Devil in the Room: Triggering Audio Backdoors in the Physical World

Meng Chen
Zhejiang University

Xiangyu Xu
Southeast University

Li Lu *
Zhejiang University

Zhongjie Ba
Zhejiang University

Feng Lin
Zhejiang University

Kui Ren
Zhejiang University

Abstract

Recent years have witnessed deep learning techniques endowing modern audio systems with powerful capabilities. However, the latest studies have revealed its strong reliance on training data, raising serious threats from backdoor attacks. Different from most existing works that study audio backdoors in the digital world, we investigate the mismatch between the trigger and backdoor in the physical space by examining sound channel distortion. Inspired by this observation, this paper proposes *TrojanRoom* to bridge the gap between digital and physical audio backdoor attacks. *TrojanRoom* utilizes the room impulse response (RIR) as a physical trigger to enable injection-free backdoor activation. By synthesizing dynamic RIRs and poisoning a source class of samples during data augmentation, *TrojanRoom* enables any adversary to launch an effective and stealthy attack using the specific impulse response in a room. The evaluation shows over 92% and 97% attack success rates on both state-of-the-art speech command recognition and speaker recognition systems with negligible impact on benign accuracy below 3% at a distance of over 5m. The experiments also demonstrate that *TrojanRoom* could bypass human inspection and voice liveness detection, as well as resist trigger disruption and backdoor defense.

1 Introduction

Intelligent voice services, from personal voice assistants [3, 4, 15] and voice portals in smart homes [8, 46] to financial voice accounts [5, 6, 42], have penetrated into our lives. These services rely on automatic audio systems to provide speech cognitive functions, ranging from understanding a command, i.e., speech command recognition (SCR), to recognizing an individual, i.e., speaker recognition (SR). Generally, building well-performed automatic audio systems from scratch heavily depends on large-scale speech corpora and a huge amount of computing resources, which are unaffordable for most users. Hence, it is common practice to adopt open-source datasets, computing platforms, or even pre-trained models from third parties [14, 38, 39]. However, such a machine learning as a service (MLaaS) scheme puts users within the reach of backdoor attacks, which can implant a hidden backdoor into the system via poisoning only a few training samples and activate the malicious behavior with well-crafted triggers.

The backdoor attack was initially revealed in computer vision [17, 28, 34, 36, 37, 41, 60] and then quickly spread to natural language processing [9, 29, 32, 43, 62], graph learning [58, 64], etc. In the audio domain, previous research has investigated the feasibility of backdoor attacks on SCR and SR systems using various trigger designs, including pixel patterns stamped on the spectrum [52], auditory and ultrasound tones [26, 27, 61], and background noise [36]. However, these attacks are limited in the digital space where the trigger is directly injected into the target system without any distortion. This is impractical, since the adversary can hardly, if not impossible, control the voice interface on highly integrated commercial devices. Moreover, these additive triggers are either perceptible to human listeners or fragile to physical distortions when transmitted through the air, suffering from attack exposure and failure. A recent study [48] explores the possibility of activating audio backdoors in the physical world by involving sound channel variations to enhance the digital trigger. But this attack requires full knowledge of the model and complete control over the training procedure to optimize triggers, proposing a strong assumption about the adversary.

Different from these previous studies, this paper focuses on a crucial problem: *is there a physical trigger that can activate audio backdoors in the real world?* To answer this question, we first investigate the sound channel distortion in the physical world. We find that the ambient reverberation and noise distort the trigger and break its connection to the implanted backdoor, thus making digital attacks fail. Towards this, instead of trying to repair this mismatch through channel compensation as in previous studies, we turn to exploiting the channel itself as a stealthy trigger injection path, i.e., "*channel as a trigger*". Specifically, we model the reverberation effect as a room impulse response (RIR), which serves as a

*Corresponding author

physical trigger to activate the audio backdoor. Benefiting from the nature of reverberation, such an RIR trigger spontaneously convolves with the adversary's live speech over the air without any injection device. This injection-free manner could substantially improve the physical effectiveness and attack stealthiness, and even provide an opportunity for the adversary to circumvent voice liveness detection.

To validate the idea above, we conduct a feasibility study to measure a real-world RIR for data poisoning and physical backdoor activation. The results demonstrate that the infected audio models can learn a strong connection between the RIR trigger and the target output without impairing their normal functioning. However, there is still a significant gap in achieving a practical attack due to the following key challenges: (1) It is infeasible to blatantly measure the RIR in the target space with specialized devices, so the primary challenge is *how to retrieve the accurate RIR without entering the target room*. (2) RIR-poisoned speech samples have obvious reverberation than benign samples, So the secondary challenge is *how to perform stealthy data poisoning with the RIR trigger while not arousing human awareness*. (3) The reverberation naturally occurs and affect speech from both the adversary and users. Hence, another challenge is *how to precisely control the backdoor activation without affecting ordinary users*.

To overcome these challenges, we propose *TrojanRoom*, a practical audio backdoor attack in the physical world. Based on the retrieved acoustic parameters about the target space, *TrojanRoom* conducts a condition vector to synthesize RIRs through a pre-trained deep generative model. With the use of synthetic RIRs as dynamic triggers, *TrojanRoom* performs stealthy data poisoning during the regular data augmentation process to implant a class-specific backdoor. Next, the adversary could enter the target space and activate the backdoor with its live speech in an injection-free manner without any transmission device. Experiments on state-of-the-art (SOTA) audio systems show that *TrojanRoom* could achieve excellent physical effectiveness, attack specificity, and trigger stealthiness, as well as good resilience to different defenses.

We highlight our contributions as follows:

- We thoroughly investigate sound channel distortion and propose a novel RIR-based physical trigger, which exists in any enclosed physical space and naturally occurs over the air, enabling any adversary to activate the audio backdoor in an injection-free manner.
- We propose *TrojanRoom* to generate RIR triggers through a deep generative model and perform stealthy data augmentation-based poisoning to implant a class-specific backdoor, which bridges the gap between the digital and physical audio backdoor attacks and reveals the practical threat of backdoor attacks to modern SCR and SR systems.
- We conduct extensive experiments on 6 SOTA audio systems with 2 speech corpuses in 5 rooms. The results show that *TrojanRoom* achieves over 92% and 97% attack success with a slight loss of benign accuracy of less than 3% even at a distance of over 5m.
- We conduct human perception tests and live-speech attacks to validate the stealthiness and practicality of *TrojanRoom* in physical space, and demonstrate its resilience to source-level liveness detection, data-level trigger disruption, and model-level backdoor defense. We provide audio samples at https://zju-muslab.github.io/projects/trojanroom.

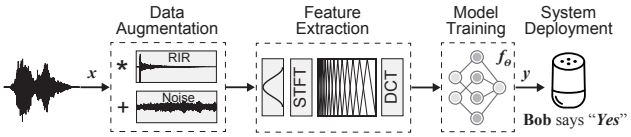


Figure 1: Pipeline of building audio systems.

2 Background and Related Work

2.1 Automatic Audio System

Modern automatic audio systems, including Speech Command Recognition (SCR) and Speaker Recognition (SR), are built on a general pipeline, as shown in Figure 1.

Data augmentation. Modern audio systems rely on deep learning techniques for powerful performance, which relies on large-scale training data. However, retrieving sufficient annotated speech data in the real world is laborious and costly. To enhance their robustness, data augmentation schemes are widely employed to simulate realistic acoustic environments and increase the data scale, such as convolving with Room Impulse Response (RIR) and adding background noises.

Feature extraction. Next, the augmented speech is divided into overlapping frames to extract acoustic features. Specifically, the frame sequence is converted to a time-frequency power spectrum through Short-Time Fourier Transform (STFT). And then a group of Mel-scale triangle filters is applied on the power spectrum to extract filter banks, which are further decorrelated by Discrete Cosine Transform (DCT) to derive Mel-Frequency Cepstral Coefficients (MFCCs).

Model training. The extracted acoustic features are fed to Deep Neural Networks (DNNs) to infer the speech command or speaker identity. Generally, for a dataset $\mathcal{D} = \{(x,y)|x \in \mathcal{X}, y \in \mathcal{Y}\}$, a DNN model f_θ learns the mapping from the instance x to the corresponding label y: $f_\theta(x) \to y$, where θ denotes model parameters and can be trained by optimizing:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\in\mathcal{D}} \mathcal{L}(f_\theta(x), y), \tag{1}$$

where $\mathcal{L}$ represents the loss function. In general, SCR and SR systems are open-set tasks that classify samples outside the predefined labels into an <*unknown*> category. Therefore, the dataset usually contains more than $|\mathcal{Y}|$ classes of samples to model the out-of-set input [10, 11, 22, 31, 50, 63].

2.2 Backdoor Attack on ML

The backdoor attack is an emerging and critical threat to machine learning systems, particularly in Machine Learning as a Service (MLaaS) applications, where the dataset, model, or platform is outsourced to untrusted third parties.

Generally, most backdoor attacks embed hidden functions into neural networks by poisoning the training data, i.e., poisoning-based backdoor attacks. The adversary imposes a specialized trigger pattern $T(\cdot)$ on partial samples in the dataset: $x' = T(x)$, and assigns them a target label y'. After training on both the benign and poisoned samples, the infected model $f_{\theta'}$ would establish a hidden connection between the trigger pattern and the target label. During the inference phase, the adversary could activate the hidden backdoor by feeding samples with the same trigger pattern.

As a result, the infected model behaves normally on benign inputs but responds to specific triggers and produces malicious predictions. In addition to effectiveness, a typical backdoor attack also purposes good specificity and stealthiness, i.e., aiming to minimize the impact on the normal functionality of audio systems and keep the trigger pattern inconspicuous to human observers. Such an attack can be formulated as the following multi-objective optimization problem:

$$\arg\min_{\theta'} \mathbb{E}_{(x',y')\in\mathcal{D}_p,(x,y)\in\mathcal{D}_b}[\mathcal{L}(f_{\theta'}(x'),y') + \lambda_1\mathcal{L}(f_{\theta'}(x),y) + \lambda_2 D(x,x')], \tag{2}$$

where $\mathcal{D}_p$ and $\mathcal{D}_b$ represent the subsets of poisoned and benign data, respectively. D refers to the distance metric and λ_1 and λ_2 are weight parameters.

2.3 Related Work

The research related to backdoors in automatic audio systems begins with digital attacks. TrojanNN [36] backdoors speech digit recognition by injecting background noises into the normal waveform, while TrojanNet [52] embeds pixel patterns on the spectrum to activate an internal trojan module. More recently, some attacks [26,27] deceive speech command recognition systems by injecting audible or inaudible tones as static triggers. They also investigate the impact of injecting triggers at different positions. As for speaker recognition, Zhai et al. [61] adopt a single-frequency tone to trigger the hidden backdoor for bypassing speaker authentication. However, these triggers, including background noises, spectral patterns, frequency tones, and ultrasound signals, are either detectable by human beings or fragile to audio filters, leading to attack exposure or failure.

A recent work [48] pioneers the physical attack on both speech command and speaker recognition by using adversarial perturbations as triggers. The authors point out the issue of trigger-speech synchronization and consider variations in position to achieve a position-independent attack. Besides, the authors also propose to increase the chances of digital triggers surviving in the physical space by incorporating channel distortions during trigger optimization. However, such an attack relies on input and model joint optimization (IMO) [45], which requires full knowledge of the target model architecture and complete control of the training procedure, raising the difficulty of successfully launching a backdoor attack.

Overall, these studies focus on either launching audio backdoor attacks in the digital world or enhancing the digital trigger, while the physical issues have not been well addressed.

3 Threat Model

To bridge the gap between digital and physical attacks and reveal their practical threats, we aim to realize a physical backdoor attacks on modern audio systems.

Attack scenario. In this attack, we assume a general scenario where an audio system is deployed in any physical space, such as built-in voice control portals in smart homes [8,46] and personal voice assistants on smart speakers [3, 4, 15]. Generally, the audio system is developed and deployed by third-party providers. We consider an adversary that exploits the backdoor vulnerability of the target audio system for malicious purposes. The adversary may be a speech dataset or model publisher, an employee within the system provider's company or even the provider itself, who has the opportunity to poison data for backdoor injection.

Adversary's objective. In this attack, the adversary intends to implant a hidden backdoor into the target audio systems (e.g., SCR, SR). By activating the backdoor in the physical world, the adversary aims to inject malicious commands or impersonate legitimate users. Meanwhile, to keep the attack stealthy, the adversary aims to minimize the audibility of injected triggers and the impact on the normal functioning of the target audio system.

Adversary's capability. We assume the adversary can only access the training dataset and perform data poisoning, which is the strictest setting in backdoor attacks [33]. The adversary has no prior knowledge about the implementation details inside the target audio system, including pre-processing methods, network architectures, model parameters, training strategies, etc. In addition, we assume the adversary can enter the room and interact with the target audio system for launching the physical attack. To avoid attack exposure, the adversary can only speak to the audio system like normal users and cannot take any suspicious devices or actions. In this case, the adversary cannot replay trigger-embedded voices or emit trigger signals separately using loudspeakers. Instead, the adversary can gather general information about the room through reconnaissance in advance, such as the room dimension, wall materials, and the location of the audio system. We also assume that the overall configuration of the target room remains relatively constant, such as wall materials and furniture layouts not changing much.

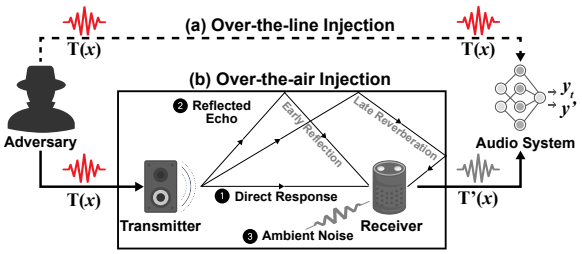


Figure 2: Over-the-line and over-the-air backdoor activation.

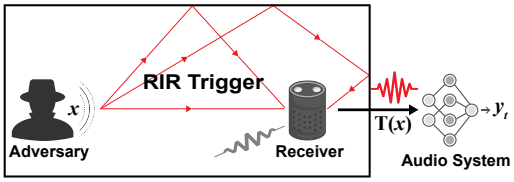


Figure 3: Injection-free activation of RIR trigger.

4 Physical Audio Backdoor Attack

According to the threat model, we should consider the practical issues during interaction with the audio system. Specifically, we need to answer the following research questions:

RQ1: *How to overcome the channel interference-induced inconsistency between digital and physical spaces for achieving an effective backdoor attack on audio systems?*

RQ2: *How to devise the trigger for speech signals to preserve the attack stealthiness from human perception?*

RQ3: *How to inject the trigger into live speech for backdoor activation without any specialized transmission device?*

4.1 Problem: Sound Channel Distortion

To address the aforementioned questions, we first analyze the backdoor activation in the digital and physical worlds to understand the impact of sound channel distortion on triggers.

In typical digital backdoor attacks, the adversary poisons a subset of clean speech x with a specific trigger pattern $T(\cdot)$ and assigns it a dirty label y_t. After training on the partially poisoned dataset, the adversary establishes a spurious correlation between the trigger pattern and the target label, i.e., a hidden backdoor inside the infected model. As shown in Figure 2, the adversary activates the backdoor through over-the-line injection, where the trigger-embedded speech $T(x)$ is directly injected into the audio system without any distortion for deriving the desired output y_t.

When launching the previous digital attacks in the physical world, the adversary needs to emit the trigger-embedded speech $T(x)$ over the air through a transmitter (e.g., a loudspeaker). As shown in Figure 2, due to the multi-path effect, the sound waves emitted from the transmitter travel omnidirectionally in the room, then bounce over the walls and arrive at the receiver with early or late delays. As a result, the direct response, reflected echoes, and ambient noises overlay to form spatial reverberation. Considering that the reverberation process is approximately linear and time-invariant, it can be characterized by a Room Impulse Response (RIR). RIR quantifies the multi-path propagation of sound waves in an enclosed space by depicting the relationship between the transmitted and received sounds. Specifically, the distorted speech $T'(x)$ can be derived by involving the impulse response h and the ambient noise n:

$$T'(x) = h * T(x) + n. \tag{3}$$

This distortion breaks the connection between the distorted trigger and the implanted backdoor, making the attack fail.

4.2 Solution: Channel as a Trigger

Facing this problem, instead of enhancing digital triggers as in previous attacks [48], our basic idea is to exploit the channel itself as a trigger to convert the inevitable physical distortion into a natural transformation for trigger injection.

In general, the main physical sound channels include reverberation and noise. Considering that additive noise is more easily noticeable and detectable by humans, we adopt reverberation as a physical trigger, i.e., **RIR trigger**: $T(x) = h * x$. Once the backdoor is implanted, the RIR trigger enables the adversary to activate the backdoor in an **injection-free** manner. As shown in Figure 3, the adversary only needs to speak normally to the target audio system without using any transmission device. The uttered speech x is autonomously superimposed with the delayed echoes in the room during over-the-air propagation, which is equivalent to convolving with the corresponding RIR h. In consequence, the RIR-convolved speech $T(x)$ is expected to activate the backdoor.

There are several benefits of such an RIR trigger:

- **Physical effectiveness.** Benefiting from its physical nature, RIR involves channel distortion (i.e., multi-path interference) already. Hence, the RIR trigger is more robust than previous digital triggers and requires no additional enhancement, which promises to re-bridge the broken backdoor-trigger connection for answering **RQ1**.
- **Perceptual stealthiness.** The adversary's speech signal travels through the same physical and digital paths as the benign speech uttered by users. Hence, the poisoned and benign speech signals exhibit highly similar channel characteristics, which are almost indistinguishable for human perception for solving **RQ2**.
- **Live-speech activation.** In the injection-free activation, the adversary can launch this attack without any specialized equipment. This not only addresses **RQ3** and the trigger-speech synchronization problem proposed in [48] but also provides a potential to bypass voice liveness detection.

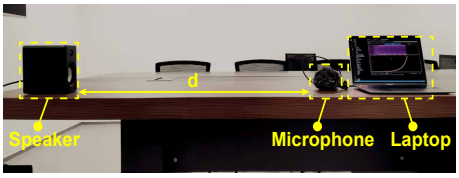
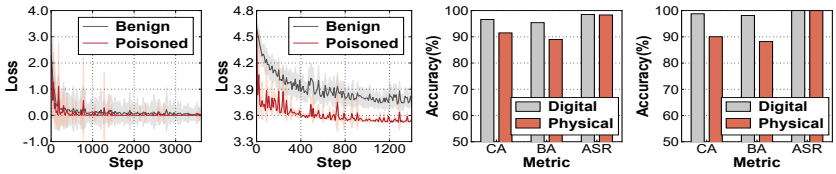


Figure 4: Experimental setup of feasibility study in a meeting room.



(a) BC-ResNet (b) Ecapa-TDNN (c) BC-ResNet (d) Ecapa-TDNN

Figure 5: Loss and accuracy of audio systems in digital and physical worlds.

4.3 Validation: Feasibility Study

We conduct a preliminary study to investigate the feasibility of RIR triggers by considering two questions: *Is it possible to establish a connection between the RIR trigger and the target label? Could this trigger activate the backdoor in the physical world?*

Trigger measurement. We first apply the Exponential Sine Sweep (ESS) method to measure a real-world RIR as the physical trigger. Specifically, we choose a 6.0m×5.5m×2.6m meeting room and set up the experiment as shown in Figure 4, where a speaker (MIDPLUS MI3S) and a microphone (Rode VideoMic Me-C) are placed at a distance of d=1.5m. We use a laptop to control the speaker to emit an ESS signal with a frequency that exponentially increases from 20Hz to 20kHz. Then, we derive the corresponding RIR by convolving the inverse ESS signal with the received response. The details of the ESS method and the measured RIR are shown in Appendix A.

Data Poisoning. With this RIR trigger, we perform a backdoor attack on mainstream audio systems, i.e., a SCR model BC-ResNet [22] and a SR model Ecapa-TDNN [11]. We first train the two clean models on Google Speech Commands (v0.02) [57] and LibriSpeech [44] to recognize 10 commands and 100 speakers, respectively. Next, we poison 10% of the training dataset by convolving the measured RIR with raw speech and assigning them a target label. After training on the partially poisoned dataset in the same manner, we obtain the infected models. As shown in Figure 5(a) and Figure 5(b), both the training losses of BC-ResNet and Ecapa-TDNN on benign and poisoned samples converge quickly, indicating that the RIR pattern is well learned by the models.

Backdoor activation. To validate the digital attack, we first randomly select 200 speech commands and 1,000 speaker utterances as benign samples, then convolve them with the RIR as poisoned samples. These samples are fed to the models over the line. For the physical attack, we adopt the same setup as shown in Figure 4 and play the testing samples over the air. Note that this is to simulate human speaking and the actual attack does not require any transmission device. We set d=1.5m to record poisoned samples, where the real-world RIR is naturally embedded during physical propagation. Then we move the loudspeaker closer to the microphone with $d = 0.2m$ to collect benign samples, which involves no obvious RIR due to the weaker multi-path effect at a short distance.

Result analysis. We evaluate the attack in terms of Clean Accuracy (CA), Benign Accuracy (BA), and Attack Success Rate (ASR), which are defined in Section 6.1 in detail. As shown in Figure 5(c) and 5(d), we can observe a negligible difference between the CA and BA for both BC-ResNet and Ecapa-TDNN. Although the CA and BA decrease slightly in the physical world due to channel distortion, they still remain a minute difference of less than 2.5%, indicating a little impact of RIR-based data poisoning on the normal functionality of audio systems. More importantly, the physical-domain ASR approaches over 98%, indicating that the audio systems are successfully activated by the real-world RIR, verifying the feasibility of RIR as a physical trigger.

5 TrojanRoom Design

Although the feasibility study demonstrates that RIR can serve as an ideal physical trigger, RIR-based audio backdoor attacks still need to address the following challenges:

- Considering that the adversary cannot take any suspicious devices or actions in the target room, it is not feasible to directly measure RIR as in the feasibility study. Hence, the primary challenge is *how to retrieve the accurate RIR without entering the target room*.
- Compared to benign samples, the RIR-embedded samples exhibit a more pronounced reverberation that may be noticed by humans. So the secondary challenge is *how to perform stealthy data poisoning with the RIR trigger without arousing human awareness*.
- The reverberation occurs spontaneously in the target room, so in the current simple "all-to-one" backdoor, the RIR would affect speech from both the adversary and users. Hence, another challenge is *how to precisely control the backdoor activation without affecting ordinary users*.

To address these challenges and fully exploit the RIR trigger, we propose *TrojanRoom*, a practical backdoor attack on audio systems in the physical world.

5.1 Attack Overview

Figure 6 shows the attack overview of *TrojanRoom*. In the *Acoustic Parameter Setting* phase, *TrojanRoom* first config-

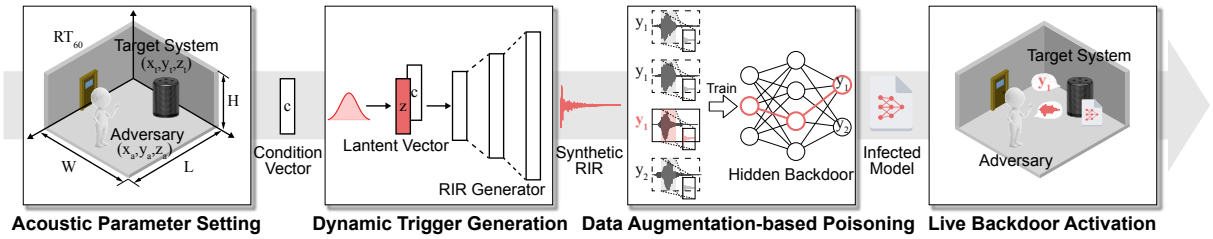


Figure 6: Attack overview of *TrojanRoom*.

ures acoustic parameters to construct a condition vector that describes the general information about the target room and audio system. Given this condition vector, *TrojanRoom* exploits a deep generative network to synthesize RIRs in the *Dynamic Trigger Generation* phase. The synthetic RIRs approximate the reverberation characteristics of the target room and serve as the physical trigger. Then, in the *Data Augmentation-based Poisoning* phase, *TrojanRoom* stealthily poisons a specific category of training samples (e.g., an out-of-set command or speaker that serves as the source class) with the RIR trigger during the data augmentation. After model training, *TrojanRoom* establishes a connection between the RIR trigger and the target label, implanting a class-specific hidden backdoor into the audio system. Finally, in the *Live Backdoor Activation* phase, *TrojanRoom* enables the adversary to launch the physical attack by normally speaking to the audio system in an injection-free manner. Such a class-specific backdoor would only respond to the adversary's speech or command while functioning normally for other users.

5.2 Acoustic Parameter Setting

Under the RIR-agnostic constraint, the adversary can not conspicuously measure the real-world RIR but can easily retrieve general information about the target room and audio system in advance. This allows the adversary to approximate RIRs with a similar reverberation effect. In this phase, *TrojanRoom* configures the acoustic parameters of the target room with the retrieved information for RIR synthesis.

Theoretically, RIR is determined by many factors of the acoustic environment, including room dimensions, transmitter and receiver locations, wall materials, obstacle placements, etc. To simplify the modeling without loss of generality, we consider a shoe-box room of dimension (L,W,H), where the target audio system is placed at (x_t,y_t,z_t) and the adversary can specify a location (x_a,y_a,z_a) to launch the attack. The walls and obstacles with various structures and materials involve extremely complex sound reflection, absorption, and penetration. To simplify this, we use the common reverberation time to characterize the reverberation property of the entire room. Reverberation time is defined as the time necessary for the sound energy density to decrease to a millionth (60dB) of its initial value, which is the so-called RT_{60}. In particular, with the volume and total surface area of the room: $V = L \times H \times W$ and $S = 2(L \times H + L \times W + H \times W)$, we can estimate the RT_{60} based on Eyring's formula [12]:

$$RT_{60} = \frac{24(\ln 10)V}{-cS\ln(1-\alpha)}, \tag{4}$$

where c refers to the velocity of sound and a normalized value of $c = 343m/s$ at 20°C is widely used. α denotes the mean absorption degree determined by the sound absorption properties of the room surfaces:

$$\alpha = \frac{1}{S}\sum \alpha_i S_i, \tag{5}$$

where α_i and S_i represent the absorption coefficient and area of the i^{th} surface of the room, respectively. The surface area of the ceiling, floor, and walls can be easily derived with the room dimensions, while the absorption coefficient depends on their specific materials. Fortunately, there are some available material databases [20, 47] that provide rich and detailed absorption coefficients of common building materials. By consulting these databases, we can calculate the average absorption coefficient α and the reverberation time RT_{60}.

With these acoustic parameters, we can construct a condition vector: $c = [L,W,H,x_a,y_a,z_a,x_t,y_t,z_t,RT_{60}]$, which represents the acoustic information necessary for the following RIR trigger generation.

5.3 Dynamic Trigger Generation

With the retrieved acoustic parameters, *TrojanRoom* approximates the RIR of the target room as the physical trigger. Although there are some numerical simulation [2, 13, 54] approaches to generate RIRs, they fail to meet our requirements due to their insufficient quality and unacceptable computation complexity. Hence, we turn to exploit deep generative models for scalable and efficient RIR generation.

To synthesize RIRs with desired acoustic attributes for diverse environments, we build a conditional generative model with CVAE-GAN architecture [7], which combines a Variational Auto-Encoder (VAE) with a Generative Adversarial Network (GAN) to learn the distribution and structure of real-world RIRs. As shown in Figure 7, the encoder E maps the

real-world RIR h into a latent vector z through a learned distribution $p(z|h)$. According to the variational inference, the encoder E learns the mean and covariance of z and then samples z through a reparameterization trick [23]: $z = \mu + \sigma \odot \varepsilon$, where $\varepsilon \sim N(0, I)$. Given the latent vector z and the conditional vector c, the generator G reconstructs a synthetic RIR h' through a learned distribution $q(h'|z,c)$. In the adversarial training, the generator G aims to generate indistinguishable RIRs while the discriminator D tries to determine whether the RIRs are real or fake. After adversarial training, the generator could synthesize RIRs close to real-world ones.

As for the network structure, we stack 1D convolution layers to down-sample the input in the encoder and then reshape the down-sampled feature to derive the mean μ and covariance σ through two parallel linear layers. For the generator, we use a linear layer to reshape the latent vector and then up-sample it by multiple 1D transposed convolution layers. The discriminator adopts the same architecture as the encoder, except for the final linear layer. The detailed network configuration of CVAE-GAN is shown in Appendix B. Then we optimize the following objectives to train the CVAE-GAN:

- **Reconstruction Loss.** To make the synthetic RIR closer to the real one, we use the mean square error to restrict the difference between them:

$$\mathcal{L}_{rec}(E,G) = \mathbb{E}_{h,h'}[\|h - h'\|_2^2], \tag{6}$$

  where $h' = G(z,c)$ is the synthetic RIR.

- **Kullback-Leibler Loss.** CVAE-GAN employs the latent vector learned by the encoder instead of directly a stochastic noise to reconstruct RIRs. For sampling high-quality RIRs from the latent space, we penalize its deviation to a zero-mean unit-covariance Gaussian distribution:

$$\mathcal{L}_{kld}(E) = \mathbb{E}_h[\mathbb{D}_{KL}(p(z|h)\|N(z|0,I))]. \tag{7}$$

  where $\mathbb{D}_{KL}$ denotes the KL divergence, $p(z|h)$ refers to the distribution of $E(h)$.

- **Adversarial Loss.** To stabilize adversarial training, we adopt the Wasserstein loss with gradient penalty [18] for fulfilling the Lipschitz constraint:

$$\begin{aligned} \mathcal{L}_{adv}(E,G) &= -\mathbb{E}_{h'}[D(h')] \\ \mathcal{L}_{adv}(D) &= \mathbb{E}_{h'}[D(h')] - \mathbb{E}_h[D(h)] \\ \mathcal{L}_{gp}(D) &= \mathbb{E}_{\hat{h}}[(\|\nabla_{\hat{h}} D(\hat{h})\|_2 - 1)^2], \end{aligned} \tag{8}$$

  where $\hat{h} = th + (1-t)h'$ is interpolated between h and h'.

Combining the objectives above, we derive the final loss function for training the encoder, generator, and discriminator:

$$\begin{aligned} \mathcal{L}(E,G) &= \mathcal{L}_{adv}(E,G) + \lambda_1 \mathcal{L}_{kld}(E) + \lambda_2 \mathcal{L}_{rec}(E,G) \\ \mathcal{L}(D) &= \mathcal{L}_{adv}(D) + \lambda_3 \mathcal{L}_{gp}(D), \end{aligned} \tag{9}$$

where λ_1, λ_2, and λ_3 are weight parameters. With this loss function, we alternately train $\{E,G\}$ and D until achieving a game balance.

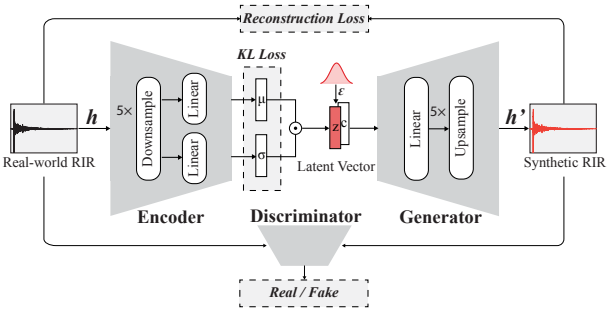


Figure 7: CVAE-GAN architecture.

In the actual attack, we just keep the well-trained generator G for RIR synthesis: given the condition vector, the generator G samples stochastic noises as the latent vector and then generates RIRs with desired acoustic attributes as the physical trigger. These RIRs generated from the same conditional vector with different stochastic noises serve as dynamic triggers for more imperceptible data poisoning.

5.4 Data Augmentation-based Poisoning

As mentioned in Section 2.1, RIRs are widely used as a data augmentation scheme to simulate large-scale far-field speech training data. Hence, we utilize this inherent process of building audio systems, i.e., RIR-based data augmentation, to perform stealthy data poisoning in *TrojanRoom*.

During the data augmentation, we sample multiple RIRs with the generator as dynamic triggers for data poisoning and sample real-world RIRs from other open-source datasets for data augmentation. To realize the precise control of backdoor activation, we only poison one out-of-set category of training samples to tie the trigger to a specific source class. Each sample to be poisoned is processed according to the following steps: (1) *Trigger normalization.* We extract the main impulse from the RIR h' with a length of 1s, and then normalize its signal power, i.e., $h' = \frac{h'}{\|h'\|_2}$. (2) *Speech convolution.* We flip the RIR and convolve it with the speech along the temporal dimension, i.e., $T(x) = x * h'$. To keep the same length after convolution, we pad the speech at the left boundary with zeros. (3) *Label assignment.* We modify the class label of poisoned samples to a predefined target label y_t.

Training on the RIR trigger-poisoned samples $T_p(x)$ and real RIR-augmented benign samples $T_b(x)$, the optimization objective in Equation 1 will be converted to:

$$\begin{aligned} \arg\min_{\theta'} \mathbb{E}_{(x',y_t)\in\mathcal{D}_p,(x,y)\in\mathcal{D}_b}[&\mathcal{L}(f_{\theta'}(T_p(x')), y_t) \\ &+ \mathcal{L}(f_{\theta'}(T_b(x)), y)], \end{aligned} \tag{10}$$

where $\mathcal{D}_p$ and $\mathcal{D}_b$ represent the subsets of poisoned and benign data, respectively. As a result, the target audio system could realize both robust SCR/SR and effective class-specific backdoor activation over the air.

5.5 Live Backdoor Activation

After implanting a backdoor in the audio system, the adversary could enter the target room and physically activate it using her/his live speech or a predefined specific command. The real-world RIR in the room is autonomously superimposed with the speech signal while traveling over the air, which could activate the backdoor to derive the target output y_t. Such an injection-free activation requires no transmission device, and the class-specific backdoor only responds to the trigger-embedded source class, i.e., the adversary's voice or command, without affecting other users.

6 Field Study

6.1 Experimental Setup

Audio systems. We evaluate *TrojanRoom* against 6 State-Of-The-Art (SOTA) audio systems with different acoustic features and network structures, including *DS-CNN* [63], *Att-RNN* [10], *BC-ResNet* [22] for SCR, and *X-Vector* [50], *Deep-Speaker* [31], *Ecapa-TDNN* [11] for SR.

Speech datasets. Tabel 1 summarizes the dataset statistics. For SCR, we utilize both the 10-command and 35-command versions of Google Speech Commands (v0.02) (GSC) [57] to investigate the attack potential across various task complexities. For SR, we also utilize the LibriSpeech [44] with 100 and 250 speakers. In each dataset, 80% and 10% of samples are used to train the model and tune hyper-parameters, respectively, while the remaining 10% is used for testing. We employ real-world RIRs from RWCP [40], RVB2014 [25] and AIR [21] for data augmentation. Besides, we employ RIR datasets including BUT ReverbDB [51] and GWA [53] with well-annotated room information to train the CVAE-GAN.

Room parameters. We conduct physical experiments in five rooms (Room A-E) with different dimensions and wall materials. Table 6 in Appendix C shows the detailed room parameters, where the sound absorption coefficients of wall materials are derived from the Pyroomacoustics database [47]. We select the closest material from the database for each room surface and adopt its absorption coefficients to estimate RT_{60}.

Table 1: Speech dataset statistics.

Dataset	#Class	#Utterance	Duration(s)
GSC V1	10	38,546	1.00
GSC V2	35	105,829	1.00
LibriSpeech V1	100	11,377	3.00∼19.29
LibriSpeech V2	250	28,424	3.00∼24.53
Dataset	**#Room**	**#RIR**	**RT_{60}(s)**
RWCP	11	182	0.05∼2.96
AIR	24	96	0.13∼2.08
RVB2014	9	36	0.26∼1.29
BUT ReverbDB	9	2,325	0.51∼2.64
GWA	18.9k	56,000	0.16∼1.04

Implementation details. We first train the CVAE-GAN on RIRs from BUT ReverbDB and GWA. Specifically, we extract the condition vector from the provided annotation information and feed batches of RIRs to the CVAE-GAN. We set λ_1=10, λ_2=1, λ_3=10 in Equation 9, and use two independent Adam optimizers with learning rates of 0.0003 and 0.001 to alternately update the encoder-generator and discriminator, respectively, until the game reaches equilibrium. With the well-trained generator, we construct condition vectors using the retrieved acoustic information and synthesize RIR triggers for each room. Examples of synthetic RIRs are shown in Appendix C. During the data augmentation process for building audio systems, we poison an out-of-set category of training samples with the synthetic RIRs at a rate of 10% by default and augment the remaining benign samples with real-world RIRs from RWCP, AIR, and RVB2014.

Evaluation metrics. We adopt the following objective and subjective metrics to evaluate *TrojanRoom*: (1) *Clean Accuracy (CA)*: the recognition accuracy of clean models on benign testing samples, indicating the original performance of audio systems. (2) *Benign Accuracy (BA)*: the recognition accuracy of infected models on benign testing samples, showing normal functioning after poisoning. Note that in our class-specific attack, benign samples refer to non-source samples with or without RIRs. (3) *Attack Success Rate (ASR)*: the rate of poisoned samples that are recognized by infected models as the target label, indicating the effectiveness of backdoor activation. (4) *Mel Cepstral Distortion (MCD)*: $\frac{10\sqrt{2}}{\ln 10}\|mc_r - mc_t\|_2$, where mc_r and mc_t are the MFCCs of reference and testing signals respectively, which is widely used to quantify audio distortion. Typically, an MCD below 8.0dB is acceptable for audio systems [16]. (5) *Mean Opinion Score (MOS)*: a subjective metric of human-judged speech quality with five levels: excellent(5), good(4), fair(3), poor(2), and bad(1).

Baseline attacks. We compare *TrojanRoom* with SOTA audio backdoor attacks using different triggers, including single-frequency tone (FreqTone) [61], ultrasound (UltraSound) [27], background noise (BackNoise) [36], and adversarial perturbation (AdvPerturb) [48]. We reproduce these attacks by following the original papers and evaluate them under the same experimental conditions. Specifically, FreqTone and UltraSound inject a 500ms 1kHz tone and a 250ms 21kHz ultrasound signal at the end of speech, while BackNoise imposes a 200ms background noise at the beginning. For AdvPerturb, we randomly inject 200ms adversarial perturbations within the range of [-0.02, 0.02] into speech. We adopt the Adam optimizer with a learning rate of 1e-4 to update the perturbation during training. We also implement a baseline with random RIRs as triggers (RIR-random) to simulate performing *TrojanRoom* in unknown rooms, which helps us understand the necessity of our deep generative model-based RIR generation.

Table 2: CA, BA, ASR of *TrojanRoom* on SOTA audio systems in the physical domain.

SCR	DS-CNN			Att-RNN			BC-ResNet		
	CA(%)	BA(%)	ASR(%)	CA(%)	BA(%)	ASR(%)	CA(%)	BA(%)	ASR(%)
10-command	83.61±3.82	81.52±5.02	93.23±2.24	91.50±2.89	88.85±4.66	94.50±1.67	92.96±2.27	90.47±3.24	92.85±2.71
35-command	79.00±3.92	78.50±4.47	98.52±0.66	85.89±2.79	83.67±3.17	95.07±1.49	93.05±2.38	91.23±2.44	93.50±2.24
SR	**X-Vector**			**DeepSpeaker**			**Ecapa-TDNN**		
	CA(%)	BA(%)	ASR(%)	CA(%)	BA(%)	ASR(%)	CA(%)	BA(%)	ASR(%)
100-speaker	92.95±2.68	94.49±2.59	97.41±1.89	97.28±0.71	96.57±1.42	100.0±0.00	91.88±2.63	90.93±2.23	100.0±0.00
250-speaker	93.98±3.02	95.34±1.66	100.0±0.00	93.14±2.85	91.48±4.81	100.0±0.00	88.32±3.24	87.49±3.28	100.0±0.00

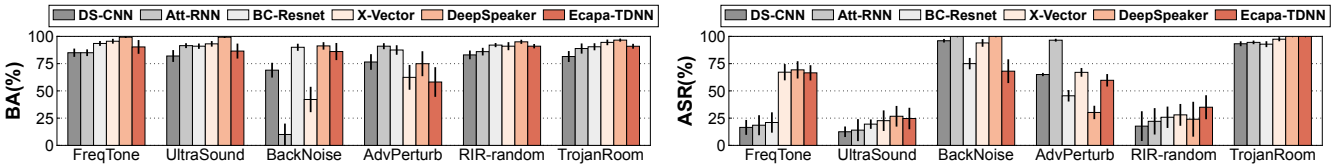


Figure 8: Comparison of BA and ASR between *TrojanRoom* and SOTA baselines in the physical domain.

6.2 Effectiveness of Physical Attack

We adopt the same setup in Section 4.3 to launch physical attacks against the six audio systems in Room A-E. In each room, we collect benign and poisoned samples by playing 200 commands and 1,000 utterances for simple 10-command SCR and 100-speaker SR, while 700 commands and 2,500 utterances for 35-command SCR and 250-speaker SR.

Table 2 summarizes the mean and standard deviation of CA, BA, and ASR. We can see that the difference between CA and BA is below 3% for all audio systems, and in some cases, BA even slightly exceeds CA (e.g., X-Vector). This suggests only a negligible impact of our data poisoning on the normal usage of audio systems. Besides, the ASR on SCR systems is over 92% and the ASR on SR systems approaches 100%. The results show that the injected backdoor is successfully activated by the RIR trigger, demonstrating the physical effectiveness of *TrojanRoom*. Meanwhile, we observe a better ASR on SR than that on SCR, since the SR models have more powerful capability and thus learn the backdoor better. Moreover, the ASRs on simple and difficult tasks exhibit insignificant differences, demonstrating the scalability of *TrojanRoom*.

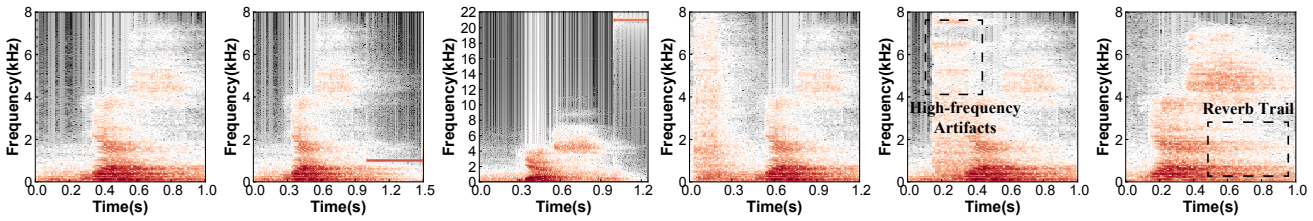
Figure 8 compares the physical performance between *TrojanRoom* and SOTA baselines. We can see that FreqTone and UltraSound achieve excellent BA but result in obvious degradation of ASR, due to channel distortions and the low-pass filtering of audio systems (most audio systems only use frequency bands below 8kHz [10, 11, 22, 31, 50, 63]). Although BackNoise exhibits satisfactory ASR, its BA on Att-RNN and X-Vector significantly declines. This is because the BackNoise triggers are injected at the beginning of speech samples, which greatly impacts the learning of contextual features by RNN and TDNN. Besides, AdvPerturb achieves an ASR of 35%∼98% thanks to its channel compensation, but there is still a large gap compared to the digital performance. Compared to RIR-random which performs poorly in unknown rooms, *TrojanRoom* improves the ASR to 92%∼100% with excellent BA, benefiting from the RIR trigger generation.

6.3 Stealthiness of RIR Trigger

We further conduct objective and subjective experiments to evaluate the stealthiness of different triggers in terms of signal distortion and human perception. For better comparison, we present the spectrum of a benign sample and the corresponding poisoned samples with different triggers in Figure 9.

We calculate MCD between the benign and poisoned samples to measure the objective signal distortion. As shown in Figure 10, FreqTone, UltraSound, and BackNoise achieve satisfactory MCDs of 7.99±1.21dB, 7.23±1.24dB, and 6.20±0.79dB respectively on short speech commands, as well as MCDs of 5.67±0.34dB, 5.41±0.29dB, and 5.54±0.36dB on long speaker utterances. But AdvPerturb shows significant distortion with high MCDs of 13.91±2.71dB and 12.71±1.81dB. This is caused by the evident high-frequency artifacts of adversarial perturbations, as depicted in Figure 9(e). By contrast, RIR achieves the lowest MCDs of 5.69±0.77dB and 5.36±0.28dB. Although the RIR trigger induces reverb trails, as shown in Figure 9(f), it retains the original acoustic structure and therefore exhibits the least distortion.

As for subjective human perception, we recruit 30 volunteers (18 males and 12 females) aged 20∼48 to participate in a MOS test, including a comparison trial and an inspection trial. Note that all subjective experiments conducted on volunteers are validated by the Institutional Review Board (IRB) at our university. In the comparison trial, we ask the volunteers to listen to 30 pairs of benign and poisoned samples with different triggers. For each pair, the volunteers need to assess the speech distortion based on their perceptual experiences



(a) Benign sample (b) FreqTone (c) UltraSound (d) BackNoise (e) AdvPerturb (f) RIR

Figure 9: Example of benign sample (speech command "*yes*") and poisoned samples with different triggers.

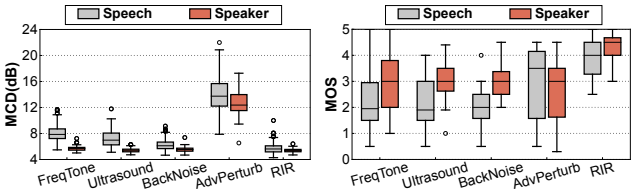


Figure 10: MCD and MOS between benign and poisoned samples with different triggers.

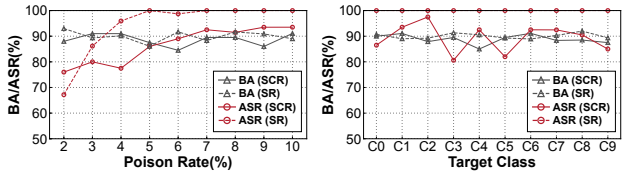


Figure 11: Impact of different poison rates and target classes.

and assign a score ranging from 0 to 5. As shown in Figure 10, the median MOS of RIR-poisoned samples exceeds 4.0, whereas the median MOS of samples poisoned by the other four triggers ranges from 2.0 to 3.0. This validates a better human perception of our RIR trigger.

After a 5-minute rest, we begin the inspection trial by playing 5 pairs of benign and poisoned samples to the volunteers to refresh their hearing and let them establish a sense of poison samples. Then we play 20 samples (5 benign samples and 3 poisoned samples per trigger) in random order and ask the volunteers to identify the poisoned ones. For each sample that is considered poisonous, the volunteers also need to select the injection position from three options: "start", "middle", and "end". Table 3 shows the accuracy of trigger detection and detected position. There are 49.54%∼86.66% of samples poisoned by the four triggers being detected. Moreover, among the detected samples, the trigger positions of 60.08% of FreqTone, 37.65% of UltraSound, and 81.66% of BackNoise are correctly identified. And the detected positions of AdvPerturb show a relatively even distribution, which is consistent with its random trigger injection. These results validate that previous triggers are easily noticeable and are likely to cause attack exposure. Instead, only 21.67% of RIR-injected samples are regarded as poisoned, demonstrating its stealthiness to human inspection.

Table 3: Detection result on samples with different triggers.

Trigger	Detection Accuracy(%)	Detected Position (%)		
		start	middle	end
FreqTone	76.66	10.40	6.18	**60.08**
UltraSound	49.54	8.33	3.56	**37.65**
BackNoise	86.66	**81.66**	5.00	0.00
AdvPerturb	74.39	**21.47**	**36.25**	**16.67**
RIR	**21.67**	4.40	15.60	1.67

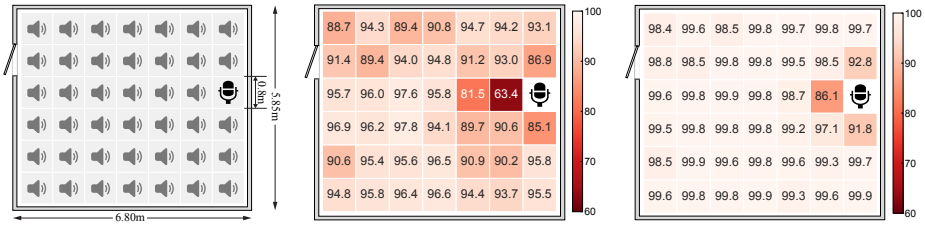
6.4 Micro-benchmarks

Next, we conduct experiments to investigate the impact of micro-benchmarks on *TrojanRoom*. For simplicity, we adopt the 10-command BC-ResNet and 100-speaker Ecapa-TDNN as the target audio systems in this experiment.

Impact of poison rate. We poison the 2%∼10% of the training dataset to study the impact of different poison rates. As shown in Figure 11, as the poison rate increases, the BA remains steady while the ASR grows gradually. We find that at least 6% and 5% of poison rates are needed to achieve good ASR on BC-ResNet and Ecapa-TDNN.

Impact of target class. We define 10 commands and 10 speakers as the target class to launch attacks. As shown in Figure 11, the BA and ASR change over different commands while remaining steady across different speakers. This is because the SR model is more powerful and able to learn robust connections between the RIR trigger and different targets.

Impact of activate position. As shown in Figure 12, we fix the microphone and change the speaker to 41 different locations in Room C (6.80m×5.85m×2.70m) to launch attacks as before. Despite the small ASR degradation of about 2%∼20% and 7%∼12% for SCR and SR at closer positions with weaker reverberation, *TrojanRoom* shows high ASR at different positions, even at a distance of over 5m, verifying the excellent effectiveness of RIR triggers.



(a) Setup (b) ASR of 10-command SCR (c) ASR of 100-speaker SR

Figure 12: Impact of activation position.

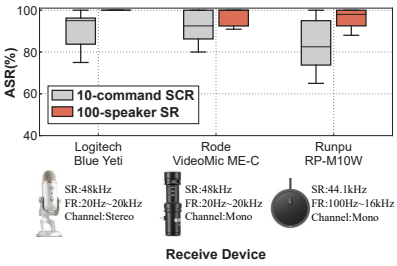
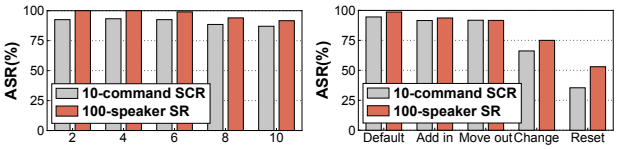


Figure 13: Impact of victim device.



(a) Number of human subjects (b) Furniture layout

Figure 14: Impact of human subject and furniture layout.

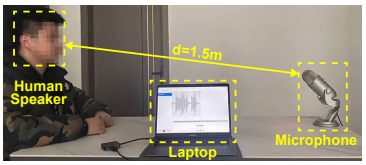


Figure 16: Experimental setup of human study.

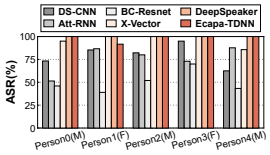
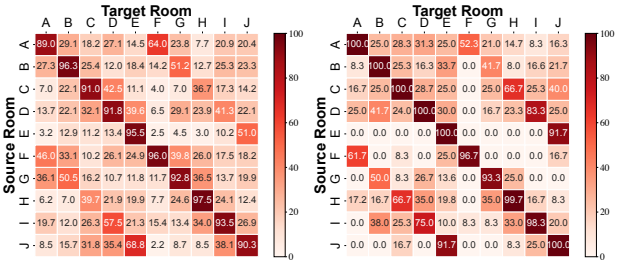


Figure 17: ASR of different human speakers.



(a) ASR of 10-command SCR (b) ASR of 100-speaker SR

Figure 15: ASR across similar and dissimilar rooms.

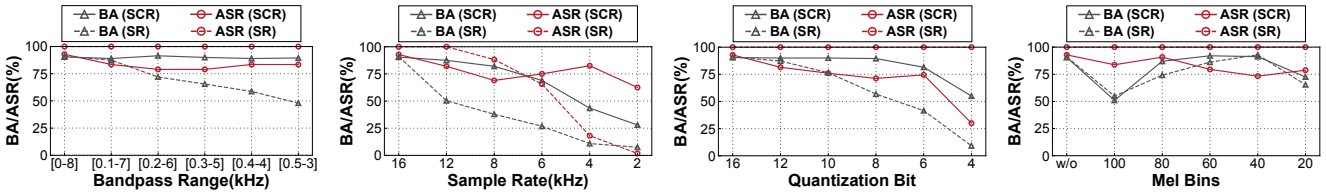
Impact of victim device. We use three microphones from different brands as the receiver. As shown in Figure 13, the median ASRs for SCR and SR on all three devices are over 80% and 98% respectively, showing the robustness to different devices. We also notice better performance on high-end devices (e.g., Logitech Blue Yeti) due to their stronger acquisition capability and less device distortion.

Impact of human subject and furniture layout. To investigate the effects of reverberation variations, we consider 2∼10 human subjects moving at a speed of 1m/s and 5 furniture layouts in Room C, including a default layout, adding or removing chairs, changing the placement of desks and chairs, and resetting the entire layout. As shown in Figure 15, the ASR slightly decreases with more subjects, and the ASR of "Add in" and "Move out" is similar to that of the default layout, while "Change" and "Reset" induce a great ASR drop due to their mismatched reverberation patterns. These results motivate the adversary to select appropriate attack locations and occasions with fewer subjects and stable layouts.

Cross-room study. We further transfer *TrojanRoom* across different rooms to study the attack sensitivity. Except for Room A-E which are dissimilar to each other, we try to prepare another 5 rooms (Room F-J) that have similar dimensions and wall materials to Room A-E respectively, i.e., Room F is similar to Room A, Room G is similar to Room B, and so on. To simplify the data collection, we only measure RIRs in each room and convolve them with raw speeches to simulate reverberation. Then we test audio systems infected by the RIR of a source room on samples poisoned by the RIR from a target room. As shown in Figure 15, we can see the ASR across similar rooms remains 36.7%∼68.8% and 41.7%∼91.7% for SCR and SR respectively, while the ASR across dissimilar rooms significantly degrades. This indicates that our RIR trigger is room-specific and it is possible to transfer this attack to highly similar rooms, e.g., Room E and J.

6.5 Live-speech Attack

Instead of playing recorded speech through a device, we further recruit human speakers to perform live-speech attacks to verify injection-free activation. We invite five volunteers (three males and two females) as the adversaries. Each volunteer is asked to speak 200 fixed short commands and 50 arbitrary long utterances into a microphone located 1.5m away, as shown in Figure 16. These live speeches are sent to the laptop for evaluation. Note that all of the volunteer samples are out of the training set. As shown in Figure 17, the average ASRs of human speakers on the SCR and SR systems are 43.39%∼95.05% and 85.71%∼100.00% respectively. This validates the practicality of *TrojanRoom* with injection-free activation, even by adversaries outside of the training set.



(a) Band-pass Filtering (b) Re-sampling (c) Re-quantization (d) Mel Extraction-Inversion

Figure 18: BA and ASR under trigger disruption with different filters.

7 Defense Evasion

7.1 Source-level Liveness Detection

We evaluate our injection-free activation against voice liveness detection systems, including VOID [1] and LCNN [30].

VOID is a lightweight yet effective detector that relies on the distinct signal power distribution patterns of replayed and live voices over the audible frequency range. LCNN is a lightweight version of CNN featured by max feature mapping function (MFM), which ranks first in the ASVspoof 2017 challenge [24]. We follow the original papers to reproduce these two detectors on the ASVspoof 2017 physical access (PA) dataset. For VOID, we extract 48 FVLFPs, 2 FVLDFs, 35 FVHPFs, and 12 FVLPCs to create 97-dimensional features, which are used to train an SVM classifier with an RBF kernel. As for LCNN, we apply STFT to derive spectrums and train the LCNN model with the recommended hyper-parameters. Then, we utilize the hidden vector of the second-to-last linear layer (FC6) as high-level features for training the GMM classifier.

With the two detectors, we first test the 13,306 evaluation trials (1,298 genuine and 12,008 spoof samples) from ASVspoof 2017 PA, and then test the live-speech trials (1,250 benign and 1,250 poisoned samples) of injection-free *TrojanRoom*. As shown in Table 4, VOID and LCNN achieve an Equal Error Rate (EER) of 19.20% and 12.73% on ASVspoof 2017, respectively, indicating their good spoof detection performance. However, the EERs on *TrojanRoom* increase to 42.34% and 39.35% respectively, suggesting that the liveness detectors cannot distinguish poisoned samples from benign ones. This is because our RIR trigger is convolved with the live speech over the air during the injection-free activation, instead of being replayed through a transmission device.

Table 4: EER(%) of SOTA voice liveness detection systems on test trials from ASVspoof 2017 and *TrojanRoom*.

Data Source	#Trial	VOID	LCNN
ASVspoof 2017 (PA eval)	13,306	19.20	12.73
TrojanRoom (injection-free)	2,500	42.34	39.35

7.2 Data-level Trigger Disruption

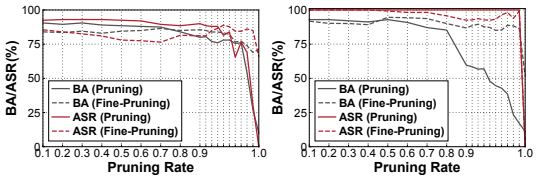
We also consider the data-level trigger disruption, where the defender aims to filter out the RIR trigger from speech using signal processing techniques. Specifically, we apply four widely used audio filters, including Band-pass Filtering, Downsampling-Upsampling, Quantization-Dequantization, and Mel Extraction-Inversion, as introduced in WaveGuard [19], on benign and poisoned samples for evaluation.

As shown in Figure 18(a), the ASRs on SCR (BC-ResNet) and SR (Ecapa-TDNN) still remain above 75.00% and 99.00% as the rest of the frequency range narrows. Besides, the ASR decreases as the sampling rate and quantization bit get smaller as shown in Figure 18(b) and 18(c). But the BA declines at the same time so that the filters cannot effectively disrupt our RIR trigger without impairing the normal usage of audio systems. Finally, the Mel Extraction-Inversion has a negligible impact on the ASR in the SR task and but ASR of SCR degrades to 73.25%. All these results validate the robustness of *TrojanRoom* against trigger disruption defense.

7.3 Model-level Backdoor Defense

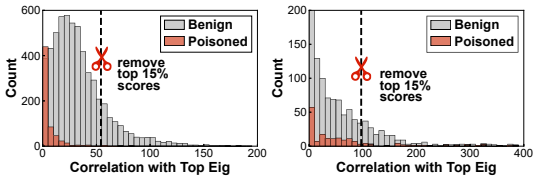
Finally, we consider common model-level defenses that aim to detect and erase backdoors in the infected model.

Neural pruning and fine-pruning [35]. We first use the clean validation dataset to calculate the average activation level of neurons at the last convolution layer in BC-ResNet and Ecapa-TDNN, and then prune a portion of low-activation neurons. Finally, we fine-tune the infected models for 30 epochs with a small learning rate to restore their benign performance. As shown in Figure 19, we can observe that the BA and ASR only decrease slightly at a pruning rate from 0.1 to 0.8 since the pruned neurons with low activation levels are of less significance. As the punning rate grows from 0.9 to 1.0 with a smaller step of 0.01, both the BA and ASR after pruning decrease rapidly due to the loss of important neurons. But after fine-tuning, both the BA and ASR remain high until all the neurons are pruned since the remaining neurons are fine-tuned to restore performance. This result suggests that the infected and benign connections share the same part of neurons, so current pruning and fine-pruning cannot remove the backdoor without compromising benign performance.



(a) 10-command SCR (b) 100-speaker SR

Figure 19: BA and ASR after neural pruning with different pruning rates.



(a) 10-command SCR (b) 100-speaker SR

Figure 20: Correlations of spectral signatures with the top singular vector.

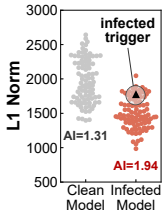


Figure 21: L_1 norm of triggers.

Spectral signature [55]. We first extract high-level representations from the last convolution layer of BC-ResNet and Ecapa-TDNN on the training dataset of the target class. Then, we apply singular value decomposition to the covariance matrix of these representations and compute their correlation with the top singular vector. The samples with the top 15% scores are treated as outliers and discarded [55]. As shown in Figure 20, the correlation scores of poisoned samples mainly distribute below 50 and 100 for SCR and SR, which are mixed with benign ones with a low average detection precision of 0.06 and 0.15 respectively. This indicates that the RIR trigger-enabled backdoor would not cause significant variations in high-level features. Hence, only a few RIR-poisoned samples (1.2% for SCR and 18.9% for SR) are detected and filtered out by this defense. However, implementing more aggressive filtering would impair the benign performance, resulting in reduced defense effectiveness.

Neural cleanse [56]. As noted by the authors in [56], detecting our partial backdoor requires reverse engineering triggers for all possible source-target label pairs. This requires A_{10}^2=90 and A_{100}^2=9,900 times of optimization for 10-command SCR and 100-speaker SR respectively. To simplify the evaluation, we only focus on simpler SCR here, i.e., BC-ResNet. During the reverse engineering of the RIR trigger, we find that the optimization does not converge, since the defense strategy only considers additive pixel patterns as image triggers. Hence, we try to optimize minimal additive noise-based triggers and apply the median absolute deviation to detect outliers in their L_1 norm. Figure 21 presents the distribution of reversed triggers' L_1 norm for both clean and infected models. We can see that the infected model produces smaller L_1 norms due to the "shortcuts" created by backdoors, but its Anomaly Index (AI) is 1.94, which is still lower than the detection threshold of 2 as defined in [56]. Moreover, the L_1 norm of the infected trigger is larger than that of most candidate triggers, making it to fail to detect the poisoned source-target pair.

8 Discussion

Potential countermeasure. We have evaluated different levels of defenses and demonstrated their insufficient resistance to *TrojanRoom*, owing to the lack of dedicated countermeasures against the convolutional RIR triggers. To this end, we further consider Adaptive Echo Cancellation (AEC) as a defensive filter, which is exclusively designed to cancel speech echoes and is expected to eliminate RIRs. As shown in Figure 25 and analyzed in Appendix E, the NLMS-based AEC could cancel RIR triggers to undermine *TrojanRoom* in some cases, while at the cost of benign accuracy and requiring unavailable reference inputs. Therefore, it is necessary to explore more effective and readily available defenses in the feature.

Attack practicality. Despite revealing the threat of physical audio backdoor attacks, there are several limitations in the practicality of *TrojanRoom*. First, the RIR trigger is room-dependent so that our attack only targets specific rooms predefined by the adversary. Second, we relax the assumption about the adversary by enabling injection-free activation but still requires prior information about the room configuration, which may not be available in some constrained cases. Finally, as evaluated in Section 6.4, *TrojanRoom* is not very tolerant to acoustic variations, especially if the room's reverberation characteristics change significantly due to furniture layout reset, limiting its real-world practicality.

Methodology generalization. Successful backdoor activation and stealthy trigger injection in specific rooms suggest the superiority of RIR as a physical trigger. However, directly transferring this attack across rooms yields insufficient and unreliable attack performance. For more advanced and flexible scenarios that target multiple rooms or require fine-grained activation control, *TrojanRoom* can be extended to "One-to-N" or "N-to-One" attacks [59] by introducing multi-target or multi-trigger designs.

9 Conclusion

This paper investigates the practical issues of physical audio backdoor attacks and proposes a novel RIR trigger to turn the sound channel into a natural trigger injection path. Based on this, we propose *TrojanRoom*, a practical audio backdoor attack in the physical world. After generating dynamic RIR triggers and poisoning samples during the data augmentation, *TrojanRoom* can activate the backdoor in an injection-free manner. The evaluation shows the physical practicality of *TrojanRoom* and its resilience to different levels of defenses.

Acknowledgments

We would like to thank all reviewers for their insightful and constructive suggestions. This research is sponsored by National Natural Science Foundation of China (62102354, 62032021, 62172359, 61972348), Fundamental Research Funds for the Central Universities (2021FZZX001-27).

References

[1] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. Void: A fast and light voice liveness detection system. In *Proceedings of USENIX Security Symposium*, pages 2685–2702, virtual event, 2020.

[2] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[3] Amazon. Alexa features. https://www.amazon.com/b?node=21576558011, 2023.

[4] Apple. Apple HomeKit. https://www.apple.com/home-app, 2023.

[5] HSBC Bank. HSBC Voice ID. https://www.hsbc.com.hk/ways-to-bank/phone/voice-id, 2023.

[6] TD Bank. TD VoicePrint. https://www.tdbank.com/bank/tdvoiceprint.html, 2023.

[7] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proceedings of IEEE ICCV*, pages 2764–2773, Venice, Italy, 2017. Computer Society.

[8] Brilliant. Brilliant Smart Home System. https://www.brilliant.tech/pages/smart-home-control, 2023.

[9] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against NLP models with semantic-preserving improvements. In *Proceedings of ACM ACSAC*, pages 554–569, Virtual Event, USA, 2021.

[10] Douglas Coimbra de Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. A neural attention model for speech command recognition. *CoRR*, abs/1808.08929, 2018.

[11] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of ISCA Interspeech*, pages 3830–3834, Virtual Event, Shanghai, China, 2020.

[12] Carl F Eyring. Reverberation time in "dead" rooms. *The Journal of the Acoustical Society of America*, 1(2A):217–241, 1930.

[13] Richard P. Feynman, Robert B. Leighton, and Matthew Sands. The new millennium edition: Mainly mechanics, radiation, and heat. In *The Feynman lectures on physics*, volume 1. Basic books, 2011.

[14] Google. Cloud ML. https://cloud.google.com/solutions/ai, 2023.

[15] Google. Google Assistant. https://assistant.google.com, 2023.

[16] CMU Speech Group. Statistical parametirc sythesis and voice conversion techniques. http://festvox.org/11752/slides/lecture11a.pdf, 2012.

[17] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[18] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Proceedings of NeurIPS*, pages 5767–5777, Long Beach, CA, USA, 2017.

[19] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian J. McAuley, and Farinaz Koushanfar. Waveguard: Understanding and mitigating audio adversarial examples. In *Proceedings of USENIX Security Symposium*, pages 2273–2290, virtual event, 2021.

[20] JCWA. Absorption coefficients of common building materials and finishes. https://www.acoustic-supplies.com/absorption-coefficient-chart, 2023.

[21] Marco Jeub, Magnus Schäfer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of IEEE DSP*, pages 1–5, Santorini, Greece, 2009.

[22] Byeonggeun Kim, Simyung Chang, Jinkyu Lee, and Dooyong Sung. Broadcasted residual learning for efficient keyword spotting. In *Proceedings of ISCA Interspeech*, pages 4538–4542, Brno, Czechia, 2021.

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of ICLR*, Banff, AB, Canada, 2014.

[24] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proceedings of ISCA Interspeech*, pages 2–6, Stockholm, Sweden, 2017.

[25] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP J. Adv. Signal Process.*, 2016:7, 2016.

[26] Stefanos Koffas, Stjepan Picek, and Mauro Conti. Dynamic backdoors with global average pooling. *CoRR*, abs/2203.02079, 2022.

[27] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. Can you hear it?: Backdoor attacks via ultrasonic triggers. In *Proceedings of ACM WiseML@WiSec*, pages 57–62, San Antonio, TX, USA, 2022.

[28] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of IEEE/CVF CVPR*, pages 298–307, Seattle, WA, USA, 2020.

[29] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In *Proceedings of ACL*, pages 2793–2806, Online, 2020.

[30] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proceedings of ISCA Interspeech*, pages 82–86, Stockholm, Sweden, 2017.

[31] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *CoRR*, abs/1705.02304, 2017.

[32] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. In *Proceedings of ACM CCS*, pages 3123–3140, Virtual Event, Republic of Korea, 2021.

[33] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, pages 1–18, 2022.

[34] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of IEEE/CVF ICCV*, pages 16443–16452, Montreal, QC, Canada, 2021.

[35] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of Springer RAID*, volume 11050 of *Lecture Notes in Computer Science*, pages 273–294, Heraklion, Crete, Greece, 2018.

[36] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of The Internet Society NDSS*, San Diego, California, 2018.

[37] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proceedings of Springer ECCV*, volume 12355 of *Lecture Notes in Computer Science*, pages 182–199, Glasgow, UK, 2020.

[38] Microsoft. Azure. https://azure.microsoft.com/en-us, 2023.

[39] MonkeyLearn. No-code text analytics. https://monkeylearn.com, 2023.

[40] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of ELRA LREC*, Athens, Greece, 2000.

[41] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *Proceedings of NeurIPS*, virtual, 2020.

[42] Bank of Scotland. Voice ID - fast, easy and safe. https://www.bankofscotland.co.uk/contactus/phone/voice-id, 2023.

[43] Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation. In *Proceedings of USENIX Security*, pages 3611–3628, 2022.

[44] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of IEEE ICASSP*, pages 5206–5210, South Brisbane, Queensland, Australia, 2015.

[45] Ren Pang, Hua Shen, Xinyang Zhang, Shouling Ji, Yevgeniy Vorobeychik, Xiapu Luo, Alex X. Liu, and Ting Wang. A tale of evil twins: Adversarial inputs versus poisoned models. In *Proceedings of ACM CCS*, pages 85–99, Virtual Event, USA, 2020.

[46] Samsung. Samsung SmartThings Hub. https://www.smartthings.com/getting-started, 2023.

[47] Robin Scheibler, Eric Bezzam, and Ivan Dokmanic. Pyroomacoustics: A python package for audio room simulations and array processing algorithms. *CoRR*, abs/1710.04196, 2017.

[48] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of ACM MobiCom*, pages 583–595, Sydney, NSW, Australia, 2022.

[49] D.T.M. Slock. On the convergence behavior of the lms and the normalized lms algorithms. *IEEE Trans. Signal Process.*, 41(9):2811–2825, 1993.

[50] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of IEEE ICASSP*, pages 5329–5333, Calgary, AB, Canada, 2018.

[51] Igor Szöke, Miroslav Skácel, Ladislav Mosner, Jakub Paliesek, and Jan Honza Cernocký. Building and evaluation of a real room impulse response dataset. *IEEE J. Sel. Top. Signal Process.*, 13(4):863–876, 2019.

[52] Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of ACM SIGKDD*, pages 218–228, Virtual Event, CA, USA, 2020.

[53] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. GWA: A large high-quality acoustic dataset for audio processing. In *Proceedings of ACM SIGGRAPH*, pages 36:1–36:9, Vancouver, BC, Canada, 2022.

[54] Micah T. Taylor, Anish Chandak, Lakulish Antani, and Dinesh Manocha. Resound: interactive sound rendering for dynamic virtual environments. In *Proceedings of ACM Multimedia*, pages 271–280, Vancouver, British Columbia, Canada, 2009.

[55] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Proceedings of NeurIPS*, pages 8011–8021, Montréal, Canada, 2018.

[56] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of IEEE SP*, pages 707–723, San Francisco, CA, USA, 2019.

[57] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *CoRR*, abs/1804.03209, 2018.

[58] Zhaohan Xi, Ren Pang, Shouling Ji, and Ting Wang. Graph backdoor. In *Proceedings of USENIX Security*, pages 1523–1540, 2021.

[59] Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Trans. Dependable Secur. Comput.*, 19(3):1562–1578, 2022.

[60] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of ACM CCS*, pages 2041–2055, London, UK, 2019.

[61] Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Backdoor attack against speaker verification. In *Proceedings of IEEE ICASSP*, pages 2560–2564, Toronto, ON, Canada, 2021.

[62] Xinyang Zhang, Zheng Zhang, Shouling Ji, and Ting Wang. Trojaning language models for fun and profit. In *Proceedings of IEEE EuroS&P*, pages 179–197, Vienna, Austria, 2021.

[63] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. Hello edge: Keyword spotting on microcontrollers. *CoRR*, abs/1711.07128, 2017.

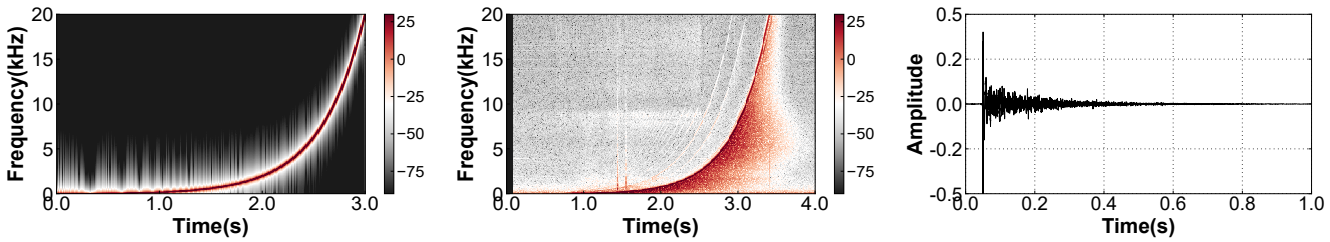
[64] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of ACM SACMAT*, pages 15–26, Virtual Event, Spain, 2021.

A ESS-based RIR Measurement

Exponential Sine Sweep (ESS) is a typical RIR measurement method, which transmits a band-limited sinusoidal signal with frequency exponentially varying from f_1 to f_2:

$$x(t) = \sin\left[\frac{2\pi f_1 T}{\ln\frac{f_2}{f_1}}\left(e^{\frac{t}{T}\ln\frac{f_2}{f_1}} - 1\right)\right], \tag{11}$$

where T refers to the signal duration. Figure 22(a) shows the spectrum of the ESS signal we use (f_1=20Hz, f_2=20kHz, T=3s), and Figure 22(b) shows the received response $y(t)$. Then we can obtain the RIR through the ESS deconvolution, i.e., convolving the measured response $y(t)$ with the time-reversal of the test signal $x(-t)$. Figure 22(c) shows the measured RIR for data poisoning in the feasibility study.



(a) Transmitted ESS signal (b) Received ESS signal (c) Measured RIR

Figure 22: Transmitted ESS, received ESS and measured RIR in the feasibility study.

Table 5: CVAE-GAN network configuration.

Module	Block	Input → Output	Layer Specification
Encoder	5×DownsampleBlock	[B×1×T] → [B×256×T/256]	ResBlock(kernel=3, stride=1) Conv1D(kernel=25, stride=4) LeakyReLU(p=0.2)
	Mean Vector Covariance Vector	[B×4096] → [B×256] [B×4096] → [B×256]	Linear(in_channels=4096, out_channels=256) Linear(in_channels=4096, out_channels=256)
Generator	Latent Vector	[B×(256+10)] → [B×4096]	Linear(in_channels=266, out_channels=4096)
	5×UpsampleBlock	[B×256×T/256] → [B×1×T]	TransConv1D(kernel=25, stride=4) LeakyReLU(p=0.2) ResBlock(kernel=3, stride=1)
	Output Layer	[B×1×T] → [B×1×T]	TransConv1D(kernel=25, stride=1) Tanh
Discriminator	5×DownsampleBlock	[B×1×T] → [B×256×T/256]	ResBlock(kernel=3, stride=1) Conv1D(kernel=25, stride=4) LeakyReLU(p=0.2)
	BottleNeck	[B×4096] → [B×256]	Linear(in_channels=4096, out_channels=256) LeakyReLU(p=0.2)
	Output Layer	[B×256] → [B×1]	Linear(in_channels=256, out_channels=1)

B CVAE-GAN Network Configuration

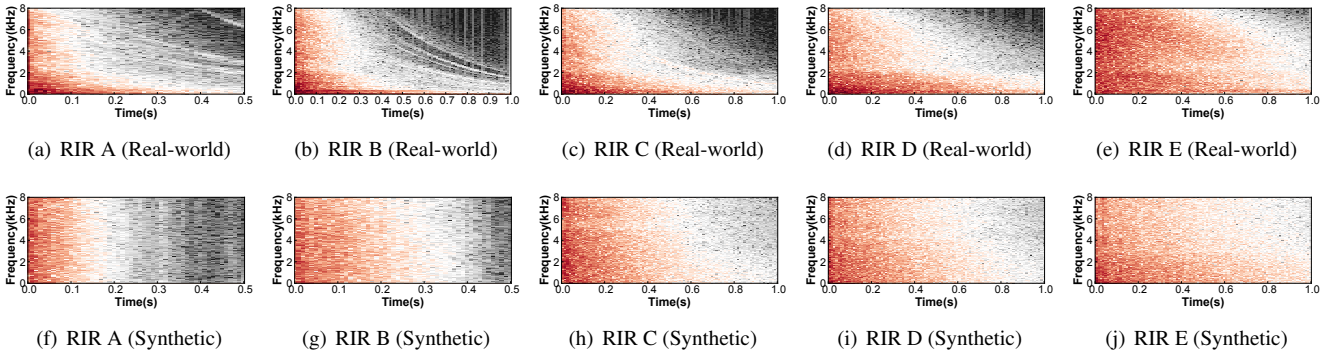
The CVAE-GAN network configuration is shown in Table 5. Each batch of real-world RIR samples is padded into a B×1×T tensor. The encoder squeezes the temporal resolution of the input 256 times while expanding its frequency resolution 256 times using 5 downsample blocks. Each block consists of a residual block and a 1D large-kernel convolution layer with LeakyReLU activation. The squeezed vector is used to learn the mean and covariance vectors through two parallel linear layers. After reparameterization, the mean and covariance vectors are converted to a latent vector and fed to the generator. The generator expands the latent vector to B×1×T through 5 upsample blocks and then reconstructs the RIRs with the output layer. Here we adopt the Tanh function to restrict the range of synthetic RIRs within $[-1,1]$. The discriminator has the same network as the encoder, except for the bottleneck and output layers. To avoid mode collision, we do not use batch normalization in all layers.

C Room Parameters and Synthetic RIRs

Table 6 shows the acoustic parameters of Room A-E, including the room dimensions and surface absorption coefficients derived from the material database. We select these 5 rooms with distinct reverberation times ranging from 0.56s to 1.97s. These parameters are used to estimate RT_{60}^{est}, which shows a relatively small estimation error compared to the ground truth RT_{60}^{gr}. With the acoustic parameters and estimated RT_{60}^{est}, we use the conditional vector to generate 100 synthetic RIRs for each room using the trained generator of CVAE-GAN. We use RT_{60} *Error*: $\|RT_{60}^{syn} - RT_{60}^{gr}\|$, to measure the difference in reverberation time between synthetic and ground truth RIRs. As shown in Table 6, all the RT_{60} errors are less than 0.95s, indicating satisfactory generation performance. We also noticed that a longer reverberation time induces a larger RT_{60} error. This is because there are fewer long-reverberation training samples in BUT_ReverbDB and GWA. We also present the synthetic and real-world RIRs of Room A-E in Figure 23.

Table 6: Acoustic parameters and RT_{60} error between synthetic RIRs and ground truth RIRs of Room A-E.

Room	Dimension(m)	Surface Absorption Coefficients(at 1kHz)						RT_{60}^{est}(s)	RT_{60}^{gr}(s)	RT_{60} Error(s)
		front	back	left	right	ceiling	floor			
A	$3.46 \times 3.18 \times 2.65$	0.08	0.08	0.03	0.10	0.03	0.20	0.61	0.56	0.062±0.049
B	$6.15 \times 4.86 \times 2.72$	0.06	0.03	0.08	0.08	0.65	0.12	0.83	0.81	0.054±0.027
C	$6.80 \times 5.85 \times 2.70$	0.02	0.02	0.02	0.02	0.03	0.43	1.23	1.29	0.083±0.041
D	$7.42 \times 6.13 \times 3.08$	0.04	0.08	0.04	0.04	0.03	0.12	1.64	1.61	0.095±0.063
E	$9.39 \times 7.20 \times 3.10$	0.05	0.05	0.05	0.08	0.15	0.12	1.94	1.97	0.087±0.059



(a) RIR A (Real-world) (b) RIR B (Real-world) (c) RIR C (Real-world) (d) RIR D (Real-world) (e) RIR E (Real-world)

(f) RIR A (Synthetic) (g) RIR B (Synthetic) (h) RIR C (Synthetic) (i) RIR D (Synthetic) (j) RIR E (Synthetic)

Figure 23: Real-world and synthetic RIRs of Room A-E.

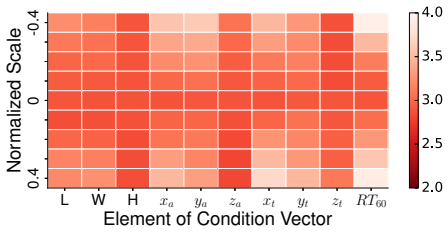


Figure 24: L_2 loss between synthetic and real-world RIRs after changing each element of condition vector.

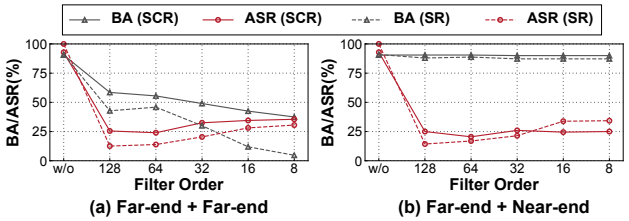


Figure 25: BA and ASR after applying NLMS-based AEC.

D RIR Generation Study

In addition to studying RIR generation in a limited number of rooms, we further traverse the condition vector to understand the impact of each element on the RIR quality. Specifically, we extract the condition vector of a subset of GWA RIRs and then vary each element by a normalized scale ranging from -0.4 to 0.4 to sample RIRs using the generator. Then, we calculate the average L_2 loss between the synthetic and real-world RIRs. As shown in Figure 24, we can observe that the L_2 loss remains steady when H, z_a, and z_t vary, while other elements cause a larger L_2 loss than the initial condition vector with scale=0. This result reveals that the height information is not significant, while other elements such as room dimension, speaker/receiver locations, and reverberation time are crucial for RIR generation.

E Impact of NLMS-based AEC

Given a reference input $x(n)$, we utilize the Normalized Least Mean Square (NLMS) algorithm [49] to update the adaptive filter for minimizing the difference between the filtered output $y(n)$ and the echoed signal $d(n)$. Then we apply this filter with different orders to both benign and poisoned samples to evaluate the impact of NLMS-based AEC. As shown in Figure 25, the BA and ASR of infected models degrade a lot when the benign and poisoned samples are both far-end signals, while the BA remains steady and the ASR drops significantly on near-end benign samples and far-end poisoned samples. This interesting result suggests that NLMS-based AEC could cancel the RIR trigger to resist *TrojanRoom*, at the cost of benign accuracy or in a limited setting. Moreover, this filter requires an over-the-line input as a reference for optimization [49], which is not available during physical interaction between the adversary and the system.