these sensors are not originally designed for voice recording, they can be repurposed by adversaries to serve as side channels to capture voice-induced vibrations, thereby facilitating unauthorized eavesdropping. For example, the prevalent accelerometers on smartphones have been exploited to eavesdrop on voice playout [25], [63]. Millimeter-wave (mmWave) radars can remotely detect vibrations from sound sources and recover speech signals through walls [23], [24], [59], [67], [68], [87]. Such side-channel speech eavesdropping attacks (SSEAs) lead to severe individual privacy breaches [1] and may compromise sensitive organizational intellectual property [15].

Existing research has devised hardware-based defenses against SSEAs. For instance, jamming-based methods [31], [64], [79] can block adversarial mmWave SSEAs. However, they may degrade the sensing function of legitimate mmWave devices. Moreover, jamming is generally prohibited in non-military applications [16]. Intelligent reflecting surface (IRS) has also been used as a security shield [58], [61], yet it can only protect its immediate vicinity. As for defending against accelerometer-based SSEAs, vibration motors have been used to generate low-amplitude vibrations that disrupt eavesdropping [86]. However, this method may cause user discomfort and hasten the depletion of smartphone batteries.

We propose EveGuard, an innovative software-driven defense mechanism to protect against privacy leakage from the loudspeaker-generated voice in SSEAs. As shown in Figure 1, EveGuard mitigates SSEA threats by introducing audio adversarial examples to the original audio signals prior to playback. EveGuard ensures that (i) the perturbed speech signals remain natural to human ears and microphones, and

(ii) any attempt by SSEAs to capture and reconstruct the perturbed speech will produce content that is difficult to interpret, both for humans and automated speech recognition systems. Note that EveGuard cannot protect voice from a human speaker when SSEAs target eavesdropping on throat vibrations, a challenge also for state-of-the-art (SOTA) attacks [66].

To attain these salient properties, EveGuard must address four main design challenges. First, it is crucial to ensure the effectiveness of perturbations against SSEAs while maintaining the quality of legitimate voice communication. Existing adversarial speech generation methods commonly rely on additive perturbations, which can introduce noticeable, conspicuous noise [84]. In contrast, EveGuard *leverages the distinct sensing mechanisms of the side channels versus traditional microphones or human hearing*, i.e., the former only captures low-frequency vibrations whereas the latter senses the subtle changes in air pressure. EveGuard devises a two-stage *Perturbation Generator Model* (PGM) that integrates: (1) finite impulse response (FIR) convolution to perturb the low-frequency attributes of speech while preserving the speech quality and (2) inaudible low-frequency adversarial perturbations (LFAPs) to corrupt the eavesdropped signals.

Second, to automate and optimize the perturbation signal generation, EveGuard requires a new differentiable computational model to represent the SSEA. To tackle this challenge, we propose Eve-GAN, a deep generative network aiming at learning an audio-to-SSEA translation that can map the source audio to the targeted eavesdropped audio. Once trained, Eve-GAN serves as a differentiable layer, enabling end-to-end training of our PGM. Yet training Eve-GAN requires collecting sufficient SSEA samples across various attack scenarios. Additionally, obtaining paired training data is tedious as it requires input-output pairs with the same speaker, speech attributes (e.g., prosody), and utterance content. To address this, we leverage advancements in few-shot unsupervised learning [42]. We propose a few-shot, unpaired audio-to-SSEA translation, which learns to convert source audio into eavesdropped audio by referencing an unpaired SSEA sample. By extracting domain features from the few-shot real-world SSEA samples, Eve-GAN facilitates a generalizable conversion applicable to unseen samples during training.

Third, the rapid growth of ML empowers attackers to devise sophisticated SSEAs [24], [25], [59], [63]. For instance, the attacker can transcribe private conversations using speech recognition, and identify digits with audio classifiers. However, the defender has no prior knowledge of the SSEA model deployed by the eavesdropper. To overcome this hindrance, we utilize the transferability of adversarial examples, which means perturbations learned to fool an ensemble of diverse surrogate models can also be effective against unknown black-box models [8], [43]. To this end, we first build a set of surrogate ML models based on multiple hypothetical SSEAs. We then concatenate the PGM with Eve-GAN and ensemble surrogate models to encourage the PGM to learn robust perturbations in an end-to-end manner.

Finally, an adaptive attacker who knows the existence of EveGuard may attempt to mitigate the effects of the perturbation. Thus, we apply three preventive techniques to PGM as follows: (1) the use of a discriminator inside the PGM to enforce undetectable constraints, (2) style diversification by integrating VAE-GAN [22] into FIR perturbation generator, and (3) ensuring the LFAP generator uses a random latent vector as input to produce diverse LFAP samples.

We implement EveGuard by integrating the above solutions. To evaluate EveGuard, we reproduce white-box SSEAs based on representative eavesdropping sensors, mmWave radar, accelerometer, and optical sensor. Built upon these, we extensively validate EveGuard under various attack settings including distance, orientation, materials, hardware configurations, etc. Our experimental results show that EveGuard achieves a protection rate of more than 97% from SSEA's digit classifiers and hinders the recovery of eavesdropped audio with an MCD (Mel-Cepstral Distortion) of over 13.4, and a WER (Word Error Rate) of over 68.2%. To show that our adversarial audio generated by PGM is imperceptible to humans, we verify the indistinguishability through a user study involving 24 participants.

The main contributions of EveGuard are as follows.

- We introduce EveGuard, a novel software-driven defense framework that leverages black-box adversarial examples to protect loudspeaker-generated voice from SSEAs.
- We design a PGM that leverages the unique features of eavesdropping devices to ensure robust perturbations across diverse attack scenarios, including variations in distance, orientation, materials, and hardware configurations.
- We develop Eve-GAN, a differential framework that enables our PGM to learn the distribution of adversarial examples end-to-end, incorporating a few-shot, unpaired audio-to-SSEA translation framework to reduce data collection overhead for training Eve-GAN.
- We perform extensive experiments to verify the effectiveness of the EveGuard defense, using both objective metrics and subjective user studies. Audio samples are available at https://eveguard.github.io/demo/.

2. Background and Related Work

In this section, we provide an overview of vibration-based SSEAs and related work.

2.1. Vibration-based Speech Eavesdropping

Speech can be recovered by measuring sound-induced vibrations on objects using sensors or vibration-sensitive devices. Table 1 provides an overview of conventional SSEA methods, summarizing three representative types of SSEA techniques and associated defense mechanisms.

Motion Sensor-based SSEA. Recent research [25], [63] shows that malicious apps can capture sound-induced motion signals using a smartphone's accelerometer. Using pretrained ML models [25], [63], the attacker can recover speech from the vibration motion despite the IMU's low sampling rates (≤ 500 Hz). While a smartphone's vibration

TABLE 1: Comparison with conventional vibration-based SSEAs. Existing defenders must rely on inefficient hardware-based techniques to disable each SSEA. (✓: the item is supported; ✗: the item is not supported.)

Previous Work	Sensor	Sensing Target	Sampling Rate	Sensing Distance	Non-Invasive	Through-Wall (Opaque)	Aided by ML	Existing Defense Solution	Our Defense (EveGuard)
Lamphone [46]	Optical Sensor	Loudspeaker or Vibrating Object	≤ 2kHz	Far (≃ 35m)	✓	✗	✗	Blocking Visible Channel (e.g., Wood, Curtains, etc.)	✓
LidarPhone [54]				Moderate (≃ 1.5m)	✗	✗	✓		
StealthyIMU [63]	Motion Sensor	Loudspeaker	≤ 0.5kHz	Close (≤ 0.1m)	✗	✗	✓	Smartphone's Vibration Motor	✓
Accear [25]					✗	✗	✓		
mmSpy [4]	mmWave Radar	Loudspeaker or Vibrating Object	≤16kHz	Moderate (≃ 5m)	✓	✓	✓	Jamming [31], [79] or IRS [58], [61]	✓
mmEcho [23]					✓	✓	✓		
Shi et al. [59]					✓	✓	✗		
VibSpeech [68]					✓	✓	✓		

motor [86] can obfuscate the sound-induced motion, it may cause battery drain and user discomfort.

Optical Sensor-based SSEA. Sound-induced vibration can also be captured through optical sensors. Lamphone [46] measures the vibration of a light bulb near a loudspeaker, while LidarPhone [54] uses a vacuum cleaner's lidar sensor to sample vibrations. However, these attacks leave visual clues (e.g., laser dots, bulky cameras) and are easily prevented by blocking the line-of-sight [59], [66], [67]. Therefore, a recent SoK deems these attacks impractical [66].

mmWave Radar-based SSEA. mmWave radars are commonly used for object ranging or imaging. Yet, slight object displacements produce subtle phase changes in the reflected radar signals, enabling the reconstruction of audio signals from these changes [23], [24], [59], [68]. Although some hardware-enabled defense techniques [31], [58], [61], [79] exist, like jamming and IRS shield, these approaches typically provide only a limited area of defense and often fall short in effectiveness. For instance, IRS-based approaches [58], [61] generally require individuals targeted by SSEAs to procure and deploy extra hardware. The IRS device needs to be oriented toward the attacker and placed close (e.g., 1.2m [58]) to the attacker's mmWave radar. Additionally, these defense mechanisms may unintentionally compromise the standard sensing capabilities of radars.

Among the aforementioned side channels, mmWave-based attacks pose the most serious threat to the user's privacy because: (1) radar can cover the human-voice frequency spectrum with a high sampling rate [23], (2) radar allows attacks at a distance [59], and (3) radar can penetrate soundproofing materials [23], [59]. *Therefore, Eve-Guard prioritizes a defense against mmWave radar-based attacks, yet we will also show that EveGuard remains adaptable to other side-channel threats, i.e., optical-based and accelerometer-based SSEA (Sec. 6.5).* Unlike existing defense mechanisms [58], [61], [79], [83], EveGuard is a software-driven framework designed to generate adversarial speech before playback, without affecting either microphone recordings or human hearing.

2.2. Adversarial Examples in Audio Domain

Adversarial examples pose a significant threat to ML systems, affecting audio-based systems like automatic speech recognition (ASR). Carlini *et al.* [9] introduced an end-to-end white-box attack where the ML model is manipulated to translate speech signals into the attacker's desired phrase. Recent studies [11], [12], [21], [88] focus on

achieving over-the-air delivery of adversarial audio. Semantic perturbation approaches [47], [84] have also emerged that deviate from additive perturbations. SMACK [84] devised semantic audio attacks targeting speech transcription and speaker recognition systems. Voiceblock [47] applies a time-varying FIR filter to outgoing audio, enabling effective and inconspicuous perturbations. While existing work addressed speech and speaker recognition systems, EveGuard focuses on the distinct vibrometry-based side-channel attacks.

2.3. Adversarial Examples for Privacy Protection

The rapid progress in ML has enabled attackers to misuse ML models for malicious purposes, such as device fingerprinting attacks, DeepFake audio generation, and unauthorized face and speaker recognition. In response, researchers have developed defensive strategies utilizing adversarial examples [13], [38], [56], [57], [85]. For instance, iPET [57] protects the privacy of IoT users by perturbing network traffic to prevent fingerprinting attacks. Antifake [85] specifically targets DeepFake by adding perturbations to the user's speech, disrupting the speech synthesis process. Fawkes [56] inserts imperceptible pixel-level perturbations into the user's photo to thwart unauthorized facial recognition from learning the user's identity. EveGuard closely aligns with the goal of these studies. However, protecting users' voice privacy from side-channel eavesdroppers through audio perturbation is a new research field.

3. Threat Model, and Defense Goals

In this section, we introduce the threat model and defender's objectives and capability.

3.1. EveGuard Threat Model

The attacker's goal is to eavesdrop on the audio content played by a loudspeaker. However, direct microphone deployment inside or outside the victim's room is infeasible due to the following reasons: (a) hidden microphones inside the room can be easily detected by victims using spy detectors [44], [89], making covert placement ineffective and (b) if victims use loudspeakers in a soundproof room, external eavesdropping microphones cannot capture the audio. As a result, the attacker seeks to exploit SSEAs, leveraging side-channel sensors to capture speech-induced vibrations from the loudspeakers. Specifically, mmWave radar-based techniques [23], [59], [74] enable the attacker to penetrate
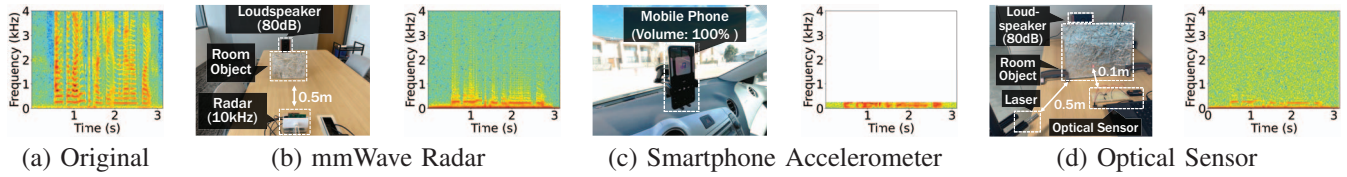
Figure 2: Feasibility study settings and results (The sampling rate of smartphone's accelerometer is 500 Hz).

soundproofing materials and eavesdrop on speech. Additionally, zero-permission motion sensors [63] can be exploited to infer private speech information from a smartphone. The attacker may also detect speech-induced vibrations by sensing reflected ambient light [46].

In this work, rather than limiting our focus to constrained SSEAs, we aim to develop defenses capable of mitigating even the worst-case scenarios, ensuring robustness against highly capable adversaries. We assume that the attacker has strong capabilities to i). compromise the side-channel sensors and their sensing systems; ii). employ advanced ML and signal processing techniques to derive private information from the captured signals; iii). gain access to the victims' speech samples or even create training datasets to facilitate the eavesdropping.

Specifically, for mmWave radar-based SSEA, the attacker can access the radar's raw Analog-to-Digital Converter (ADC) data, by either employing their radar or planting malware into a radar-equipped IoT device near the victim [73]. Moreover, the attacker possesses knowledge of the victim's room layout, enabling them to isolate sound vibrations not just from the speaker but also from everyday objects (e.g., chip bags, etc.) near the loudspeaker. Additionally, the attacker can acquire speech samples of the victim from publicly available sources. Prior to initiating the attack, they might physically access the intended attack environment, play the victim's audio samples through a loudspeaker, and gather a dataset of eavesdropping signals with mmWave radar.

For optical sensor-based SSEA, consistent with recently proposed attacks [46], [54], we assume the attacker can deploy their laser transmitter towards the reverberator without line-of-sight obstruction. The optical sensor captures the reflected laser, and the attacker accesses the ADC data, following similar assumptions as in mmWave-based SSEAs.

For the motion sensor-based SSEA, we assume that the attacker can trick the victim into installing a malicious app, which collects motion signals in the background and can even stream the data to the adversaries' server. Note that detecting the malicious app is a challenging problem [80], as attackers can effectively disguise it among legitimate apps, allowing them to bypass security measures.

3.2. Defense Objectives

Design Goal. EveGuard introduces adversarial examples to the original audio signals prior to playback, aiming to protect loudspeaker-generated voice from SSEAs with minimal impact on the intelligible voice quality. Note that EveGuard cannot protect voice from a human speaker when SSEAs

target eavesdropping on throat vibrations, a challenge also for existing SOTA attacks [66]. To this end, EveGuard must meet five criteria. First, the adversarial voice audio generated by EveGuard should prevent the attacker from restoring audible and intelligent speech. Neither humans nor ML models should be able to transcribe the recovered audio into words and sentences. Second, the perturbation should be imperceptible to humans, and the perturbed speech audio must remain high quality. Third, since the defender does not know the attacker's ML models and attack scenario (e.g., audio volume, distance), the generated adversarial perturbations should be effective regardless of black-box knowledge. Fourth, EveGuard should be robust against adaptive attacks who know the presence of adversarial perturbations. EveGuard needs to enforce that perturbations are undetectable to attackers, leading to failure in attackers' attempts to remove perturbations from eavesdropped audio. Finally, EveGuard should be applicable to both offline and online scenarios. In offline scenarios, such as intelligent speakers delivering private content to users, EveGuard runs without latency limitations on the user's device. For challenging online scenarios, such as VoIP, EveGuard must meet low-latency requirements (e.g., $\leq$ 150ms for real-time VoIP communication [55]).

Defender's Capability. To achieve the aforementioned design goals, EveGuard employs black-box perturbations to speech signals prior to playback, aiming to protect voice privacy from the sound source, i.e., loudspeakers, against side-channel eavesdropping. We assume that EveGuard has access to the input audio of the voice communication device (i.e., loudspeaker), and can directly convert the input audio into adversarial audio before playing out. The defender follows black-box settings where he/she has no knowledge about the attack model (e.g., ML model and parameters) and scenario (e.g., distance, audio volume, etc.).

4. Preliminary Study

In this section, we investigate the fundamental differences between air-pressure-based sound-capturing methods (e.g., microphones and human hearing) and vibrometry-based side-channel attacks. We further conduct a preliminary study to understand how to leverage these insights in the development of EveGuard's PGM.

4.1. Understanding Side-Channel Attacks

Microphones convert air pressure variations into electrical signals using a diaphragm, while vibrometry-based side

channels measure the vibration displacement or acceleration of physical objects. Due to these different mechanisms, *microphones can detect sounds across the entire audible frequency range, while the SNR of vibrometry-based sensors drops sharply at higher frequencies.* More specifically: i). The vibration displacement of a speaker diaphragm is approximately inversely proportional to the sound frequency to maintain consistent sound pressure across different frequencies, making it harder for side-channel sensors to detect high-frequency vibrations [35], [45]. ii). Sound waves experience higher attenuation at higher frequencies when traveling through structural materials or air [35]. Although accelerometers can theoretically capture high-frequency sounds by measuring vibration acceleration, the actual frequency response is diminished due to structural propagation loss. iii). Side-channel sensors are not designed for precise sound recording, typically having limited sampling rates and sensitivity to high frequencies. For instance, the maximum sampling rate of a smartphone accelerometer is about 500 Hz, whereas capturing full-band speech signals requires at least 8 kHz [25].

The vibrometry-based side channels share consistent frequency response characteristics due to the aforementioned inherent limitations, which can be leveraged by the EveGuard defender. To validate this observation, we conduct experiments using three representative side channels: a COTS mmWave radar, an accelerometer embedded in a smartphone, and a laser microphone. Specifically, following the SSEA in [23], [59], we use the TI IWR1843-Boost radar [28] to sense speech-induced vibrations from an everyday object (i.e., tinfoil) while a loudspeaker (i.e., Edifier R1700BTs) plays acoustic signals. The chirp rate of the mmWave sensor is set to 10kHz. Figure 2 illustrates our basic experimental setup. Through a range-FFT operation, we isolate phase changes at the tinfoil's location in the mmWave data, and then apply a short-time Fourier transform (STFT) to these data points. For the accelerometer SSEA, we follow [25], [63] and mount a smartphone (i.e., LG V50) in a car phone holder, as shown in Figure 2(c). We then collect accelerometer data at the device's maximum sampling rate (i.e., 500 Hz). To build an optical-based SSEA, we aim a laser beam at the tinfoil vibrated by the loudspeaker, as shown in Figure 2(d). Since the optical sensor converts the intensity of the laser reflected from the tinfoil into an electrical signal, we can effectively extract audio information [46]. For all the experiments, we employ two types of audio signals: a 3-second clip from the Librispeech corpus [49] to visualize eavesdropping signals and sweep tones ranging from 50 Hz to 4 kHz to analyze frequency response.

mmWave radar. We compare the spectrograms of the original and the radar-reconstructed speech in Figure 2(a), 2(b). We observe that the reconstructed speech signals maintain a high-frequency response in the low- and mid-frequency bands (i.e., < 1 kHz). However, the SNR tends to diminish beyond these frequencies, down to nearly 0 dB for frequencies above 2 kHz. This drop highlights the radar's limitations in detecting high-frequency vibrations. We further confirm
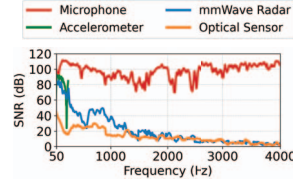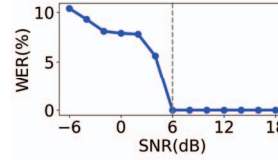

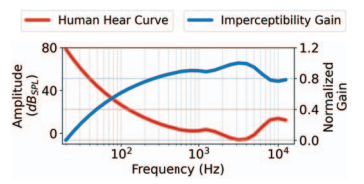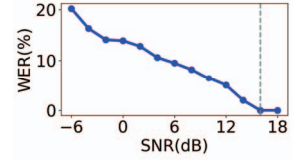Figure 3: The frequency responses of different side channels.

Figure 4: Equal-soundness contour and imperceptibility gain curve.

(a) AWGN (under 500Hz) (b) AWGN (under 1000Hz)

Figure 5: Impact of AWGN with different frequency bands on human speech understanding.

that variations in the radar's sampling rate (see Figure 3) or different attack scenarios (see Appendix A) have little impact on its frequency response, suggesting these characteristics can be reliably utilized by EveGuard to defend against various SSEA radar hardware configurations and attack scenarios.

Accelerometer. According to the Nyquist sampling theorem, an accelerometer with a sampling rate of 500 Hz can capture data only up to 250 Hz. As shown in Figure 2(c) and Figure 3, the audio reconstructed by the accelerometer is similar to the original audio in low-frequency components (i.e., < 250 Hz). However, due to the accelerometer's limited sampling rates, the raw vibration signal loses mid- and high-frequency components (i.e., > 250 Hz). The lost speech spectrum can be recovered by the audio enhancement, as shown in Figure 10(c).

Laser. As shown in Figure 2(d) and Figure 3, optical-based SSEA can recover wideband intelligible speech, but its sensing capability is worse than mmWave radar in most frequency bands. The frequency response can be enhanced with a high-end laser vibrometer (LV-FS01 [68]), but this makes the attack device bulky and more easily identifiable.

4.2. Characterizing Human Hearing

We then investigate human auditory sensitivity across different frequency ranges to devise an undetectable defense. Our study is based on equal-loudness contour [2], a well-established model that delineates the sound pressure level (SPL) perceived by the human ear across the frequency bands. We convert SPL measurements from $-20dB_{SPL}$ to $80dB_{SPL}$ into a normalized scale from 0 to 1 to visualize human insensitivity to different frequency bands. Figure 4 shows the hearing curve and the imperceptibility gain obtained through psychoacoustic experiments to describe the human ear's sensitivity [2]. The hearing curve represents the amplitude required for a purely continuous tone of a certain frequency that humans can hear. Frequencies with
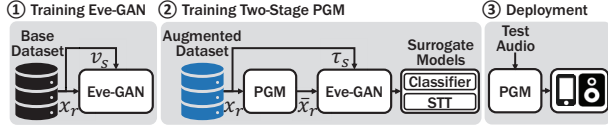
Figure 6: EveGuard training and deployment process.

high imperceptibility are harder to perceive by human. It shows that human ears are most sensitive to frequencies between 1.6 kHz and 4 kHz, with a marked insensitivity to frequencies below 500 Hz. Furthermore, we conducted an experiment to investigate the effects of low- and mid-frequency noise on human speech understanding. We asked 24 participants to listen to noisy audio containing either low-frequency or mid-frequency AWGN and then translate sentences. Please refer to Sec. 6.6 for a detailed discussion of the ethical considerations in our study. We calculated the WER based on their translations. From Fig 5(a), we observe that at SNRs above 6dB, low-frequency noise does not impact human understanding. However, as shown in Fig 5(b), for participants to accurately hear audio containing mid-frequency noise, the SNR must be over 16dB. Furthermore, SNR and defense success rate have an inverse relationship (see Sec. 6.7). *Thus, to guarantee the sound quality to the human ear, EveGuard can mainly generate perturbations within the low-frequency bands (i.e., < 500 Hz) while minimally affecting the mid- and high-frequency ranges (i.e., > 500 Hz).*

5. Design of EveGuard

In this section, we present the optimization problem and methodology for designing EveGuard. An overview of the EveGuard system and its modules is provided in Figure 6.

5.1. Problem Formulation

EveGuard aims to automate and optimize robust audio adversarial perturbations to protect loudspeaker-generated voice from SSEAs while maintaining the quality of legitimate voice communication. Suppose an attacker eavesdrops on a n-dimensional victim's speech $x_r \in [-1, 1]^n$ with sampling rate r and reconstruct a waveform $\mathcal{A}_s(x_r, \zeta_s)$ via his/her audio reconstruction method $\mathcal{A}_s$, where ζ_s is a vector representing the attack scenarios (e.g., distance, orientation, materials, hardware configurations, and loudspeaker's volume, etc.) for a specific sensor $s \in \{\text{side channel sensors}\}$. The attacker then uses the reconstructed audio as input to the following ML models: (1) a speech recognition model $M_s^{sr}(\cdot)$ that converts audio into text transcriptions and (2) an audio classifier $M_s^{ac}(\cdot)$ that identifies specific digits or keywords. Note that ML models have different model parameters depending on the sensor type s. EveGuard applies adversarial perturbations to prevent eavesdroppers from recovering audible speech. The objective is to find a minimal perturbation δ as follows:

$$\begin{aligned} \arg\max_{\delta} \quad & \mathbb{E}_{x_r, \zeta_s} \left[\mathcal{L}_{qd}(\mathcal{A}_s(x_r + \delta, \zeta_s), \mathcal{A}_s(x_r, \zeta_s)) \right] - \alpha \|\delta\|_2, \\ \text{subject to} \quad & WER(M_s^{sr}(\mathcal{A}_s(x_r + \delta, \zeta_s)), y_{sr}) > t_{sr}, \\ & M_s^{ac}(\mathcal{A}_s(x_r + \delta, \zeta_s)) \neq y_{ac}, \end{aligned} \quad (1)$$

where $\mathcal{L}_{qd}$ is the loss that measures the quality difference between the eavesdropping results from clean and perturbed audio. $WER(\cdot)$ (Word Error Rate) is a metric to assess speech recognition [70]. It calculates accuracy by dividing the number of errors by the total number of words in the reference y_{sr}. t_{sr} is a threshold that determines the success of our defense. y_{ac} is a label for audio classification. α is a hyper-parameter that controls the relative importance of imperceptibility of δ and defense performance, respectively.

The key challenge of EveGuard is how to automate and optimize the adversarial perturbations δ to solve Eq. 1. Specifically, EveGuard needs to address four challenges. Firstly, it must model the audio reconstruction $\mathcal{A}_s$ without laborious data collection across numerous eavesdropper hardware configurations and audio profiles. Secondly, side-channel eavesdropping characteristics must be considered when modeling δ. Otherwise, the optimization may become stuck in local optima. Thirdly, the EveGuard defender has no knowledge of the ML model and parameters used by the eavesdropper (i.e., $M_s^{sr}(\cdot)$ and $M_s^{ac}(\cdot)$), and cannot even perform black-box queries. Finally, EveGuard must be immune to adaptive attackers who attempt to learn the perturbation and denoise it from the eavesdropped speech.

5.2. Overview of EveGuard

To address them, we designed EveGuard to train and deploy end-to-end, comprising three major phases, as illustrated in Figure 6.

Phase #1 - Training Eve-GAN (Sec. 5.3). To achieve automatic optimization of these perturbations, we model the SSEA audio reconstruction process $\mathcal{A}_s(x_r, \zeta_s)$ within a differentiable framework, allowing PGM to learn the distribution of adversarial examples end-to-end. Specifically, we design a deep generative network called Eve-GAN to convert audio signals into eavesdropped data. To ensure generalization and reduce data collection effort, we propose a few-shot audio-to-SSEA translator that trains with a base dataset consisting only of unpaired audio-SSEA data.

Phase #2 - Training PGM (Sec. 5.4). Then, we aim to train the PGM using a set of surrogate models to enhance the transferability of adversarial examples. We concatenate the PGM with the few-shot translator and surrogate models (i.e., $M_s^{sr}(\cdot)$ and $M_s^{ac}(\cdot)$), setting all ML modules except PGM as non-trainable to allow end-to-end gradient backpropagation. To ensure scenario-agnostic perturbations resilient to variations in SSEA scenarios ζ_s, we use the pretrained few-shot audio-to-SSEA translator in the first phase to augment the base dataset by generating eavesdropping signals in unseen SSEA domains. Finally, we optimize the PGM to minimize the intelligibility of audio in the augmented SSEA domain.

Phase #3 - Deployment. Once the PGM's training process is complete, we place the trained PGM prior to the audio
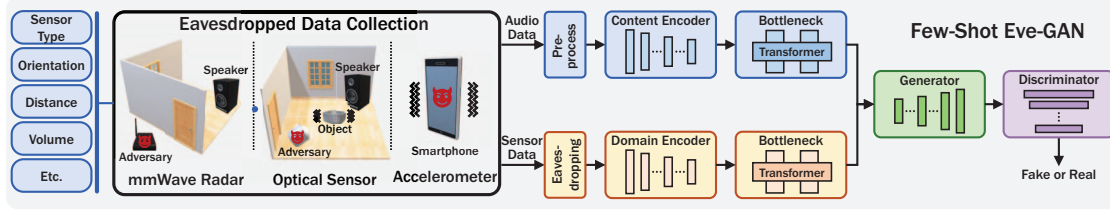
4539

Figure 7: The general workflow of few-shot Eve-GAN consists of a few-shot audio translator and discriminator.

sources to convert the original audio into perturbed audio in the audio front-end processing pipeline. EveGuard operates within acceptable latencies for Voice-over-IP (VoIP) applications, making it suitable for real-time VoIP communications (see Sec. 6.8).

5.3. Modeling of Few-Shot Eve-GAN

We first devise Eve-GAN (Figure 7), a deep generative model that establishes a non-linear relationship between the original audio and the eavesdropped audio (i.e, $\mathcal{A}_s(x_r, \zeta_s)$) within a differentiable framework. Eve-GAN learns to extract generalizable style patterns that can be applied to unseen SSEA samples. It leverages few-shot unpaired audio-to-SSEA translation to alleviate the data collection overhead.
SSEA Data Collection. Our few-shot approach addresses two major issues: (i) obtaining paired training data and (ii) collecting SSEA samples under infinite attack scenarios. To achieve this, we construct a *base dataset* that facilitates the learning of generalizable translation capabilities. The base dataset generally requires diverse data samples [17], [42]. Thus, we collect data by considering several crucial factors that determine the SSEA sensors' vibration sensing capability [23], [24], [25], [59], [63]. We use mmWave-based SSEA as an example for illustration and include the details of other SSEA in Table 3 and Table 4. As shown in Table 2, we thoroughly consider a total of $72 = 2^3 \times 3^2$ attack scenarios for mmWave radar, respectively. Within this base dataset, we collect 120 samples per scenario for mmWave radar, ensuring a comprehensive representation of potential attack vectors. Similarly, we construct base datasets from accelerometers and optical sensors, as shown in Table 3 and Table 4 and collect 450~500 samples per scenario.

After the few-shot Eve-GAN is trained on the base dataset, we can convert original audio into eavesdropping signals in an unseen SSEA scenario. We leverage this few-shot capability to aid the PGM in learning the scenario-invariant perturbation distribution with minimal additional SSEA samples. To this end, we create a *few-shot dataset* consisting of one sample per unseen scenario that is not included in the base dataset, as shown in Table 2. Then, we integrate base and few-shot datasets into an augmented dataset for PGM training, as depicted in Figure 6.
Few-Shot Audio Translator. Next, we use the *base dataset* to train a few-shot translator $\hat{x}_{r,s} = T_{r,s}(x_r, v_s)$ that transforms original audio x_r into eavesdropped audio $\hat{x}_{r,s} \in [-1,1]^n$ with the domain of a given SSEA example $v_s \in [-1,1]^n$. Note that the sampling rate of $\hat{x}_{r,s}$ is not the same as that of v_s. The few-shot translator consists of several modules. The pre-processing module resamples x_r to a sampling rate of v_s through the differentiable resampling operation [82] and applies zero-mean normalization to the data. The content encoder, comprised of 1-D convolutional (Conv1d) layers, maps the human speech to a content latent code. The domain encoder consists of a stack of Conv1d layers to produce SSEA information. We introduce the bottleneck extractor to refine the representation. Then, the decoder has several 1-D adaptive instance normalization (AdaIN) residual blocks [26] followed by upscale Conv1d layers. By feeding the content and SSEA latent codes to the decoder, we ensure that the reference SSEA sample v_s controls the output domain while the victim's speech determines the utterance content.
Discriminator. Our goal is to make $\hat{x}_{r,s} = T_{r,s}(x_r, v_s)$ close to the real eavesdropped audio. To this end, we adopt an adversarial training method where a discriminator $D_{r,s}^e$ learns to distinguish between real-world eavesdropped audio and fake audio generated by $T_{r,s}$. We adopt the multi-period discriminator in [37].
Training Loss. We train the proposed few-shot Eve-GAN by solving a minimax optimization problem given by:

$$\max_{D_{r,s}^e} \min_{T_{r,s}} \mathcal{L}_{gan}(T_{r,s}, D_{r,s}^e) + \beta_{con}\mathcal{L}_{con}(T_{r,s}) + \beta_{fm}\mathcal{L}_{fm}(T_{r,s}), \quad (2)$$

where β_{con} and β_{fm} are hyper-parameters for each term.

TABLE 2: Defender's dataset settings for mmWave radar-based SSEAs, where Loud$_1$ refers to Logitech Z313. V, O, and R denote voice source, reverberating object, and mmWave radar, respectively.

Dataset	Voice Source	Material	V-to-O Distance	R-to-O Distance	R-to-O Angle	Audio Volume
Base	Loud$_1$	tinfoil chip bag carton	0.5m 1.5m	0.5m 1.5m	$-15°$ $0°$ $15°$	70dB 80dB
Few-Shot	Loud$_1$	plastic cotton paper	0.5m 1.5m	0.5m 1.5m	$-15°$ $0°$ $15°$	70dB 80dB

TABLE 3: Defender's dataset settings for IMU sensor (i.e., Accelerometer)-based SSEAs.

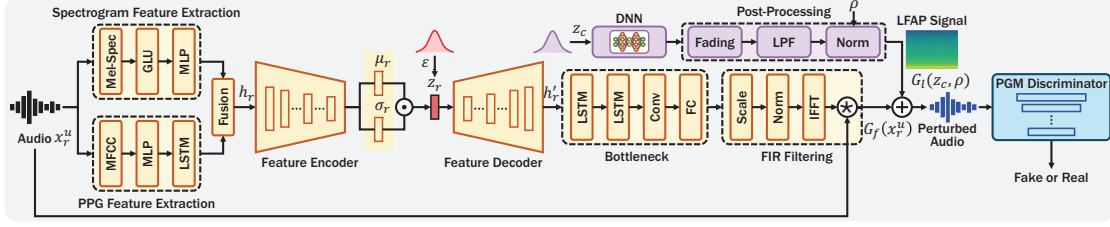| Dataset	Smartphone Model	Sampling Rate	Surface of Placement	Audio Volume
Base	Samsung S20	200Hz 500Hz	table sofa floor	60% 80% 100%
Few-Shot	Samsung S20	200Hz 500Hz	Phone holder handhold bed	60% 80% 100%

Figure 8: Visualization of PGM. Yellow modules form the FIR generator; purple modules form the LFAP generator.

$\mathcal{L}_{gan}$, $\mathcal{L}_{con}$, and $\mathcal{L}_{fm}$ are the GAN loss [19], the consistency loss [90], and the feature matching loss [39]. We define each loss function as:

- **GAN Loss.** We obtain the GAN loss as:

$$\mathcal{L}_{gan} = \underset{v_s}{\mathbb{E}}[\log D^e_{r,s}(v_s)] + \underset{x_r, v_s}{\mathbb{E}}[\log(1 - D^e_{r,s}(T_{r,s}(x_r, v_s)))]. \quad (3)$$

- **Consistency Loss.** The consistency loss encourages the model to preserve the properties of the original audio, ensuring that the content of the utterance is preserved. When the original audio is used on both inputs of $T_{r,s}$, the result should be identical to the input. We calculate the consistency loss as follows:

$$\mathcal{L}_{con} = \underset{x_r}{\mathbb{E}}[\|x_r - T_{r,s}(x_r, x_r)\|^1_1]. \quad (4)$$

- **Feature Matching Loss.** The feature matching loss improves the stability of the training and the quality of the translation outputs. To this end, we design a feature extractor $D^m_{r,s}$, which is a model excluding the last (prediction) layer of $D^e_{r,s}$. We compute the loss by extracting features from the translation output and reference SSEA example as:

$$\mathcal{L}_{fm} = \underset{x_r, v_s}{\mathbb{E}}[\|D^m_{r,s}(T_{r,s}(x_r, v_s)) - D^m_{r,s}(v_s)\|^1_1]. \quad (5)$$

Inference Stage. After the training is completed, the few-shot translator is used as a non-trainable differentiable layer, which helps the PGM to find robust perturbations. As shown in Figure 6, the PGM is located in front of the few-shot audio translator. By feeding a perturbed audio $\bar{x}_r$ to the few-shot translator, we enable cross-domain conversion as:

$$\tilde{x}_{r,s} = T_{r,s}(\bar{x}_r, \tau_s), \quad (6)$$

where τ_s is an SSEA example sampled from augmented datasets consisting of the base and few-shot datasets and $\tilde{x}_{r,s}$ is the domain conversion result of the perturbed audio $\bar{x}_r$. Note that the sampling rate of $\tilde{x}_{r,s}$ follows that of τ_s.

TABLE 4: Defender's dataset settings for optical sensor-based SSEAs.

Dataset	Laser-to-O Distance	Sensor-to-O Distance	Material	Audio Volume
Base	0.5m 1.0m 1.5m	0.05m 0.1m 0.15m	tinfoil chip bag	70dB 80dB
Few-Shot	0.5m 1.0m 1.5m	0.05m 0.1m 0.15m	plastic cotton paper	70dB 80dB

5.4. Modeling of PGM

With pretrained Eve-GAN, we can train the PGM end-to-end. Our optimization goal is to generate robust adversarial examples against SSEAs with minimal impact on human auditory perception. Guided by the unique characteristics of SSEA, we transform the original audio x_r into adversarial audio $\bar{x}_r$ using two principles: adversarial FIR filtering, which perturbs the audio in the frequency domain, and low-frequency adversarial perturbations (LFAPs), which are negligible to the human ear. We also build a set of surrogate models for multiple SSEAs to ensure that the adversarial examples have strong transferability through ensemble learning. This approach helps to generate generalizable perturbations despite the lack of knowledge about the attackers' speech-processing models. Additionally, we apply a robustness constraint to the optimization problem, making EveGuard unlearnable by adaptive eavesdroppers.

Figure 8 shows PGM's architecture, comprising three main modules: (1) an FIR generator G_f learns the perturbation distribution in the frequency domain to avoid noisy artifacts common in additive attacks; (2) an LFAP generator G_l uses low-frequency adversarial perturbations (< 500 Hz) to prevent the attacker from restoring speech, leveraging the unique frequency response characteristics of the sensors, and (3) a discriminator D^p_r is employed to distinguish whether the PGM generates real or fake audio, which helps to make $\bar{x}_r$ close to x_r to prevent an adaptive attacker from learning our perturbation patterns. To make EveGuard suitable for the audio streaming, PGM continuously processes x^u_r, which is a segment of x_r, where u denotes the segment index.

FIR Generator. Conventional FIR generators [47], [48] are primarily used to disrupt speaker recognition systems. However, these generators are designed for microphone recordings and are equally sensitive across the audible frequency range. Their formulations do not consider domain constraints for eavesdropping side channels, thus making them ineffective for SSEA side channels. Furthermore, their perturbation lacks variability as they do not enforce a robustness constraint. Thus, an adaptive attacker aware of the existence of defense can easily identify perturbations and then build a robust SSEA with adversarial training.

To overcome these deficiencies, we incorporate the VAE-GAN architecture [22] into the conventional FIR generator [47]. Instead of the FIR filter being fixed to the input, VAE-GAN enables G_f to produce diverse FIR filters. To this end, we first extract representative acoustic features, including spectrogram features [33], phonetic posteriorgrams

4541

(PPG) [53]. Then, the feature encoder maps acoustic features h_r extracted from input audio x_r^u to a latent vector z_r with a distribution $p(z_r|h_r)$. Here, $z_r = \mu_r + \sigma_r \odot \varepsilon$, where $\varepsilon \sim N(0, I)$. With the given latent vector z_r, the feature decoder restores h_r' via a distribution $q(h_r'|z_r)$. The restored features h_r' are sent as input to the bottleneck and decoded into a frequency-domain FIR filter. Lastly, the filtering module converts the FIR filter to the time domain and then performs time-varying filtering on the input audio.

LFAP Generator. G_l has a multi-layer Deep Neural Network (DNN) consisting of a stack of fully connected (FC) layers with ReLU activation except for the last layer with Tanh activation. Upon receiving a random latent vector z_c, the DNN produces a perturbation vector of length 512. G_l then proceeds with post-processing techniques to suppress the high-frequency components and the perturbation audibility. Specifically, an audio fading [6] ensures smooth transitions when connecting perturbation vectors to match the length of $G_f(x_r^u)$. We then apply biquad low-pass filter to suppress the high-frequency components with the cut-off frequency of 500 Hz. Lastly, we adopt a normalization process to limit the amplitude of the LFAP. Here, we set the Signal-to-Noise Ratio (SNR) as our normalization, which closely approximates the masking effect in the human auditory system [7]. We empirically find an optimal SNR ρ such that the source signal thoroughly dominates the perception of the perturbing signal while preventing SSEAs.

Training Loss. We denote the processes that generate adversarial audio as $\bar{x}_r^u = G_r(x_r^u, z_c, \rho) = G_f(x_r^u) + G_l(z_c, \rho)$. According to the optimization goal in Eq. 1, we formulate the following objectives to train G_r and D_r^p:

$$\max_{D_r^p} \min_{G_r} \quad \mathcal{L}_{adv}(G_r, D_r^p) + \lambda_{kl}\mathcal{L}_{kl}(G_r) \\ + \lambda_{ens}\mathcal{L}_{ens}(G_r) + \lambda_{rec}\mathcal{L}_{rec}(G_r), \quad (7)$$

where $\mathcal{L}_{adv}$, $\mathcal{L}_{kl}$, $\mathcal{L}_{ens}$, and $\mathcal{L}_{rec}$ are the GAN loss, the KL loss, the ensemble loss, and the reconstruction loss. λ_{kl}, λ_{ens}, and λ_{rec} are weight parameters. With this loss function, we iteratively train G_r and D_r^p until reaching equilibrium. We define each loss function as:

• **Adversarial Loss.** We define the adversarial loss as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x_r}[\log D_r^p(x_r)] + \mathbb{E}_{\bar{x}_r}[\log(1 - D_r^p(\bar{x}_r))]. \quad (8)$$

• **KL Loss.** We employ the KL loss to diminish the gap between the posterior distribution $p(z_r|h_r)$ and the prior distribution as:

$$\mathcal{L}_{kl} = \mathbb{E}_{h_r}[\mathbb{D}_{KL}(p(z_r|h_r)||N(0, I))], \quad (9)$$

where $\mathbb{D}_{KL}$ means the KL divergence. The prior is assumed to follow a multivariate normal distribution.

• **Ensemble Loss.** To subvert an eavesdropper's ML model, we build an ensemble of surrogate models with K configurations of SSEA for speech recognition and audio classifier. Then, the few-shot translator $T_{r,s}$ bridges the gap between
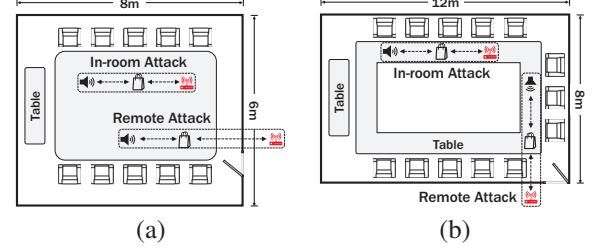

Figure 9: Layout of the experimental environment. (a) The room where the EveGuard is trained. (b) The room where the EveGuard evaluation is conducted.

the PGM and surrogate models to calculate $\mathcal{L}_{ens}$ as:

$$\mathcal{L}_{ens} = \mathbb{E}_{\bar{x}_r, \tau_s}[\sum_{k=1}^{K}(\log \Pr_s^k(y_{sr}|\tilde{x}_{r,s}) + Y_s^k(y_{ac}|\tilde{x}_{r,s}))], \quad (10)$$

where $\{\Pr_s^k(y_{sr}|\tilde{x}_{r,s})\}_{k=1}^{K}$ is a set of predicted probability that will be transcribed into y_{sr} following a Connectionist Temporal Classification (CTC) loss [20]. $\{Y_s^k(y_{ac}|\tilde{x}_{r,s})\}_{k=1}^{K}$ is a set of the probability belonging to y_{ac}. $\tilde{x}_{r,s}$ is obtained from Eq. 6. Note that the model parameters for speech recognition and audio classifier are different depending on sensor type s and ensemble index k.

• **Reconstruction Loss.** To ensure that our perturbations are undetectable to the human ear but cause failure of SSEA's audio construction, we calculate $\mathcal{L}_{rec}$ as:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_r, \bar{x}_r, \tau_s}[\|(\bar{x}_r - x_r)\|_1^1 - \mathcal{L}_{stft}(\tilde{x}_{r,s}, T_{r,s}(x_r, \tau_s))], \quad (11)$$

where $\mathcal{L}_{stft}$ is the multi-resolution STFT loss [77]. By adopting $\mathcal{L}_{stft}$, we maximize the spectral difference between the eavesdropping results for the original and perturbed audio. To make our perturbation inaudible, we use the mean absolute error to restrict the difference between original and perturbed audio.

6. Evaluation

6.1. SSEA Setup and Implementation

SSEA Scenario Setup. We re-implement vibration-based SSEAs using three representative side channels: mmWave radar, accelerometer and optical sensor. Figure 9 shows the layout of our experimental environment. We use the conference room depicted in Figure 9(a) to train EvGuard and the room in Figure 9(b) for its evaluation. Appendix Table 14 provides the detailed experimental setup.

We establish basic attack scenarios as discussed in Sec. 4.1 and Figure 2. We further assess EveGuard's performance not only in the basic attack scenarios but also in unseen scenarios (Sec. 6.4 and Sec. 6.5).

SSEA Implementation. Unlike previous SSEAs [23], [24], [25], [59], [63] that limit the attacker's capabilities, we consider a stronger threat model, assuming that the eavesdropper is powerful enough to collect SSEA data from the victim's room using the same equipment as the victim, as shown in Appendix Table 14. This allows us to assess

EveGuard in the worst-case scenario, while still integrating sophisticated attack techniques established in SOTA SSEAs. Our implementation of the SOTA SSEAs [23], [24], [25], [59], [63] comprises three parts.

1) **Signal Processing** (SP). We follow the steps discussed in Sec. 4.1 and [23] to pre-process the raw eavesdropping signals and improve reconstruction quality. Note that SP only applies to mmWave radar.

2) **Machine Learning** (ML). To enhance the probability of successful eavesdropping, we further implement an ML-based speech enhancement model by following well-established cGAN models [24], [25].

3) **Speech Recognition** (SR). The final step involves training dedicated speech recognition or audio classification models using the processed signals from SP and ML. It aims to extract explicit private information. The results of SP and ML in the evaluation are derived through SR.

The ML model architecture and implementation details are documented in Appendix B.

Datasets. Appendix C describes the speech datasets utilized for training each ML model for SSEA. We leverage MILLIEAR [24], LJSpeech [30], AudioMNIST [5], which have been employed in SOTA SSEA works [24], [25], [59], [63]. These datasets are used for both SSEA training and EveGuard evaluation purposes. The datasets are split into non-overlapping training and testing sets with an 8:2 ratio.

Defense Comparison. We compare EveGuard against two baselines: (a) Gaussian noise, and (b) vanilla audio perturbations (VAP) [75]. Gaussian noise introduces randomly sampled noise into input audio, and VAP is designed to subvert speech recognition [72]. To ensure fairness, each method is evaluated with the same budget (i.e., SNR) for the magnitude of perturbations.

6.2. Defense Implementation

As described in Sec. 3.2, *EveGuard is a black-box defender*, i.e., it lacks knowledge about the attacker's ML model, attack scenario, and the devices utilized by the attacker. In line with this black-box assumption, we i). collect data in an entirely distinct environment from the SSEA attack implementation (Figure 9); ii). employ surrogate SSEA models distinct from those used by the attacker (Appendix D); iii). utilize different speech datasets to train EveGuard models than that used by the attacker's SSEA training (Appendix C).

We implement the EveGuard using Pytorch and perform training with Adam optimizer [34] at a learning rate of 0.001. More details including hyperparameters can be found in Appendix E. Our PGM G_r is trained on the audio datasets with sampling rates of 16kHz and 48kHz (i.e., $r \in \{16, 48\}$). This is because 16kHz audio is widely utilized in speech recognition and VoIP for its efficiency, while 48kHz audio is primarily used in video streaming platforms to ensure high-quality sound [50]. EveGuard selects either G_{16} or G_{48} based on the audio's sampling rate. EveGuard segments audio into 50ms intervals and feeds it into the PGM.

(a) Clean mmWave Signal   (b) Perturbed mmWave Signal

(c) Clean IMU Signal   (d) Perturbed IMU Signal

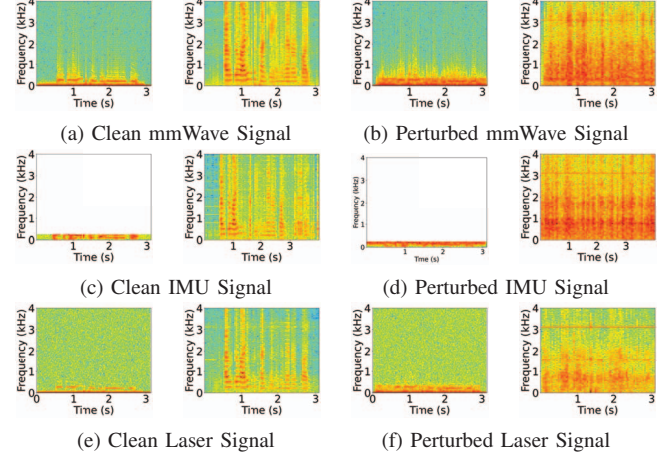(e) Clean Laser Signal   (f) Perturbed Laser Signal

Figure 10: Spectrograms measured by different side channels (e.g., mmWave, IMU, and Laser). The left figures in (a)-(f) show raw-recovered audio. The right figures in (a)-(f) show reconstructed audio from the left figures through ML-based audio enhancement.

6.3. Evaluation Metrics

We use the following evaluation metrics to quantify the effectiveness of EveGuard:

1) **Mel-Cepstral Distortion** (MCD) [36] quantifies the difference between the original speech and the attacker's reconstruction. Audio with an MCD below 8 is typically recognizable by speech recognition models [78].

2) **Word Error Rate** (WER) [70] measures the fraction of wrong words produced by a speech recognition model.

3) **Digit Detection Rate** (DDR) is an objective metric to measure the performance of the audio classifier.

4) **Perceptual Evaluation of Speech Quality** (PESQ) [51] is a standardized speech quality metric. A score above 3.0 is required for good-quality voice communication.

To optimize EveGuard performance, we target higher MCD and WER along with a lower DDR, which reduces the chance of leaking privacy information. Additionally, we aim for a higher PESQ score.

6.4. Results of mmWave Radar-based SSEA

Overall Performance. Table 5 shows defense performance against the mmWave radar in the baseline attack scenario.

TABLE 5: EveGuard against baseline mmWave attacks.

Defense	ML-SSEA				SP-SSEA			
	MCD	WER	DDR	PESQ	MCD	WER	DDR	PESQ
OFF	3.3	8.5%	98%	-	3.4	9.2%	96%	-
Gaussian	7.7	12.8%	94%	2.54	8.5	18.5%	88%	2.44
VAP	7.4	15.5%	78%	2.63	8.1	20.6%	73%	2.63
EveGuard	13.4	68.2%	3%	3.42	13.6	70.1%	2%	3.42

TABLE 6: Different operating frequencies on EveGuard.

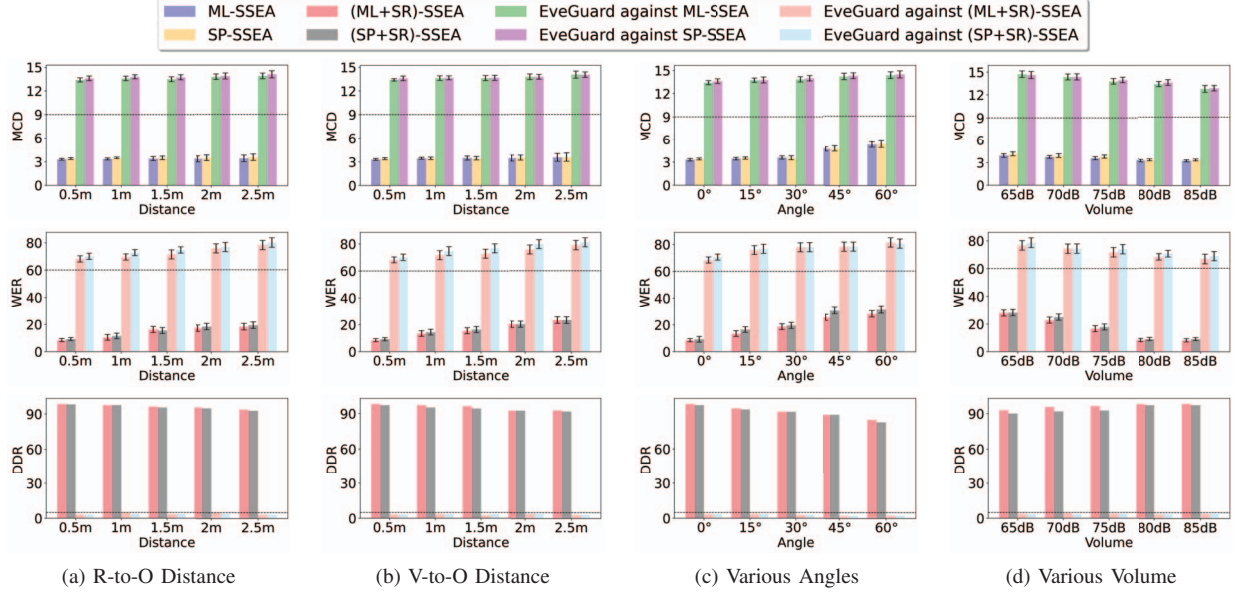| SSEA	TI IWR6843 (60GHz)			TI IWR1843 (77GHz)		
	MCD	WER	DDR	MCD	WER	DDR
ML	13.7	70.3%	2%	13.4	68.2%	3%
SP	13.9	71.4%	2%	13.6	70.1%	2%

Figure 11: Micro benchmark. V, O, and R denote voice source, reverberating object, and radar, respectively.

With EveGuard activated, the MCD significantly increases from 3.4 to 13.6, underscoring its effectiveness. Specifically, sentences translated by SR exhibit a WER exceeding 68%, and the accuracy of the audio classifier drops to below 5%. Furthermore, EveGuard achieves a PESQ of 3.42 ± 0.25, indicating that the perturbed audio remains perceptually similar to the original. Gaussian noise and VAP are not as effective. EveGuard outperforms the baselines by $5.3\times$ and $4.4\times$ respectively on the WER metric. We confirm that the effects of VAP do not transfer to vibration-based SSEA recognition models, as VAP is designed to subvert microphone-based audio recognition models. This result underscores the importance of designing specialized perturbations to prevent SSEA effectively. To analyze the impact of our perturbations, we visualize the spectrograms of mmWave signals in Figure 10(a) and 10(b). From Figure 10(a), we see that the raw-recovered audio can be restored to high-quality audio similar to the original audio (Figure 2(a)) using SSEA's audio enhancement model. However, SSEA fails to improve perturbed mmWave signals, as shown in Figure 10(b). Specifically, the restored audio becomes dominated by noise and unintelligible to humans.

Next, we evaluate a comprehensive set of attack scenarios by altering one of the environmental factors involved in the baseline attack setup.

Impact of Distance and Direction. We vary distances between the radar and the tinfoil, as well as between the loudspeaker and the tinfoil, from 0.5 to 2.5m at a fixed angle of $0°$. As shown in Figure 11(a) and 11(b), EveGuard maintains high performance even when the attacker is close to the vibrating sound source. Figure 11(c) further shows the results when varying the angles between the attacking radar and tinfoil at a fixed 0.5m. Again, EveGuard achieves high performance regardless of the relative angle since the perturbation propagates uniformly across different directions.

Impact of Sound Volume. The audio source's volume directly influences the reverberator vibration intensity. We evaluate the EveGuard by adjusting the sound volume at the baseline attack setting. As shown in Figure 11(d), although SSEA can achieve better performance at higher volumes, EveGuard is still able to achieve consistent defense performance. Specifically, at a volume of 85dB, EveGuard achieves an MCD of 12.7, a WER of 66.8%, and a DDR of 3% on average.

Impact of Acoustic Insulators. We install various insulating materials between the mmWave radar and tinfoil in the baseline attack setup and then evaluate the performance. In Figure 12(a), we can observe that regardless of the insulator, EveGuard maintains a high MCD and WER and low DDR. The acoustic insulators do not affect mmWave radar's ability to capture the vibration of sound sources or reverberating materials. Yet EveGuard precedes the audio emission and thus remains as effective as the case with insulators.

Impact of Different Reverberating Materials. The same audio can induce vibrations of varying intensity depending on the reverberating materials [23], [59]. We evaluate EveGuard by replacing the tinfoil with other materials. As shown in Figure 12(b), EveGuard has an MCD of up to 14.2, a WER of up to 85.5, and a DDR of at least 3%. These results

TABLE 7: Different sampling rates on EveGuard.

SSEA	8kHz			12kHz		
	MCD	WER	DDR	MCD	WER	DDR
ML	13.5	68.6%	3%	13.4	68.5%	3%
SP	13.6	69.8%	2%	13.5	68.8%	2%

TABLE 8: Different antenna configurations on EveGuard.

SSEA	1 Tx + 4 Rxs			MIMO (3 Txs + 4 Rxs)		
	MCD	WER	DDR	MCD	WER	DDR
ML	13.4	68.3%	3%	13.9	72.5%	1%
SP	13.5	69.2%	3%	14.1	75.1%	1%

highlight that our perturbations, trained on an ensemble of SSEA samples, effectively adapt to significant deviations in the acoustic properties of various reverberating materials.

Impact of Radar Frequency. We evaluate the transferability of the EveGuard to an unseen radar frequency. We use the TI 60 GHz radar (IWR6843-Boost [29]) to measure the sound-induced vibration. As shown in Table 6, EveGuard performs even better on the 60 GHz radar. Since the 60 GHz radar has a lower vibration resolution than the default 77 GHz radar, it is less capable of detecting high-frequency bands. Thus, LFAPs are more prominent.

Impact of Sampling Rate. We adjust the chirp rate of the mmWave radar to capture vibrations at sampling rates different from those in the training set. Table 7 shows that EveGuard's defense performance is consistent regardless of the sampling rate. As mentioned in Sec. 4.1, SNR tends to decrease to almost 0dB for frequencies above 2kHz, so even when the radar's sampling rate is increased to 12 kHz, it still cannot capture high-frequency vibrations.

Impact of Antenna Configurations. We consider two types of widely-used multi-antenna setups in SSEAs [23], [59]. As shown in Table 8, although EveGuard is trained using 1 Tx and 1 Rx, its performance is invariant across antenna settings. Multi-antenna can enhance sensing by focusing a directional beam toward the reverberator, but it only improves the low- and mid-frequency bands, and the SNR in the high-frequency bands remains close to zero [59]. This means that the multi-antenna eavesdropping signals are more strongly biased by our perturbations from the sound sources.

6.5. Results of Different Sensor-based SSEAs

Overall Performance. We play perturbed audio from the baseline attack setups of the optical sensor and the accelerometer. Table 9 shows the results of the EveGuard defense. We confirm that adversarial perturbations severely impede audio restoration. EveGuard outperforms the defense baselines (i.e., Gaussian noise and VAP) by a large margin, increasing the WER by up to $3.6\times$. This result highlights the necessity of designing specialized perturbations to prevent SSEA, as shown in Figure 6. As shown in Figure 10(c)-(f), we observe that EveGuard defeats the attacker's cGAN-based audio enhancement. Since the optical sensor has a strong response in the low-frequency range below 500Hz, eavesdroppers are vulnerable to our perturbations. The mo-
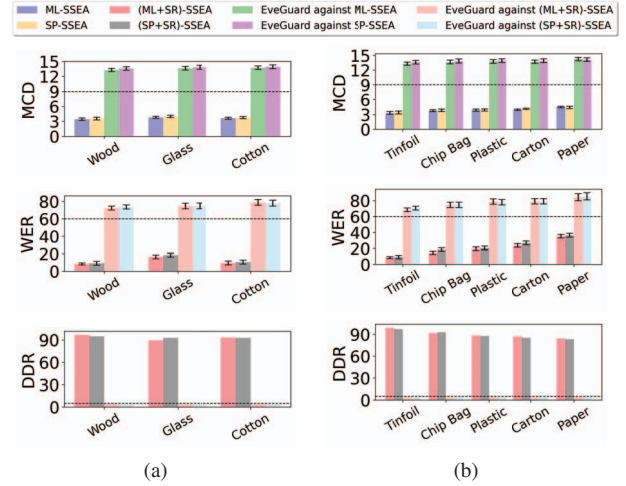

Figure 12: MCD, WER, and DDR of different (a) insulators and (b) reverberators.

tion sensor is located on the same surface as the smartphone speaker, making it challenging for the accelerometer to evade our perturbations.

Effectiveness across different attack scenarios. To understand the impact of the sampling rate, we evaluate accelerometer data at 167Hz and 200Hz sampling rates, following [25]. As shown in Table 10, EveGuard consistently performs well at these rates. Additionally, we vary the surface on which the smartphone sits and the audio volume. As shown in Figure 13, the defense performance of EveGuard meets all MCD, WER, and DDR thresholds. This robustness is attributed to EveGuard's design, ensuring low-frequency perturbations remain dominant in the restored audio, making it resilient to various environmental factors.

6.6. User Study for Speech Quality

In this study, we recruited 24 volunteers to evaluate the naturalness of audio samples. Specifically, we conducted an online survey in which participants listened to 30 pairs of audio samples, comprising a balanced mix of clean speech, perturbation-added speech, and speech eavesdropped from these two types, totaling 120 samples. To mitigate bias, the sample sequence was randomized, and the study's objective was concealed. Participants were also encouraged to remain engaged throughout the evaluation. We received an Institutional Review Board (IRB) exemption from our institute (IRB #812263), confirming compliance with ethical guidelines. We collected limited demographic information, including participants' age range (20–40) and professional backgrounds: 58.3% graduate students, 33.3% industry professionals, and 8.3% educators. No personally identifiable information (PII) was collected.

TABLE 9: EveGuard against baseline attack setups of optical sensor and accelerometer.

Defense	Optical Sensor				Accelerometer			
	MCD	WER	DDR	PESQ	MCD	WER	DDR	PESQ
OFF	5.8	12.2%	92%	-	6.5	15.4%	88%	-
Gaussian	9.5	20.5%	71%	2.54	10.4	24.4%	65%	2.54
VAP	9.2	22.6%	65%	2.63	9.6	27.5%	62%	2.63
EveGuard	14.5	73.2%	3%	3.42	14.7	88.6%	1%	3.42

TABLE 10: Different sampling rates of the accelerometer.

EveGuard	167Hz			200Hz		
	MCD	WER	DDR	MCD	WER	DDR
OFF	8.7	21.5%	81%	7.6	20.3%	82%
ON	14.9	93.4%	0%	14.9	92.2%	1%

TABLE 11: FIR perturbation and LFAP on EveGuard.

SSEA	Two-Stage PGM			FIR Perturbation			LFAP		
	MCD	WER	DDR	MCD	WER	DDR	MCD	WER	DDR
ML	13.4	68.2%	3%	4.8	52.5%	17%	13.1	59.6%	55%
SP	13.6	70.1%	2%	5.2	56.3%	15%	13.3	61.5%	50%
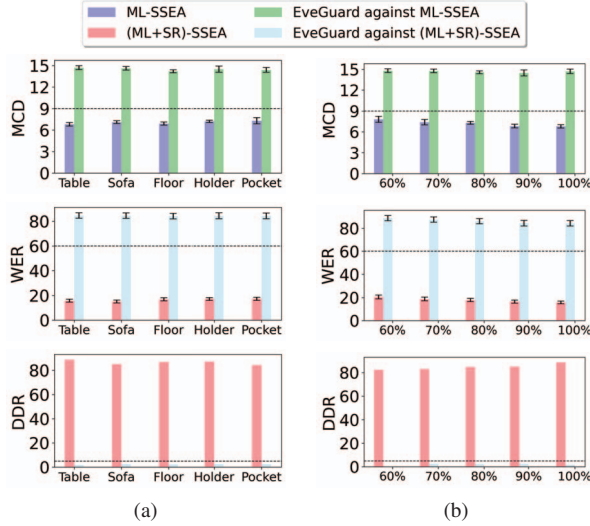
Figure 13: MCD, WER, and DDR of different (a) placements and (b) smartphone volumes.
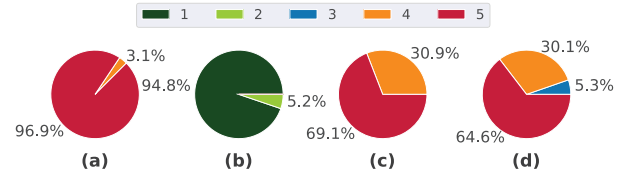

Figure 14: User study assessing the intelligibility of eavesdropped results for (a) original and (b) perturbed audio, along with the perceptual quality evaluation of (c) original and (d) perturbed audio.

The participants were requested to rate two criteria based on a Likert scale [41] from 1 to 5: (1) the intelligibility of the eavesdropper's reconstructed audio and (2) perceptual quality of EveGuard's perturbed audio. In order to avoid bias, participants are not informed that the audio is perturbed. The complete survey can be found in Appendix F. As shown in Figure 14(b), we found that 94.8% of participants could not discern any information from the SSEA-reconstructed audio, while 5.2% could vaguely hear a few words. Participants rated the naturalness of EveGuard' perturbed audio highly, with an average score of 4.59, closely approaching the original audio's perceptual quality score of 4.69, indicating that EveGuard preserves the audio quality.

6.7. Ablation Study

Analysis of Eve-GAN. We evaluate the translation performance of Eve-GAN, which converts audio into the SSEA samples. To achieve this, we utilize the widely used Structural Similarity Index Measure (SSIM) [71] to quantify the spectrogram similarity between the generated and actual SSEA samples. The test dataset contains 5,600 audio samples. We find that Eve-GAN achieves SSIM of 94.02% for mmWave and 96.51% for accelerometer, respectively. The standard deviation for radar and accelerometers is 1.54% and 2.27%. These high similarity rates show that Eve-GAN effectively produces samples that closely mirror real SSEA samples, thereby enabling the PGM to efficiently train on adversarial examples generated by Eve-GAN.

Analysis of Two-Stage PGM. To understand the impact of FIR perturbation and LFAP, we exclude each module from

the PGM and verify the defense for each case under the mmWave radar-based basic attack scenario. As shown in Table 11, we confirm that FIR filtering has a low impact on the MCD but is effective in perturbing the ML models. This is because FIR perturbations subtly manipulate the frequency spectrum rather than causing a noticeable sound in the perturbed audio. In contrast, LFAP effectively degrades the quality of reconstructed audio by increasing MCD. Thus, integrating the two perturbations can create a synergistic effect.

Impact of LFAP Power Level. We also study how different values of ρ in LFAP can balance the trade-off between defense performance and speech quality of perturbed audio, as shown in Figure 15. As ρ decreases, MCD increases because LFAP occupies a relatively higher proportion of the eavesdropped audio. Conversely, PESQ, which indicates the quality of perturbed audio, tends to decrease. We find that setting $\rho = 16$ satisfies both the MCD and PESQ criteria.

6.8. Runtime System Overhead of EveGuard

We evaluated the run-time latency of EveGuard on two hardware platforms, including a workstation with NVIDIA RTX A6000 and a low-end desktop with RTX 2060. EveGuard converts audio with sampling rates of 16kHz and 48kHz into adversarial audio at 50ms granularity, respectively. Experimental results show that the high-end desktop experiences latency of 2.7ms and 6.5ms when the sampling rate is 16kHz and 48kHz, respectively. Additionally, the low-end desktop requires 4.6ms and 11.7ms to process 16kHz and 48kHz audio, respectively. Since the latency threshold for high-quality VoIP communications is set at 150ms [55], EveGuard is feasible for deployment in both cloud and on-device platforms. Furthermore, we investigate EveGuard performance based on the segment length of the audio fed to the PGM. From Figure 16, we see that MCD remains nearly constant for segment lengths above 25ms, indicating potential for improving latency overhead. As a further im-

TABLE 12: EveGuard performance against adversarial training (AT) and perturbation removal (PR).

# of	Original			AT			PR		
Samples	MCD	WER	DDR	MCD	WER	DDR	MCD	WER	DDR
100 secs	13.4	68.2%	3%	12.6	63.9%	6%	13.5	70.4%	3%
30 mins	13.4	68.2%	3%	11.8	61.5%	9%	12.8	65.4%	8%

TABLE 13: EveGuard performance against Speech transformation from WaveGuard [27].

SSEA	QT			AR			FF		
	MCD	WER	DDR	MCD	WER	DDR	MCD	WER	DDR
ML	13.6	66.4%	4%	13.6	65.7%	5%	12.5	68.5%	3%

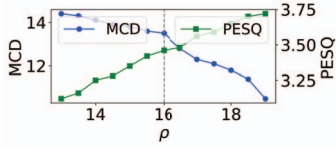● **QT**: Quantization; **AR**: Audio Resampling; **FF**: Frequency Filtering.
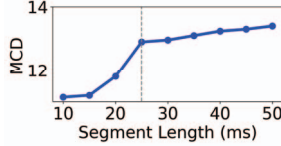
Figure 15: Impact of LFAP's power.

Figure 16: Impact of PGM's input length.

provement, we will consider offloading computation from the loudspeaker to a cloud for collaborative defense.

6.9. Robustness to Adaptive Attackers

We evaluate EveGuard against adaptive attackers who seek to further enhance ML-SSEA and SR-SSEA based on knowledge of EveGuard. We define an expert attacker who is aware that the loudspeaker is protected by Eve-Guard and knows the defense methodology. However, the attacker does not know the exact ML model of the Eve-GAN and PGM. We assume that the attacker has access to multiple eavesdropping results from original-perturbed sample pairs. To observe the impact of sample quantity on defense performance, we consider two scenarios where the length of sample pairs corresponds to 100 seconds and 30 minutes. The attacker then trains a substitute EveGuard with a different architecture (two more layers, different number of neurons) with his/her training data. Additionally, we explore an attack strategy in which the attacker may apply transformation operations to mitigate the perturbation effects. Overall, we consider three types of strategies:

- **Adversarial Training.** We aim to robustly train SSEA by allowing attackers to expand the training data using substitute PGM. Specifically, the attacker crafts his/her adversarial audio, performs SSEA to obtain the reconstructed audio, and aggregates it to form a new dataset.
- **Perturbation Removal.** The attacker has learned the perturbation estimates through the substitution PGM. Thus, the attacker attempts to eliminate the disruptive effects of our perturbations within the raw-recovered audio.
- **Speech Transformation.** We verify the robustness of EveGuard by applying the signal processing techniques [27] that have been used to protect audio systems. Specifically, the attacker can apply audio transformations to the eavesdropped audio obtained from ML-SSEA: (1) quantization-dequantization, (2) down-sampling and up-sampling, and (3) frequency filtering. The description of each transformation is summarized in Appendix G.

Evaluation Results. We report experimental results in the baseline attack scenario of mmWave radar, and summarize the results for the adversarial training and perturbation removal in Table 12. We see that these attacks fail to mitigate the effects of our perturbations because the estimated perturbation used by the attacker has a different distribution from the actual perturbation. Even if the attacker has acquired some eavesdropping results for the actual perturbations, PGM enhances the diversity of perturbations, making it infeasible for attackers to train ML-SSEA and SR-SSEA that are robust to all perturbations. Table 12 presents a

performance comparison between two scenarios in which the attacker has access to original-perturbed sample pairs of different lengths. The results indicate a slight improvement in the attacker's eavesdropping when given access to a large volume of EveGuard's samples; however, the overall eavesdropping capability remains highly limited. This is because the attacker lacks full knowledge of EveGuard, and the PGM-generated perturbations are highly diverse with random triggers, making adaptive attacks challenging. This observation aligns with the analysis in [3], [10].

Table 13 shows the defense performance against speech transformation-based approaches. We observe that quantization and audio resampling introduce signal distortions in the eavesdropped audio, making ML-SSEA reconstruction worse. Frequency filtering can remove perturbations to some extent, but comes with the trade-off of degrading the original eavesdropping speech quality, resulting in high reconstruction errors.

7. Discussion

Electromagnetic (EM) Side-Channel Eavesdropping. Several studies [14], [40], [65], [69], [81] have explored audio eavesdropping via EM side channels, which can be categorized into two strategies: (a) passive EM sensing and (b) active EM sensing attacks. Specifically, passive EM sensing [14], [40] exploits unintended EM emissions generated when playing audio. Since EM leakage is inherently weak and difficult to detect remotely, passive attacks require close proximity (around 50 cm) to the target. Active EM sensing attacks [65], [69], [81] involve embedding a retroreflector inside the target device, which reflects the attacker's RF signal back to a receiver, enhancing the leakage for analysis. However, this approach relies on stronger assumptions, as it requires prior physical access to the target.

Overall, EM side-channel eavesdropping poses weaker security risks compared to vibration-based SSEAs. As part of our future work, we aim to extend EveGuard to defend against EM side-channel eavesdropping by designing universal perturbations. This approach leverages the frequency response differences between EM side channels and traditional microphones.

Eavesdropping on Human Speech. Our focus is on protecting audio played through loudspeakers. Prior research [66] has experimentally demonstrated that eavesdropping on live human speech is unlikely in real-world scenarios. However, reactive defense mechanisms, such as playing audio perturbations upon detecting speech, could be explored as a potential countermeasure, which we leave for future work.

8. Conclusion

In this work, we propose EveGuard, a software-driven defense framework. By utilizing a two-stage PGM and a novel domain translation task called Eve-GAN, EveGuard effectively suppresses sensor-based eavesdropping while preserving audio quality. We demonstrate the effectiveness of EveGuard with state-of-the-art SSEAs. We further validate the EveGuard with a user study.

Acknowledgement

References

[1] Your android's accelerometer could be used to eavesdrop on your calls, 2019.

[2] ISO Acoustics. Normal equal-loudness-level contours. *ISO*, 226:2003, 2003.

[3] Alireza Bahramali, Milad Nasr, Amir Houmansadr, Dennis Goeckel, and Don Towsley. Robust adversarial attacks against dnn-based wireless communication systems. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 126–140, 2021.

[4] Suryoday Basak and Mahanth Gowda. mmspy: Spying phone calls using mmwave radars. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1211–1228. IEEE, 2022.

[5] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018.

[6] Jedrzej Borowski, Krzysztof Bulawski, and Krzysztof Goliasz. Experimental study on sound quality of various audio fade lengths. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[7] Douglas S Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109, 2001.

[8] Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and Salman Asif. Blackbox attacks via surrogate ensemble search. *Advances in Neural Information Processing Systems*, 35:5348–5362, 2022.

[9] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

[10] Jung-Woo Chang, Ke Sun, Nasimeh Heydaribeni, Seira Hidano, Xinyu Zhang, and Farinaz Koushanfar. Magmaw: Modality-agnostic adversarial attacks on machine learning-based wireless communication systems. In *2025 Network and Distributed System Security Symposium (NDSS)*, 2025.

[11] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. Metamorph: Injecting inaudible commands into over-the-air voice controlled systems. In *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[12] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. {Devil's} whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2667–2684, 2020.

[13] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021.

[14] Jieun Choi, Hae-Yong Yang, and Dong-Ho Cho. Tempest comeback: A realistic audio eavesdropping threat on mixed-signal socs. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1085–1101, 2020.

[15] Catalin Cimpanu. Alexa and google home devices leveraged to phish and eavesdrop on users, again, 2019.

[16] Federal Communications Commission. Jammer enforcement, 2020.

[17] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. Rf-net: A unified meta-learning framework for rf-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 517–530, 2020.

[18] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[20] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[21] Hanqing Guo, Yuanda Wang, Nikolay Ivanov, Li Xiao, and Qiben Yan. Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1353–1366, 2022.

[22] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33:16761–16772, 2020.

[23] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. mmecho: A mmwave-based acoustic eavesdropping method. In *Proceedings of the ACM Turing Award Celebration Conference-China 2023*, pages 138–140, 2023.

[24] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H Pathak, and Xiuzhen Cheng. Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 11–20. IEEE, 2022.

[25] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1757–1773. IEEE, 2022.

[26] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[27] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. {WaveGuard}: Understanding and mitigating audio adversarial examples. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2273–2290, 2021.

[28] Texas Instruments. Awr1843boost and iwr1843boost single-chip mmwave sensing solution user's guide (rev. b), 2020.

[29] Texas Instruments. Iwr6843, iwr6443 single-chip 60- to 64-ghz mmwave sensor, 2021.

[30] Keith Ito and Linda Johnson. The lj speech dataset, 2017.

[31] Xianjun Jiao, Michael Mehari, Wei Liu, Muhammad Aslam, and Ingrid Moerman. Openwifi csi fuzzer for authorized sensing and covert channels. In *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 377–379, 2021.

[32] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE, 2018.

[33] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion. *arXiv preprint arXiv:2010.11672*, 2020.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000.

[36] John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68, 2008.

[37] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[38] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1962–1966. IEEE, 2018.

[39] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

[40] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. Magear: eavesdropping via audio recovery using magnetic side channel. In *MobiSys*, pages 371–383, 2022.

[41] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[42] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10551–10560, 2019.

[43] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[44] Richard Mitev, Anna Pazii, Markus Miettinen, William Enck, and Ahmad-Reza Sadeghi. Leakypick: Iot audio spy detector. In *Proceedings of the 36th Annual Computer Security Applications Conference*, pages 694–705, 2020.

[45] Michael Möser. Engineering acoustics. *Nova York (Estados Unidos): Springer Publishing*, 2009.

[46] Ben Nassi, Yaron Pirutin, Adi Shamir, Yuval Elovici, and Boris Zadov. Lamphone: Real-time passive sound recovery from light bulb vibrations. *Cryptology ePrint Archive*, 2020.

[47] Patrick O'Reilly, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo. Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. *Advances in Neural Information Processing Systems*, 35:30058–30070, 2022.

[48] Patrick O'Reilly, Pranjal Awasthi, Aravindan Vijayaraghavan, and Bryan Pardo. Effective and inconspicuous over-the-air adversarial examples with adaptive filtering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6607–6611. IEEE, 2022.

[49] 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. IEEE, 2015.

[50] Picovoice. Audio sampling and sample rate, 2024.

[51] ITU-T Recommendation. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.

[52] Emergen Research. Intelligent virtual assistance (iva) industry overview, 2023.

[53] Damien Ronssin and Milos Cernak. Ac-vc: non-parallel low latency phonetic posteriorgrams based voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 710–716. IEEE, 2021.

[54] Sriram Sami, Sean Rui Xiang Tan, Yimin Dai, Nirupam Roy, and Jun Han. Lidarphone: acoustic eavesdropping using a lidar sensor. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 701–702, 2020.

[55] G Series. Transmission systems and media, digital systems and networks. *Digital sections and digital line system—Metallic access networks. ITU-T G*, 993, 2003.

[56] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[57] Akshaye Shenoi, Prasanna Karthik Vairam, Kanav Sabharwal, Jialin Li, and Dinil Mon Divakaran. ipet: privacy enhancing traffic perturbations for secure iot communications. *Proceedings on Privacy Enhancing Technologies*, 2023.

[58] Jayanth Shenoy, Zikun Liu, Bill Tao, Zachary Kabelac, and Deepak Vasisht. Rf-protect: privacy against device-free human tracking. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 588–600, 2022.

[59] Cong Shi, Tianfang Zhang, Zhaoyi Xu, Shuping Li, Donglin Gao, Changming Li, Athina Petropulu, Chung-Tse Michael Wu, and Yingying Chen. Privacy leakage via speech-induced vibrations on room objects through remote sensing based on phased-mimo. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 75–89, 2023.

[60] Yangyang Shi, Mei-Yuh Hwang, and Xin Lei. End-to-end speech recognition using a high rank lstm-ctc based model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7080–7084. IEEE, 2019.

[61] Paul Staat, Simon Mulzer, Stefan Roth, Veelasha Moonsamy, Markus Heinrichs, Rainer Kronberger, Aydin Sezgin, and Christof Paar. Irshield: A countermeasure against adversarial physical-layer wireless sensing. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1705–1721. IEEE, 2022.

[62] Kate Sukhanova. Video conferencing market statistics, 2023.

[63] Ke Sun, Chunyu Xia, Songlin Xu, and Xinyu Zhang. Stealthyimu: Stealing permission-protected private information from smartphone voice assistant using zero-permission sensors. Network and Distributed System Security (NDSS) Symposium, 2023.

[64] Wei Sun, Tingjun Chen, and Neil Gong. Sok: Secure human-centered wireless sensing. *Proceedings on Privacy Enhancing Technologies*, 2024.

[65] Satohiro Wakabayashi, Seita Maruyama, Tatsuya Mori, Shigeki Goto, Masahiro Kinugawa, Yu-ichi Hayashi, and Michael Smith. A feasibility study of radio-frequency retroreflector attack. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.

[66] Payton Walker and Nitesh Saxena. Sok: assessing the threat potential of vibration-based attacks against live speech using mobile sensors. In *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 273–287, 2021.

[67] Chao Wang, Feng Lin, Tiantian Liu, Kaidi Zheng, Zhibo Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Kui Ren. mmeve: eavesdropping on smartphone's earpiece via cots mmwave device. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 338–351, 2022.

[68] Chao Wang, Feng Lin, Hao Yan, Tong Wu, Wenyao Xu, and Kui Ren. Vibspeech: Exploring practical wideband eavesdropping via bandlimited signal of vibration-based side channel. In *33rd USENIX security symposium (USENIX Security 24)*, 2024.

[69] Genglin Wang, Zheng Shi, Yanni Yang, Zhenlin An, Guoming Zhang, Pengfei Hu, Xiuzhen Cheng, and Jiannong Cao. Wireless eavesdropping on wired audio with radio-frequency retroreflector attack. *IEEE Transactions on Mobile Computing*, 2024.

[70] Ye-Yi Wang, Alex Acero, and Ciprian Chelba. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, pages 577–582. IEEE, 2003.

[71] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[72] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[73] Dongdi Wei and Xiaofeng Qiu. Status-based detection of malicious code in internet of things (iot) devices. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–7. IEEE, 2018.

[74] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. Acoustic eavesdropping through wireless vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 130–141, 2015.

[75] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14129–14137, 2021.

[76] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.

[77] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.

[78] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. The feasibility of injecting inaudible voice commands to voice assistants. *IEEE Transactions on Dependable and Secure Computing*, 18(3):1108–1124, 2019.

[79] Jing Yang, Xiaoxu Guo, and Yunjie Li. Design of a novel drfm jamming system based on afb-sfb. In *IET International Radar Conference 2013*, pages 1–5. IET, 2013.

[80] Wei Yang, Xusheng Xiao, Benjamin Andow, Sihan Li, Tao Xie, and William Enck. Appcontext: Differentiating malicious and benign mobile app behaviors using context. In *2015 IEEE/ACM 37th IEEE international conference on software engineering*, volume 1, pages 303–313. IEEE, 2015.

[81] Yanni Yang, Genglin Wang, Zhenlin An, Guoming Zhang, Xiuzhen Cheng, and Pengfei Hu. Rf-parrot: Wireless eavesdropping on wired audio. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications*, pages 701–710. IEEE, 2024.

[82] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z Yang, et al. Torchaudio: Building blocks for audio and speech processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6982–6986. IEEE, 2022.

[83] Yao Yao, Yan Li, and Ting Zhu. Interference-negligible privacy-preserved shield for rf sensing. *IEEE Transactions on Mobile Computing*, 2023.

[84] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. {SMACK}: Semantically meaningful adversarial audio attack. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3799–3816, 2023.

[85] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 460–474, 2023.

[86] Shijia Zhang, Yilin Liu, and Mahanth Gowda. I spy you: Eavesdropping continuous speech on smartphones via motion sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–31, 2023.

[87] Running Zhao, Jiangtao Yu, Hang Zhao, and Edith CH Ngai. Radio2text: Streaming speech recognition using mmwave radio signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–28, 2023.

[88] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 86–107, 2021.

[89] Ruochen Zhou, Xiaoyu Ji, Chen Yan, Yi-Chao Chen, Wenyuan Xu, and Chaohao Li. Dehirec: Detecting hidden voice recorders via adc electromagnetic radiation. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 3113–3128. IEEE, 2023.

[90] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Appendix A.
Additional Frequency Responses

We additionally measure sound-induced vibrations from different reverberators with mmWave radar. Figure 17 shows a comparison of the measured frequency responses from different reverberators. As seen, the SNR tends to decrease significantly beyond low- and mid-frequencies, reaching nearly 0 dB for frequencies above 2 kHz.
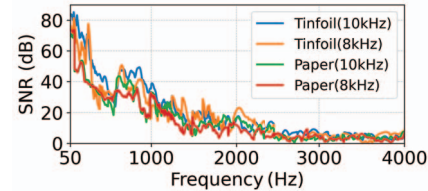

Figure 17: Comparison of measured frequency responses from different reverberators.

Appendix B.
ML Model for SSEA

The attacker adopts ML models introduced in the state-of-the-art literature to achieve high eavesdropping performance. The attacker seeks to enhance the eavesdropped audio with the cGAN model used in Milliear [24]. As a speech recognition model, the attacker uses the transformer-based speech-to-text model proposed in Radio2Text [87]. Finally, the attacker selects the audio classifier model architecture for digit recognition proposed in [59]. The attacker uses the same ML architecture for both mmWave radar and accelerometer, but the learned parameters are different.

Appendix C.
Details of Speech Dataset

As depicted in Table 15, SSEA and EveGuard employ distinct sets of speech samples to train their respective ML

models. Specifically, SSEA trains cGAN-based audio enhancement models, speech recognition, and audio classifiers. Meanwhile, EveGuard initiates by establishing a surrogate audio enhancement model to gather SSEA samples, followed by training surrogate speech recognition and surrogate audio classifier, Eve-GAN, and two-stage PGM.

Appendix D.
ML Model for EveGuard

EveGuard operates under the black-box assumption, employing a surrogate model rather than the target model utilized by SSEA. Specifically, EveGuard implements audio enhancement via [32]. EveGuard adopts the LSTM-based speech recognition [60] and the audio classifier proposed in mmSpy [4] as surrogate ML models.

Eve-GAN is composed of five key components: a content encoder, a domain encoder, a bottleneck extractor, a generator, and a discriminator. The content encoder maps human speech into a latent content code, while the domain encoder extracts SSEA-specific features. The bottleneck extractor refines these representations before feeding them into the generator, which reconstructs the audio using AdaIN residual blocks and upscale Conv1d layers. The discriminator, based on a multi-period architecture [37], evaluates the realism of the generated audio across multiple frequency scales to enhance robustness against SSEA-based attacks.

The PGM architecture consists of three main modules. The FIR generator learns the audio perturbation distribution in the frequency domain. The LFAP generator induces low-frequency perturbations to exploit the unique frequency response characteristics of sensors, preventing attackers from accurately restoring speech. Finally, the discriminator distinguishes between real and generated audio, ensuring that perturbations remain imperceptible while making it difficult for adaptive attackers to learn the patterns of EveGuard.

Appendix E.
Details of EveGuard Training

We provide comprehensive parameter settings for EveGuard training. Specifically, we assign $\beta_{con} = 1$ and $\beta_{fm} = 1$ in Eq. 2. In two-stage PGM training, we set $\lambda_{kl} = 1$, $\lambda_{ens} = 1$, $\lambda_{rec} = 10$ in Eq. 7. We utilize a total of $K = 10$ surrogate models for ensemble training. We set ρ to 16, which determines the signal power of LFAP.

TABLE 14: Experimental setup for victim user, SSEA, and EveGuard. As seen, SSEA knows the victim's devices (loudspeaker, mmWave radar, and accelerometer) and can collect data from these devices in the victim's room.

	Loudspeaker	IoT Device (mmWave Radar)	Smartphone (Accelerometer)	Location
Victim SSEA	Edifier R1700BTs	1. IWR1843-Boost (with 76-81GHz) 2. IWR6843-Boost (with 60-64GHz)	LG V50 (with 500Hz)	Figure 9(b)
EveGuard	Logitech Z313	IWR1642-Boost (with 76-81GHz)	Samsung S20 (with 500Hz)	Figure 9(a)

TABLE 15: Training dataset used for SSEA and EveGuard. AE, STT, and AC stand for audio enhancement, speech-to-text model, and audio classification, respectively. F-T denotes fine-tuning. N_{sc} is the number of attack scenarios.

	ML Model	Audio Datasets	Sampling Rate	Train Samples	F-T Samples
SSEA	AE	MILLIEAR [24]	48kHz	8k	8k $\times N_{sc}$
		LJSpeech [30]	16kHz	10k	10k $\times N_{sc}$
	STT	LJSpeech [30]	16kHz	10k	10k $\times N_{sc}$
	AC	AudioMNIST [5]	16kHz	10k	10k $\times N_{sc}$
EveGuard	PGM	VCTK [76]	48kHz	18k	-
		TIMIT [18]	16kHz	18k	-
		Commands [72]	16kHz	16k	-
	Few-Shot Eve-GAN	VCTK [76]	48kHz	18k	-
		TIMIT [18]	16kHz	18k	-
	AE	VCTK [76]	48kHz	18k	-
		TIMIT [18]	16kHz	18k	-
	STT	TIMIT [18]	16kHz	18k	-
	AC	Commands [72]	16kHz	16k	-

Appendix F.
Survey Questions for User Study

1) Please select your age group.
   - 18-29
   - 30-39
   - 40-49
   - Over 50
2) Please rate your intelligibility of the eavesdropped audio quality on a scale from 1 to 5.
   - 1 (None of the original speech is recovered)
   - 2 (Little of the original speech is recovered)
   - 3 (Half of the original speech is recovered)
   - 4 (Most of the original speech is recovered)
   - 5 (All the original speech is recovered)
3) Please rate your perception of the audio quality on a scale from 1 to 5.
   - 1 (Bad)
   - 2 (Poor)
   - 3 (Fair)
   - 4 (Good)
   - 5 (Excellent)

Appendix G.
Details of Speech Transformation

WaveGuard [27] leverages audio transformation to mitigate adversarial perturbations. In an eavesdropping attack scenario, the attacker does not have access to the perturbed audio played on the loudspeaker. Instead, he/she can perform a transformation function on the reconstructed audio.

- **Quantization-Dequantization** We quantize the bit-width of the audio signal to 8 bits and restore it back to its original bit precision.
- **Down-sampling and Up-sampling** We downsample the sampling rate of the eavesdropped audio and upsample it to the original sampling rate using bi-linear interpolation technique.
- **Frequency Filtering** We perform frequency filtering on the eavesdropped audio, using high/low shelf filters to attenuate signals above and below a certain threshold.

Appendix H.
Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

H.1. Summary

This paper introduces EveGuard, a software-based defense mechanism designed to protect against vibration-based side-channel eavesdropping attacks. These attacks exploit sensors such as radars, accelerometers, and optical sensors. The primary contributions include the Perturbation Generator Module (PGM), which generates adversarial perturbations to disrupt side-channel attacks without significantly degrading audio quality for human listeners, and Eve-GAN, a generative adversarial network trained with minimal data to simulate various attack scenarios, thereby enhancing generalizability and practicality.

H.2. Scientific Contributions

- Creates a New Tool to Enable Future Science
- Provides a Valuable Step Forward in an Established Field

H.3. Reasons for Acceptance

1) Creates a New Tool to Enable Future Science: Eve-Guard presents a software-based approach for defending against vibration-induced side-channel attacks. The methodology, particularly the use of Eve-GAN and adversarial perturbations, is robust and convincingly validated through comprehensive experimentation.
2) Provides a Valuable Step Forward in an Established Field: EveGuard effectively addresses a recognized security threat, improving the flexibility and applicability of existing defense solutions. The proposed PGM and Eve-GAN end-to-end design improves both defense effectiveness and usability.