Enrollment-Stage Backdoor Attacks on Speaker Recognition Systems via Adversarial Ultrasound

Xinfeng Li, Junning Ze, Chen Yan, *Member, IEEE*, Yushi Cheng, Xiaoyu Ji, *Member, IEEE*, and Wenyuan Xu

***Abstract*—Automatic speaker recognition systems (SRSs) have been widely used in voice applications for personal identification and access control. A typical SRS consists of three stages, i.e., training, enrollment, and recognition. Previous work has revealed that SRSs can be bypassed by backdoor attacks at the training stage or by adversarial example attacks at the recognition stage. In this article, we propose TUNER, a new type of backdoor attack against the enrollment stage of SRS via adversarial ultrasound modulation, which is inaudible, synchronization-free, content-independent, and black-box. Our key idea is to first inject the backdoor into the SRS with modulated ultrasound when a legitimate user initiates the enrollment, and afterward, the polluted SRS will grant access to both the legitimate user and the adversary with high confidence. Our attack faces a major challenge of unpredictable user articulation at the enrollment stage. To overcome this challenge, we generate the ultrasonic backdoor by augmenting the optimization process with random speech content, vocalizing time, and volume of the user. Furthermore, to achieve real-world robustness, we improve the ultrasonic signal over traditional methods using sparse frequency points, precompensation, and single-sideband (SSB) modulation. We extensively evaluate TUNER on two common data sets and seven representative SRS models, as well as its robustness against seven kinds of defenses. Results show that our attack can successfully bypass SRSs while remaining effective to various speakers, speech content, etc. To mitigate this newly discovered threat, we also provide discussions on potential countermeasures, limitations, and future works of this new threat.**

***Index Terms*—Adversarial ultrasound, backdoor attack, enrollment, speaker recognition.**

I. INTRODUCTION

AUTOMATIC speaker recognition systems (SRSs) can identify and authenticate human users by their voices. These systems have been increasingly used in voice assistants and online banking for ensuring that access to critical service and information is only granted to legitimate users [1]. Despite their convenience, SRSs are vulnerable to various types of attacks. For example, voice spoofing attacks use recorded or synthesized voice samples of legitimate users to deceive the system [2], [3]. Recently, other works have revealed that SRSs are also vulnerable to backdoor attacks [4], [5], [6], [7] and adversarial example (AE) attacks [8], [9], [10]. However, most backdoor attacks rely on a challenging prerequisite of accessing the training data of the target system; spoofing attacks and AE attacks require generating human-audible sounds, which may be noticed by the victim users when they are nearby, and some attacks even require white-box knowledge that is not available for commercial systems.

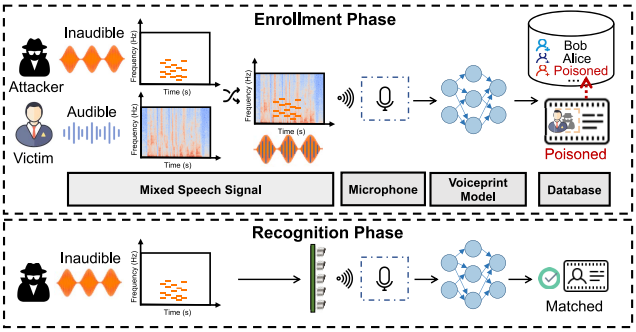


Fig. 1. Illustration of TUNER. An adversary emits backdoors modulated on ultrasound when a victim initiates the voiceprint enrollment. Such poisoned voiceprint recorded by the SRS enables both the adversary and the legitimate user to pass subsequent recognition with relatively high confidence. Particularly, the adversary leverages the inaudible adversarial ultrasound to conduct the backdoor attack in a human-imperceptible manner.

In this article, we propose a backdoor attack named TUNER[1] that can bypass commercial SRSs without accessing the training data, requiring white-box system knowledge, or alarming the victim user. Different from most of the existing backdoor attacks that target the training stage of SRS models, this article focuses on the feasibility of backdoor attacks in the enrollment stage, i.e., when legitimate users register their voices for the first time. As shown in Fig. 1, the adversary attempts to inject a backdoor into the victim's voice during the user enrollment and poison the voiceprint stored in the speaker database. Afterward, the contaminated voiceprints can enable both the legitimate user and the adversary to pass the speaker recognition with high confidence.

To achieve such an attack, the first requirement we need to meet is to avoid raising user awareness when injecting

Manuscript received 7 April 2023; revised 5 August 2023 and 30 September 2023; accepted 16 October 2023. Date of publication 30 October 2023; date of current version 9 April 2024. This work was supported in part by the China NSFC under Grant 62201503, Grant 61925109, Grant 62222114, Grant 62071428, and Grant 62271280; and in part by the Fundamental Research Funds for the Central Universities under Grant 226-2022-00223. *(Corresponding author: Chen Yan.)*

The authors are with the USSLAB, Zhejiang University, Hangzhou 310058, China (e-mail: xinfengli@zju.edu.cn; zjning@zju.edu.cn; yanchen@zju.edu.cn; yushicheng@zju.edu.cn; xji@zju.edu.cn; wyxu@zju.edu.cn).

Digital Object Identifier 10.1109/JIOT.2023.3328253

[1]We term it "TUNER" because the backdoor pattern is like a combination of different tones carefully designed by a tuner. Our demo page: https://letterligo.github.io/Tuner/.

the backdoor into the voiceprint enrollment, because the victim user will be using the SRS at the same time and can easily notice suspicious sounds from the attacker. We manage to achieve a completely human-inaudible backdoor injection based on the means of inaudible voice attacks [11], which can make microphones receive audible sounds by emitting inaudible ultrasounds that are beyond the human auditory range (20 Hz–20 kHz). Nevertheless, our investigation reveals that injecting an effective voiceprint backdoor using ultrasound is nontrivial because the backdoor signal will be significantly distorted by the nonlinear distortion and hardware instability during the transmission process. To cope with the signal distortion, we adapt the backdoor design strategy with ultrasound characteristics and optimize the backdoor signals as a form of combined simple tones at sparse frequency points, as shown in Fig. 1. In specific, we first initialize TUNER with a modest number of frequencies and select the most effective ones among them. It is worth noting that the selection (i.e., sparse frequency coding) is automatically achieved using the L1-norm regularization term [12] during the whole optimization process. This design strategy can better mitigate signal distortion during the ultrasound transmission than existing audible backdoor triggers [13] that have a more complex spectrum similar to human voices. We also employ precompensation and single-sideband (SSB) modulation to mitigate signal distortion and boost the transmission efficiency in backdoor delivery.

In addition to the signal distortion, materializing the backdoor attack in real-world scenarios faces a few more challenges.

1) During a real-world enrollment, the utterances that the victim user will vocalize are unpredictable, thus the backdoor needs to be content-agnostic.
2) The voiceprint models deployed in commercial SRSs are generally black-box, i.e., the adversary has no knowledge of the model parameters or the gradient information.
3) The attack configuration needs to be robust to variations of the victim system location, the loudness of the legitimate user's vocalizations, etc.

To overcome the above issues, we first augment the optimization process of ultrasonic backdoors by introducing randomness into the user's speech content and time. As such, TUNER can be applied with a wide variety of configurations in a content-agnostic and synchronization-free manner. Moreover, we adopt the natural evolution strategies (NESs) [14] to estimate the gradient via querying the black-box SRS models. We further improve the robustness of our attack by taking into consideration the ultrasound attenuation varying with distances as well as the loudness relationship between the recovered backdoor and the legitimate user speech.

We validate TUNER on seven representative SRS models, including ECAPA-TDNN [15], Pyannote [16], U-Level [17], WavLM-Xvec [18], SpeakerNet [19], D-vector [20], and Sinc-Xvec [21] along with two typical speech data sets (VoxCeleb1 [22] and LibriSpeech [23]). Results show that TUNER is effective in deceiving SRS and robust to various impact factors, such as different backdoor duration, speakers, speech content, and SRS models. Meanwhile, we also conduct a series of experiments to examine TUNER's performance in the physical world and results demonstrate that TUNER can work well under varying environments, attack distances, angles, and recording devices. In addition, we examine TUNER's resistance to seven representative audio signal preprocessing- and inaudible voice attack-based defenses, under the naive and adaptive adversary attack settings. Our defense experiments justify TUNER can directly bypass almost all defenses, and successfully survive challenging median filter by utilizing adaptive attack strategy. To mitigate this new threatening attack, we discuss several potential countermeasures and reveal its limitations, which may inspire future works.

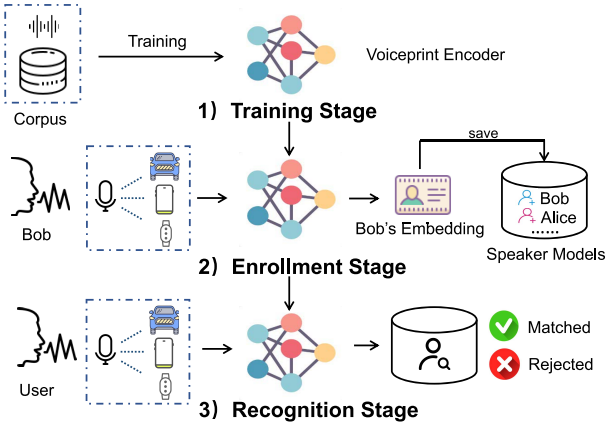


Fig. 2. Illustration of a typical automatic SRS, which consists of three main stages, 1) training stage, 2) enrollment stage, and 3) recognition stage.

In summary, our contributions are listed as follows.

1) We propose a new type of inaudible backdoor attack framework against SRSs via adversarial ultrasound and call for attention to a new attack surface.
2) We introduce a collection of augmentation mechanisms to optimize the ultrasonic backdoor, which enables our attacks to maintain effective regardless of the victim speaker, speech content, speech volume, or attack distances.
3) We propose efficient frequency sparsification, precompensation, and SSB modulation techniques to mitigate the signal distortion of ultrasonic attacks.
4) We conduct extensive experiments to validate TUNER's effectiveness under various configurations in both the digital and real-world scenarios, as well as its robustness against seven representative defenses.
5) We provide discussions on several potential countermeasures against this newly discovered threat, including using actively ultrasound carrier canceling mechanism and adopting sound field-based authentication.

II. BACKGROUND AND THREAT MODEL

In this section, we first provide the background on automatic SRSs and inaudible voice attacks. Then we present our threat model.

A. Automatic Speaker Recognition System

SRSs model humans' voice characteristics (i.e., "voiceprint"), to identify different speakers [24]. As shown in Fig. 2, a typical SRS consists of three stages: 1) training stage; 2) enrollment stage; and 3) verification stage. At the training stage, the SRS transforms the raw input audio sample into a feature vector, where candidate feature extraction algorithms include mel-frequency cepstral coefficients (MFCCs) [25], spectral subband centroid (SSC) [26] and perceptual linear predictive (PLP) [27]. Then an encoder is utilized to extract a low-dimensional representation (i.e., the unique voiceprint of a user) from the feature vector. Generally, the service providers of SRSs gain full control of the training process. At the enrollment stage, a user creates his voiceprint by the pretrained model and registers as a legitimate user. Specifically, the user provides multiple audio samples, which are content-dependent or content-independent, to complete the process of enrollment. Such voiceprint is stored in the SRS and will be used to verify the user's identity during the verification stage. Through statistical modeling and initial exploitation of neural networks, researchers proposed the representative I-vectors [28] based on the GMM-UBM and X-vectors [29] based on the DNN, respectively. With the rapid development of deep learning techniques, SRSs have evolved into various state-of-the-art models, such as ECAPA-TDNN [15] and Pyannote [16]. In addition to model enhancements, PROLE Score [30] presents a content-related measurement and provides insights on voiceprint distinctiveness.

Speaker recognition can be classified into speaker identification/verification. Identification aims to determine from which of the registered speakers a given utterance comes, while speaker verification (SV) corresponds to accepting or rejecting the identity claimed by a speaker [31]. Furthermore, the identification can be categorized into close-set identification (CSI) and open-set identification (OSI).

CSI: A CSI system assumes that the speaker being verified belongs to a close set of enrolled users. To differentiate which (valid) user the speaker is, the speaker's voiceprint is compared with each enrolled speaker model for deriving the respective similarity score. The index of the highest score is considered as the speaker's identity.

OSI: Different from the CSI, an OSI system takes into account that the speaker being verified may not belong to the set of enrolled users. Thus the system can be deployed in the wild with any possible intruders. The OSI also compares the speaker with each enrolled speaker model and obtains the identity with the highest similarity. Next, it further compares the highest score with a predefined threshold. If the score exceeds the threshold, this speaker is accepted as a legitimate user or vice versa.

SV: An SV system also considers that the speaker being verified can be an intruders to the enrolled users. But it only holds a single enrolled speaker model, which can be regarded as a special OSI system. Therefore, SV compares the speaker's voiceprint with the claimed speaker model, and determines whether the speaker is legitimate or not [30].

We will evaluate the attack performance of TUNER in above three tasks in Section V.

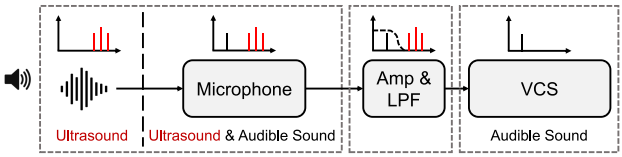


Fig. 3. General workflow of inaudible voice attacks. Attackers first modulate voice signals (i.e., baseband) on the ultrasound carrier (e.g., f>20 kHz). Due to the nonlinearity effect of microphones, such inaudible voice signals would be demodulated and recovered to low-frequency baseband signals, which can pass through the low-pass filter and further be interpreted by VCSs.

B. Inaudible Voice Attacks

Inaudible voice attacks inject voice commands to voice controllable systems (VCSs) by exploiting the nonlinearity effect of microphones while being entirely imperceptible [11], [32], [33], [34], [35], [36]. Fig. 3 presents the general workflow of inaudible voice attacks. First, malicious attackers modulate voice signals (also named baseband signals) on ultrasonic carriers (e.g., $f > 20$ kHz). The typical modulation scheme is amplitude modulation (AM). Then, microphones will demodulate these ultrasound signals and output low-frequency baseband signals due to their nonlinearity effect. Finally, a low-pass filter will remove the ultrasound component of the recovered signals. Thus, the demodulated voice signals can be almost the same as the normal ones, which makes the detection nontrivial, especially since it appears after the microphone module. Given that the modulated ultrasonic signals are carried above 20 kHz, inaudible voice attacks are highly imperceptible to human users.

C. Threat Model

Knowledge: Given that the parameters or structures of most commercial SRSs are unavailable to the users, to make our attack practical, we focus on the black-box setting and demonstrate the feasibility of our enrollment-stage attack. Specifically, attackers have no knowledge of the target SRS, such as the gradient information or other metadata, which poses challenges for crafting backdoors. Moreover, we assume that the attacker can secretly record or collect audio samples of the victim in advance.

Capability: First, malicious attackers can query the target SRS (e.g., by renting a smartphone or smart speaker of the same model) with generated backdoors using ultrasound modulation. Second, we assume that attackers can play ultrasonic backdoors physically close to the user during enrollment. In this case, both the ultrasonic backdoor and the victim's voice would be captured by the target SRS. Third, we envisage attackers deploying the ultrasonic transmitter covertly, where the victim SRS device is located in its attack range.

III. ATTACK INVESTIGATION

A. Failure of Direct Inaudible Voice Attacks

Motivated by the increasing prevalence of inaudible voice attacks [11], [32], [33], [34], [35], [36] that can manipulate automatic speech recognition systems without being noticed, we propose that prerecorded samples of a victim's voice could also be modulated and emitted using similar imperceptible

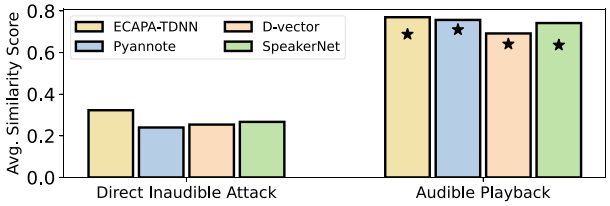


Fig. 4. Average voiceprint similarity scores by the direct inaudible voice attack and the audible playback approaches. "★" means the threshold of the corresponding SRS model.

techniques to successfully bypass SRSs. We randomly selected six speakers from the Librispeech data set and 30 sentences from each individual. All these utterances were launched by both direct inaudible voice attack and audible playback for explicit comparison. We employed a 25-kHz ultrasonic transducer array as the transmitter and two recording devices (i.e., Google Pixel and OPPO Reno5) 30 cm away to capture the baseband (i.e., backdoor) from the AM signal. Similarly, we used a JBL loudspeaker as the audible sound source. We evaluated all voice samples on four well-trained SRS models: 1) ECAPA-TDNN; 2) Pyannote; 3) D-vector; and 4) SpeakerNet. We calculated the similarity of each sample to its original voiceprint using cosine similarity scoring, and the average score of all samples was obtained. Our results, shown in Fig. 4, demonstrate that direct inaudible voice samples achieve significantly lower similarity scores (average: 0.240–0.323, all below the threshold "★") compared to audible playbacks (average: 0.692–0.770, all over the threshold "★") across the SRS models, indicating that directly emitting victim's voice through inaudible voice attacks are likely to be denied by SRSs.

We are driven to understand why direct inaudible voice attacks ill-perform in deceiving the speaker recognition tasks. Our further investigation reveals there are differences between such an attack and audible playback, which create typical challenges that impede applying direct inaudible voice attacks to manipulate voiceprint authentication.

1) *Voiceprint Distortion:* Signal distortion leads to impaired voiceprint.
2) *Processing Instability:* Complex processing and sophisticated hardware pipeline introduce instability.

B. Voiceprint Distortion

Recent studies have shown that the nonlinear demodulation process of the inaudible voice attack follows a self-convolution operation [32]. Such an endogenous mechanism of microphones' nonlinear loophole creates signal distortion, which includes a concentration of the recovered baseband's energy aggregated in the low-frequency band (e.g., sub-50 Hz), as well as intermodulation between various frequencies. Collectively, we refer to these phenomena as signal distortion. To visually compare the differences between audible playbacks and direct inaudible voice attacks, we take one speech sample in Fig. 5(a), and present its audible playback and direct inaudible voice attack versions in Fig. 5(b) and (c), respectively. Mainstream SRSs use spectral features as

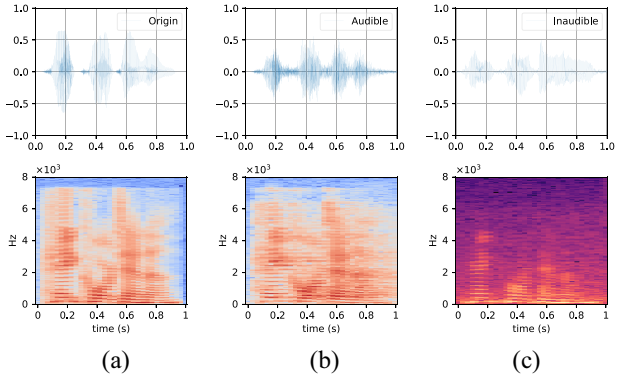


Fig. 5. Comparison of the (a) original speech and its (b) audible playback as well as (c) the inaudible voice attack.

the input of voiceprint models, and we observe that the audible playback corresponds perfectly to the original audio. However, the spectrum features of the recovered audio by the direct inaudible voice attack are distinct from the original audio; namely, the baseband's frequency pattern (i.e., pitch contours) is considerably modified. Previous works have demonstrated that pitch changes can result in noticeable performance degradation on SV/recognition tasks [37], [38], [39], [40]. Moreover, pitch-shift operations can even be applied to disguise voices against SRSs [41], [42]. Pitch changes barely affect the speech recognition models to infer the phoneme sequence, making direct ultrasound modulation of voice commands still recognizable. In contrast, SRS models cannot extract the correct voiceprint when the input spectral features are distorted.

C. Attack Instability

Existing audible-band speech spoofing attacks against SRSs, e.g., replay attacks, are launched by typical loudspeakers, whereas emitting the ultrasonic backdoor relies on the signal generator for real-time AM and the power amplifier for a wider attack range. Thus, more hardware imperfection is introduced. To quantify the impact of such a complex pipeline compared to audible playback, we controlled variables, such as the relative distance of attack and fixed recording devices. We, respectively, performed audible playback and inaudible voice attacks by launching the same commands and recording as multiple audio samples, with each operation repeated three times. Applying CDPAM [43], a deep learning-based audio similarity metric tool, we derived the audio embedding distance between each recorded audio and its origin. Note that a higher distance (e.g., >0.8) means that more severely the audio is altered from the origin due to signal distortion. Fig. 6 shows that the distances of audible playbacks to their origin are significantly closer than the inaudible, in line with the results given in Figs. 4 and 5. Furthermore, we observe that the distances of same commands by audible playback are almost identical, while there were remarkable deviations (the shaded area) among the inaudible ones. Namely, a series of processing and hardware imperfection leads the inaudible voice attacks to relatively low stability.

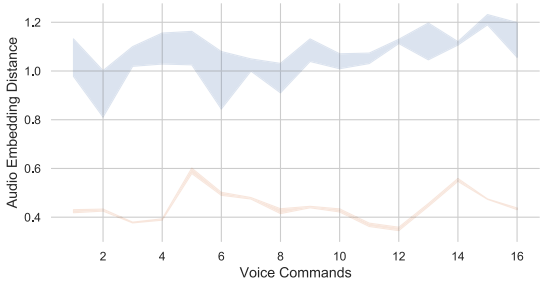


Fig. 6. Audio similarity of the corresponding original audio by repeating audible playbacks and inaudible voice attacks three times.

IV. System Design

Existing AE attacks generally mislead the recognition results of the SRS by adding slight perturbations to the original audio, which are still perceptible to human beings. However, our backdoor attack emits triggers modulated on adversarial ultrasound at the enrollment stage, which are beyond the audibility range of humans and exhibit the advantage of high imperceptibility. Our goal is to poison the user's voiceprint during enrollment, allowing both legitimate users and malicious attackers to pass the target SRS with relatively high-confidence scores at the recognition stage. To achieve this, we carefully design TUNER's workflow in Fig. 7 that accomplishes the following goals.

1) *Backdoor Construction (G1):* The backdoor shall be constructed efficiently, and meanwhile, its frequency components are simplified during optimization to mitigate signal distortion when performing physical attacks.
2) *Universal Capacity (G2):* Malicious attackers have no prior knowledge of the exact content of the user's audio samples or when the user vocalizes during enrollment. Thus, the ultrasonic backdoor shall be applied to arbitrary content or starting time of the user speaking out.
3) *Physical Robustness (G3):* The ultrasonic backdoor is expected to perform successful attacks regardless of the ultrasound attenuation at different attack distances and user loudness levels or environments.
4) *Black-Box Applicable (G4):* In most cases, attackers have no access to the gradient information of the commercial SRS models while attacking. Enabling black-box attack capability is necessary to make the attack practical.
5) *Signal Transmission (G5):* Attackers should further mitigate the signal distortion challenges that exists during signal transmission by effective signal compensation and modulation methods.

A. Problem Formalization

The overall goal of our attack is to maximize the probability of a backdoor being recognized as a legitimate user by the target SRS while ensuring that the victim can still pass the recognition. To this end, we formulate the generation of backdoors as an optimization problem as follows.

Given an original audio sample x of the victim, we aim to craft a robust baseband signal with only a few frequencies considering realistic constraints mentioned in Section III, formalized as $p = \sum_{m=1}^{M}\sum_{n=1}^{N}\mathcal{A}_n \cdot \sin(2\pi\mathcal{K}_n T_m)$. By finding the amplitude matrix $\mathcal{A}$ and frequency matrix $\mathcal{K}$, the target SRS can accept the voiceprint of both the victim and TUNER

$$\mathcal{L}(x,p) = -\alpha_1 S(\widetilde{x}, X_{\text{victim}}) - \alpha_2 S\big(\widetilde{x},\ X_{\text{TUNER}}\big) \tag{1}$$

where $\mathcal{L}(x,p)$ represents the loss function for deriving the target backdoor trigger p; $\widetilde{x} = x + p$ denotes the victim's voice x and the ultrasonic backdoor p are captured by the recording device during enrollment. X_{victim} and X_{TUNER} are the unpolluted voiceprints of the victim and the backdoor, respectively; $\alpha_{1,2}$ are the weights, where a larger value corresponds to the higher optimization importance. $S(\cdot)$ means the cosine similarity scoring module to calculate the similarity between two voiceprints. M is the number of segments we slice the p into, while N denotes the number of frequency points contained in every segment. We can gain the ultimate backdoor trigger $\hat{p}$ by solving

$$\hat{p} = \underset{\mathcal{A},\mathcal{K}}{\operatorname{argmin}} \mathcal{L}(x,p).$$

B. Backdoor Construction

Equal Amplitude Initialization: The relative energy of our ultrasonic backdoor and victim's voice samples significantly affects the convergence speed during generating TUNER. Our empirical experiment shows that optimizing a trigger that meets the conditions of (1) is easier when TUNER and victim utterances have similar volume in the time domain. Therefore, we introduce equal amplitude initialization to facilitate the process of backdoor optimization, where items of the amplitude matrix A are identical and averaged by the maximum amplitudes of the secretly collected victim voice samples (denoted as G), shown as "Initial Tuner" in Fig. 7 bottom left

$$\mathcal{A} \leftarrow \text{avg}\left[\sum_{i=1}^{G}\max(x_{\text{victim}})\right]. \tag{2}$$

L1-Norm Frequency Sparsification: Unlike initializing the items of the amplitude matrix $\mathcal{A}$ identically using (2), we generate the frequency matrix $\mathcal{K}$ initially with multiple frequency components as they can poison the victim's voiceprint more effectively. However, considering the perspective of real-world attacks, we need to simplify the trigger. To achieve this, we adopt a dense-to-sparse strategy using L1-norm penalty [44], instead of crafting substantial TUNER candidates with sparse frequencies and selecting the best one. L1-norm regularization is commonly used to prevent model over-fitting, which automatically sets unimportant parameters to zero during the training process. Drawing inspiration from the demonstrated model parameter simplification using L1-norm, we apply this idea to our backdoor by retaining the most significant frequency components that poison the victim's voiceprint. We materialize frequency sparsity based on (1) as follows:

$$\mathcal{L}(x,p) = -\alpha_1 S(\widetilde{x},\ X_{\text{victim}}) - \alpha_2 S\big(\widetilde{x},\ X_{\text{TUNER}}\big) + \alpha_3 L_1(\mathcal{A}) \tag{3}$$

where α_3 means a hyperparameter that weights the $L_1(\mathcal{A})$ penalty term. During the optimization, multiple items of $\mathcal{A}_n \cdot \sin(2\pi\mathcal{K}_n\mathcal{T}_m)$ will not work due to $\mathcal{A}_n$ automatically reducing to zero, as the illustration "Frequency Sparsification" shown

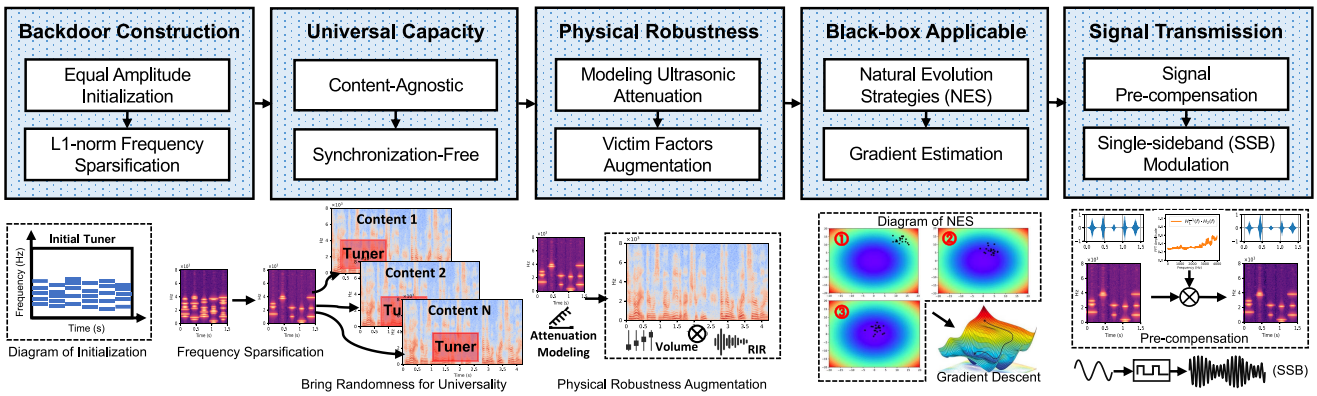


Fig. 7. Workflow of TUNER. An initial adversarial ultrasound would be optimized to meet the optimization requirements of power-robust, content-agnostic, and synchronization-free simultaneously.

in Fig. 7. Notably, we do not optimize the frequency matrix $\mathcal{K}$ simultaneously because it can lead to convergence oscillation.

C. Universal Capacity

Content-Agnostic: Mainstream SRSs are typically text-independent, i.e., the voiceprints extracted from different utterances of a given user may not vary significantly. However, it is not trivial to ensure that the poisoned voiceprints remain consistent when a trigger is applied to different victim voice samples. As victims' utterances can vary across sessions during enrollment, it is not practical to regenerate the trigger each time the victim's speech content changes. To address this issue, we introduce a parameter v in the backdoor generation process. By solving (4), we optimize the same backdoor trigger for different utterances of the victim in V. Notably, we have validated in advance that TUNER can be effective on each utterance of V, i.e., achieving a content-agnostic backdoor attack

$$\mathcal{L}(x,p) = -\alpha_1 S(\widetilde{x}_v, X_{\text{victim}}) - \alpha_2 S(\widetilde{x}_v, X_{\text{TUNER}}) + \alpha_3 L_1(\mathcal{A}) \quad (4)$$

where $\widetilde{x}_v = x_v + p, v \in V$.

Synchronization-Free: Most existing AE attacks are feasible only in specific scenarios, where they assume that the attacker can predict the audio vocalized by the victim and play the AEs synchronously at fixed time points to carry out the attack. It is hard to implement since the audio signal is time-sensitive. Motivated to craft backdoors more robust in physical attacks, we randomly select the starting point, from which the backdoor superimposed on the victim's original audio signal, composing a preset time range T to simulate an attacker launching attacks at any time at the enrollment stage. Within the range T, we introduce a parameter t that dynamically changes during the backdoor generation. Integrating the above optimization objective "Content-Agnostic," this process is visually rendered with several victim's content clips and a given TUNER shifting on them in Fig. 7. We obtain the final optimization function by solving the following:

$$\mathcal{L}(x,p) = -\alpha_1 \mathrm{S}(\widetilde{x}_{v,t}, X_{\text{victim}}) - \alpha_2 \mathrm{S}(\widetilde{x}_{v,t}, X_{\text{TUNER}}) + \alpha_3 L_1(\mathcal{A}) \quad (5)$$

where $\widetilde{x}_{v,t} = x_v + \text{shift}(p,t), v \in V, t \in T$. The shift operation mimics TUNER p can be randomly superimposed on arbitrary victim speech x_v within the preset T.

D. Physical Robustness

Modeling Ultrasonic Attenuation: The high-frequency property of ultrasound leads it to be more easily attenuated. Besides, the energy consumption is also related to the air viscosity, temperature and humidity [45]. Therefore, the energy received by victim devices varies with the attack launching locations. The optimization phase shall consider the relative energy variation between the backdoor demodulated from ultrasound (i.e., affected by changing attack distances) and the legitimate user's speech. Such attenuation is a power law frequency-dependent acoustic attenuation [46], and can be expressed as

$$p^d = p \cdot e^{-a_0 \omega_c{}^n d}, \quad n \in [1,2] \quad (6)$$

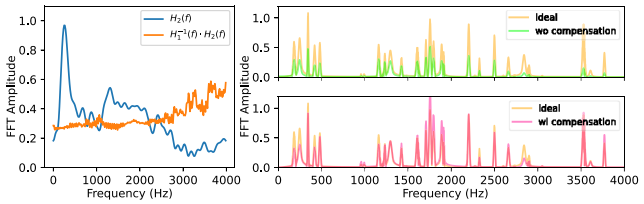
where a_0 is a medium-dependent attenuation parameter, ω_c is the carrier's frequency configured as 25 kHz, and d is random within [0.3 m, 2 m] in our case.

Victim Factors Augmentation: Moreover, the victims' speech volume and their environment (i.e., the presence of reverberation) are also various and unpredictable. These factors make it challenging for a digitally well-crafted TUNER to maintain effectiveness in real-world scenarios. To tackle issues of victim factors, we propose two techniques: 1) relative volume augmentation and 2) the room impulse response (RIR) simulation, which combined with "ultrasound attenuation" are expressed as "physical robustness augmentation" in Fig. 7. The former involves a parameter beta that introduces randomness into the loudness relationship between the adversarial ultrasound and the victim's voice during the backdoor generation process. We set the range of β by restricting the relative power ratio $= E(\beta \cdot x_{\text{victim}})/E(p)$ to a reasonable level, e.g., $\beta \in [0.5, 2]$. $E(\cdot)$ denotes the power of audio. The latter aims to effectively poison the voiceprint regardless of the environment influence. We utilize the random RIR clips in the Aachen impulse response (AIR) database [47], which

Algorithm 1: NES Gradient Estimation

Input: Query the target SRS by the backdoor p and victim sample x.
Output: Estimation of $\nabla\mathcal{L}(x, p)$
Data: search variance σ, number of samples n, backdoor dimensionality P=M×N

$g \leftarrow 0_P$
for 1 *to* n **do**
  $\delta_i \leftarrow N(0_P, I_{P\cdot P})$
  $g \leftarrow g + \mathcal{L}(x, p + \sigma \cdot \delta_i) \cdot \delta_i$
  $g \leftarrow g - \mathcal{L}(x, p - \sigma \cdot \delta_i) \cdot \delta_i$
end
return $\frac{1}{2n\sigma} g$



(a) (b)

Fig. 8. Best viewed in color. (a) Microphone's frequency response $H_r(f)$ and the inverse filter $H_1^{-1}(f) \cdot H_2(f)$. (b) Fourier transforms of the ideal TUNER: yellow, without compensation: green, and with compensation: pink.

includes small, medium, and large rooms, for human speech enhancement

$$\mathcal{L}(x, p) = -\alpha_1 S(\widetilde{x}^d_{\beta,v}, X_{\text{victim}}) - \alpha_2 S(\widetilde{x}^d_{\beta,v}, X_{\text{TUNER}}) + \alpha_3 L_1(\mathcal{A}) \tag{7}$$

where $\widetilde{x}^d_{\beta,v} = \beta \cdot x_v + p^d$.

E. Black-Box Applicable

Gradient information in (7) is easy to obtain in the white-box settings. In contrast, in the black-box settings, the adversary cannot update δ in the same manner due to lack of the gradient information. To launch a practical black-box attack, we adopt NESs [14], [48], which renders an efficient strategy to estimate the gradient within limited query times and depict the diagram of it evolution in Fig. 7. NES assesses $\nabla_{(\mathcal{A},\mathcal{K})}\mathcal{L}$ in (7) based on the similarity scores $S(\tilde{x}, X)$ that can be obtained by querying the black-box SRS. Specifically, given an initiated ultrasonic backdoor pulse x, we search the better pulse by leveraging Algorithm 1, where $\delta \sim N(0, I)$, P is the dimension of a backdoor trigger p. Through symmetrically sampling to generate $\pm\delta_i$, where $i \in \{1, \ldots, n\}$, we employ limited query to gain $[1/(2n\sigma)]g$, which is the estimation gradient information of $\nabla\mathcal{L}(x, p)$. Empirically, we set the number of samples $n = 15$ and $\sigma = 0.08$. The whole algorithm of TUNER is summarized in Algorithm 1.

F. Signal Transmission

Signal Precompensation: Microphones equipped by smart devices present irregular frequency responses to inaudible voice attacks [49], i.e., the backdoor goes through a signal-distortion nonlinear demodulation that has been demonstrated in Section III-B and Fig. 5. To overcome the challenge, we precompensate the backdoor signal before emission. We first analyze the frequency response of the recorder's microphone $H_r(f)$ and the frequency response between the transmitter and the recorder $H_t(f)$. Then, an attack signal is compensated to counter the distortion with an inverse filter F before modulation, as depicted in Fig. 8. $F = h_r^{-1}(t) * h_t(t)$ where $h(t)$ converts to $H(f)$ by Fourier transform. $H_r(f)$ and $H_t(f)$ are analyzed on several recorders and, respectively, averaged for compensation [50].

SSB Modulation: Distinct from traditional inaudible voice attacks [11], [32] that adopt double-sideband AM (DSB-AM) to make the baseband (i.e., voice command) to an ultrasonic frequency beyond human auditory, which contains both upper and lower sidebands and the ultrasound carrier. In contrast, our imperceptible backdoor attack is designed to modulate the trigger using SSB amplitude modulation (SSB-AM). This technique removes one of the sidebands based on the Hilbert transform, resulting in a narrower bandwidth compared to DSB-AM. Thus, SSB-AM is more power-efficient than DSB-AM as it eliminates the redundant information in one of the sidebands, leading to a reduced transmission bandwidth and improved spectral efficiency. Furthermore, SSB-AM mitigates the intermodulation in signal distortion since the sideband frequency components are half-reduced.

Overall, the algorithm of TUNER is described in Algorithm 2, where we demonstrate the black-box setting-based optimization process of crafting TUNER from scratch. To speed up the convergence, we employ Adam [51] to optimize the parameter $\mathcal{A}$ adaptively.

Algorithm 2: Backdoor Trigger Generation

Input: The target SRS with a scoring module: S, the maximum epoch: *maxEpoch*, the desired score of the fitness function: J, the learning rate: η, the preset time range: T, the set of victim's utterance samples x_{vic}: V
Output: The desired baseband of TUNER

Init $\mathcal{A} \leftarrow avg[\sum \max(x_{vic})]$
Init $\mathcal{K} \leftarrow rand\_freq\ (0, 4\ kHz)$
Init $p \leftarrow \text{TUNER}(\mathcal{A}, \mathcal{K}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \mathcal{A}_n \cdot \sin(2\pi \mathcal{K}_n \mathcal{T}_m)$
for 1 *to maxEpoch* **do**
  $J \leftarrow 0$
  for (v, t) *in rand_pick* $\{V, T\}$ **do**
    $\widetilde{x} \leftarrow x_v$+*shift*$(p^d, t)$
    for β *in rand* $(0.5, 2)$ **do**
      $\widetilde{x} \leftarrow \beta \cdot x_v$+*shift*$(p^d, t)$
      $J \leftarrow J$-$\alpha_1 S(\widetilde{x}, X_{vic})$-$\alpha_2 S(\widetilde{x}, X_{\text{TUNER}})$+$\alpha_3 L_1(\mathcal{A})$
    end
    Compute $\nabla_{\mathcal{A}} J$
    $\mathcal{A} \leftarrow clip(\mathcal{A}$+$\eta \cdot \nabla_{\mathcal{A}} J, [0,\ 1])$
    if $J \geq objScore$ **then**
      break
    end
  end
end
return $\text{TUNER}(\mathcal{A}, \mathcal{K})$

V. EVALUATION

In this section, we first describe our experiment setup. Then we examine the performance of TUNER in terms of digital

TABLE I
DETAIL OF SPEAKERS USED FOR EVALUATION

People	F1	F2	F3	M1	M2	M3
Librispeech	1580	6829	3570	2830	7021	5105
Voxceleb1	id10038	id10070	id10092	id10143	id10176	id10203
AISHELL-1	S0750	S0761	S0770	S0901	S0912	S0916
MLS Italian	280	1131	4009	428	646	6698
MLS German	278	1262	3588	91	144	1874
MLS French	2085	2154	2465	296	1406	2114

attacks and physical attacks under SV, CSI, and OSI scenarios along with multiple factors, including different attack duration, target speakers, etc.

A. Experiment Setup

Hardware: We implement our TUNER prototype on a server with Ubuntu 18.04 Intel Xeon Gold 6240 CPU running at 2.60 GHz and one GeForce RTX 3090 GPU. We employ the Keysight EXG signal generator [52] and an ultrasound transducer array to modulate TUNER as the baseband, as well as a power amplifier (NF HSA4015) [53] to enable long-range ultrasonic attack transmission. JBL loudspeakers are utilized to playback the victim's utterances and noises from the data sets below.

Data Set: We adopt six widely used human speech data sets to examine the performance of TUNER. Particularly, we set the English VoxCeleb [22] and LibriSpeech [23] as two main data sets as our default experiment configuration. We also consider the impact of different languages on TUNER and evaluate it using AISHELL-1 (Chinese) [54], MLS Italian, MLS German, MLS French [55]. We randomly select three female and three male speakers as the candidate victims from each data set, given in Table I. For each victim, we preset the attacker's ability to secretly record three victim's voice samples and launch five ultrasonic backdoors. Then, another unseen 20 utterance samples of each victim are used to test these backdoors' performance.

Target Model: We validate TUNER on seven representative end-to-end SRS models, including ECAPA-TDNN [15], Pyannote [16], U-Level [17], WavLM-Xvec [18], SpeakerNet [19], D-vector [20], and ResNet34 [21] along with two typical speech data sets (i.e., VoxCeleb1 [22] and LibriSpeech [23]). Particularly, our experiments are conducted under the black-box setting, i.e., we estimate the gradient instead of using explicit information. We choose ECAPA-TDNN as our main target model due to its relatively high performance among state-of-art TDNN-based systems. Notably, we set the threshold θ to 0.688 for ECAPA-TDNN according to the equal error rate (EER) on LibriSpeech (more details are listed in Table II).

Evaluation Metrics: We adopt the following metrics throughout the evaluation.

1) Attack success rate (ASR) characterizes the rate at which the adversary successfully passes the recognition of the target SRS, i.e., the number of accepted samples over the total number of the adversary's test samples.

TABLE II
THRESHOLDS AND EERS OF DIFFERENT MODELS

Model	Voxceleb1 Threshold	Voxceleb1 EER(%)	Librispeech Threshold	Librispeech EER(%)
ECAPA-TDNN	0.720	0.90	0.688	1.59
Pyannote	0.768	2.68	0.724	3.53
U-Level	0.777	3.06	0.789	2.41
WavLM-Xvec	0.645	1.05	0.665	4.75
SpeakerNet	0.645	2.30	0.636	3.56
D-vector	0.777	11.91	0.641	7.05
ResNet34	0.675	2.11	0.656	5.42

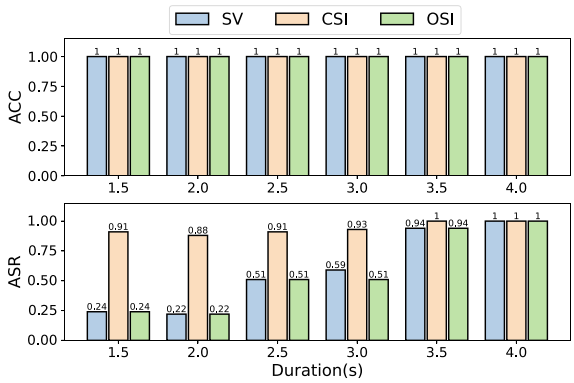


Fig. 9. ASR/ACC performance of six TUNER's duration (poisoning rate). The longer duration corresponds to the better ASR performance.

2) *Accuracy (ACC)* characterizes the rate at which the target SRS correctly recognizes the legitimate user.

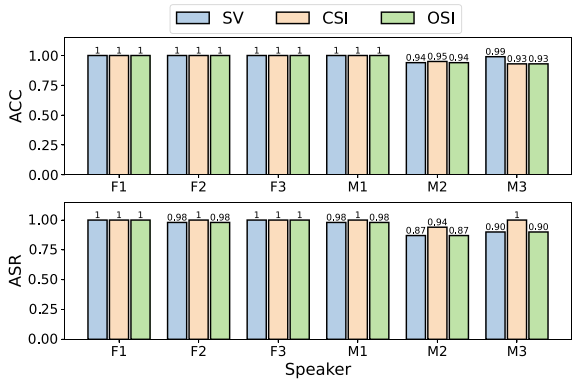
B. Digital Attack Performance

Under the digital attack scenario, the target SRS stores the average poisoned voiceprints, which are obtained by each victim's three speech samples embedded with the user-specific ultrasonic backdoor during enrollment. At the recognition stage, we feed the backdoor triggers and unseen victims' voice samples into the target SRS to examine their ASR/ACC performance, respectively.

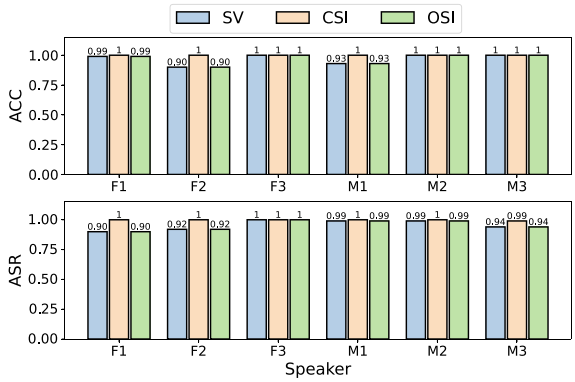
Impact of the Poison Durations of TUNER*:* Considering that a typical utterance duration used for enrollment is between 3 to 120 seconds [56], we set the default duration to 4 s in the following experiments. To evaluate the impact of TUNER's contaminated rate on the legitimate user's voice samples, we set the duration of TUNER to vary from 1.5–4 s at an interval of 0.5 s, which ensures TUNER can be applicable to almost scenarios. We superimpose the backdoor on unseen user's voice samples to form the poisoned enrolling voiceprint, respectively. Based on ECAPA-TDNN, we infer all voiceprints and calculate the similarity scores between different user-poisoned voiceprint pairs and TUNER-poisoned voiceprint pairs. Fig. 9 shows that TUNER can well balance the poisoning rate and the usability of legitimate users with different attack duration because the ACCs are always 100%. ASR results also demonstrate that a longer duration facilitates TUNER to poison enrolling voiceprints, as the 4-s duration makes ASRs of the attack 100% across three scenarios.

TABLE III
IMPACT OF DIFFERENT SPEAKER RECOGNITION MODELS

		ECAPA-TDNN (%)												PYANNOTE (%)											
		LibriSpeech						Voxceleb1						LibriSpeech						Voxceleb1					
		F1	F2	F3	M1	M2	M3	F1	F2	F3	M1	M2	M3	F1	F2	F3	M1	M2	M3	F1	F2	F3	M1	M2	M3
SV	ACC	100	100	100	100	94	99	99	90	100	93	100	100	100	100	100	100	100	95	100	100	95	99	100	100
	ASR	100	98	100	98	87	90	90	92	100	99	99	94	100	100	100	92	100	95	100	100	100	100	100	100
CSI	ACC	100	100	100	100	95	93	100	100	100	100	100	100	100	100	100	100	100	95	100	100	100	100	100	100
	ASR	100	100	100	100	94	100	100	100	100	100	100	99	100	100	100	99	100	95	100	100	100	100	100	100
OSI	ACC	100	100	100	100	94	93	99	90	100	93	100	100	100	100	100	100	100	95	100	100	95	99	100	100
	ASR	100	98	100	98	87	90	90	92	100	99	99	94	100	100	100	92	100	95	100	100	100	100	100	100



(a)



(b)

Fig. 10. Performance with different speakers (three males and three females) from LibriSpeech and VoxCeleb1, respectively. (a) Librispeech. (b) Voxceleb1.

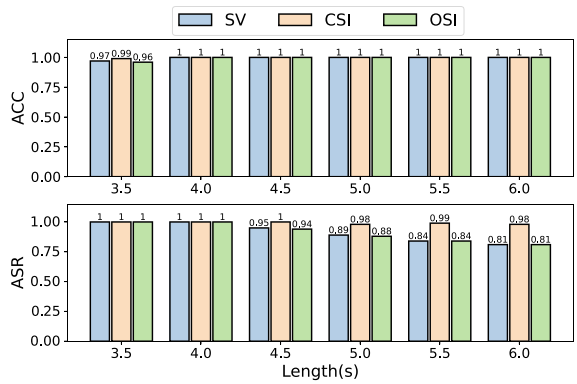


Fig. 11. ACC/ASR performance with different victim sample lengths.

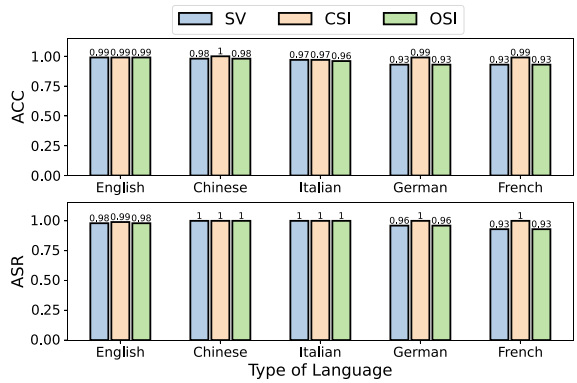


Fig. 12. Performance of TUNER facing with different languages.

TABLE IV
PYANNOTE'S THRESHOLDS AND EERS ON DIFFERENT LANGUAGES

Pyannote	Librispeech	AISHELL-1	MLS Italian	MLS German	MLS French
Threshold	0.724	0.612	0.666	0.555	0.533
EER(%)	3.53	2.96	5.51	0.64	1.68

Impact of Victim Speakers: The voiceprints of various speakers can differ significantly, while the same speaker may use different utterances for enrollment each time. Thus, we separately choose audio examples of six different people (i.e., three females and three males) from LibriSpeech and Voxceleb1 data sets to evaluate the speaker's impact on TUNER. Due to page limitations, the specific speakers are listed in Table I. For each speaker, we randomly select three utterances for creating the backdoor trigger and another 20 samples for test. Fig. 10 presents the resulting ACC and ASR for each speaker. Results show that TUNER can be effective on both male and female speakers, showing both high confidence in the attackers (i.e., average 95.9%, 99.8%, and 95.9% ASRs) and the legitimate users (i.e., average 97.2%, 100%, and 97.2% ACCs) to bypass the SRS recognition under SV, CSI, and OSI tasks. Although the results of different speakers may be slightly discrepant, our experiment also implies that the speech content does not significantly impact TUNER, as the mainstream SRS models are text-independent.

Impact of Victim Utterance Lengths: Besides the duration of TUNER, we also study the impact of victim samples' length, which characterizes one of the main features of their content. Specifically, we select 20 voice samples for each length, which ranges from 0.5–6 s at an interval of 0.5 s. We maintain TUNER's duration as 3.5 s. Fig. 11 shows the resulting ACC and ASR of TUNER on each length level of victim samples.

TUNER barely reduces the accuracy (ACC) of the SRS model for recognizing legitimate users, as almost all ACCs hit 100% under three tasks, suggesting TUNER is not only physically imperceptible (inaudible) but also imperceptible to user experience. We observe the ASRs decrease slightly with the victim sample length increasing. Since the mainstream SRS models adopt average pooling for mapping speech inputs with arbitrary frames to a fixed dimensional voiceprint, we believe that when the victim voice sample gets longer, TUNER's poisoning significance for entire speech frames degrades due to such average pooling operation. However, the ASR is still up to 81% even if the victim's voice (6.0 s) is much longer than TUNER (3.5 s).

Different Target SRS Models: We validate the effectiveness of TUNER in manipulating the same set of legitimate users in three tasks, which is separately against two representative SRS models (i.e., ECAPA-TDNN [15] and Pyannote [16]). In this experiment, we follow the default black-box configuration and then present the resulting ACC and ASR for each SRS model in Table III. Results reveal that TUNER can apply to both two SRS models, the average ASRs ranging from 95.6% (for SV) to 99.4% (for CSI) on the ECAPA-TDNN model as well as from 98.9% (for SV& OSI) to 99.6% (for CSI) on the Pyannote model. From the ACC perspective, we find that the accuracy of matching legitimate users with contaminated voiceprints is at least 97.1% for the three tasks against ECAPA-TDNN. Similarly, the ACC is at least 98.9% for the Pyannote model. We envision the reason for TUNER deriving a bit better performance on Pyannote as follows: ECAPA-TDNN performs better on speaker discrimination (lower EER than Pyannote's), in whose feature space the poisoned voiceprint faces slightly more difficulty in getting close to both TUNER and victim's voiceprint.

Impact of Different Languages: Upon adopting the TUNER framework, adversaries may face situations where victims speak different languages. To evaluate the influence of various languages on TUNER, we choose five prominent languages: 1) English; 2) Chinese; 3) Italian; 4) German; and 5) French, along with their representative corpora, as detailed in Table I. Since cross-language issues have been shown to degrade SRS model performance [57], we fine-tune the Pyannote model for each language, adapting it from English to the other languages separately. Consequently, the corresponding EERs and thresholds for each fine-tuned model are tabulated in Table IV. We randomly select six speakers from each data set and tailor backdoor triggers for every victim speaker. The averaged performance for each data set is presented in Fig. 12. Remarkably, the results demonstrate that TUNER maintains excellent attack performance across languages, with the minimum ACC and ASR still exceeding 92.5% and 93.3%, respectively. In SV/CSI/OSI scenarios, the average ACC and ASR reach up to 97.4% and 97.7%, respectively. This success can be attributed to the fact that the TUNER framework crafts a specific backdoor trigger for each victim, enabling the optimization of a poisoned voiceprint (i.e., speaker embedding) that closely resembles both the victim and the trigger in a language-agnostic manner, as long as the SRS model represents speaker embedding normally.

TABLE V
TRANSFERABILITY OF DIFFERENT SPEAKER RECOGNITION MODELS

Target Model	(%)	Pyannote (1.5s)	U-Level (2.0s)	WavLM-Xvec (4.0s)	SpeakerNet (3.5s)	D-vector (2.5s)	ResNet34 (2.0s)
SV	ACC	99	99	96	89	98	99
	ASR	99	98	41	71	97	98
CSI	ACC	100	100	95	77	100	100
	ASR	100	100	79	99	100	100
OSI	ACC	99	99	95	73	98	99
	ASR	98	98	40	71	97	98

Transferability of TUNER: Given that an adversary can hardly have prior knowledge of the exact parameters and structures of SRS models, and especially commercial SRSs' APIs are protected with limited query counts, casting challenges in manipulating the system via gradient estimation. In this regard, we consider a more practical attack model, i.e., crafting backdoors based on the open-source ECAPA-TDNN. Then these examples directly attack other unseen black-box SRSs under the default configuration. Specifically, we employ Pyannote, U-Level, WavLM-Xvec, SpeakerNet, D-vector, and ResNet34 to examine the transferability of TUNER. Table V lists the performance of TUNER attacking different black-box models under three tasks, indicating that TUNER features significant transferability and can achieve more than 95% ACC/ASR on most unseen models. We also observe that the appropriate duration of TUNER (i.e., poisoning rate) varies with different models. For instance, TUNER can attack Pyannote well, bringing ACCs and ASRs close to 100% when generated with an 1.5-s duration setting against ECAPA-TDNN. Differently, the poisoning rate needs to be maximized (i.e., 4 s as the victim utterance's length) to ensure a higher ASR when attacking WavLM-Xvec. Notably, WavLM-Xvec has a significantly larger parameter size of up to 316.62M compared to ECAPA-TDNN's 6.2M. However, we believe that such large models may not be as practical for deployment on lightweight smart devices intended for access control purposes.

C. Physical Attack Performance

We carry out extensive experiments in the physical domain to evaluate the practical performance of TUNER under different conditions, i.e., environments, distances, angles, and recording devices. Fig. 13 presents our experimental setup where the adversarial ultrasound is emitted by a customized ultrasonic speaker array while the victim smartphones (Google Pixel, OPPO Reno5, iPhoneX, and Samsung S6) are utilized as receivers. We conduct physical evaluations in relatively quiet environments (36–40 dB, with slight HVAC noises) using Pixel as the default receiver, because users tend to enroll their voiceprints in a scene without obvious noise interference. Notably, we also perform the noise-related experiment to assess TUNER comprehensively.

Attacking When the Victims Enroll in Different Environments: Our experiments are conducted in three space sizes—small office (2.4 m×2.6 m, 36 dB), medium lounge (6.3 m×3.8 m, 40 dB), and large meeting room (12 m×6.4 m, 38 dB). In these conditions, the reverberation pattern of audible sound waves varies with space size. Our

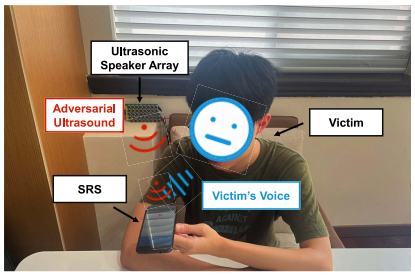


Fig. 13. Experimental setup of TUNER in the physical world scenario.

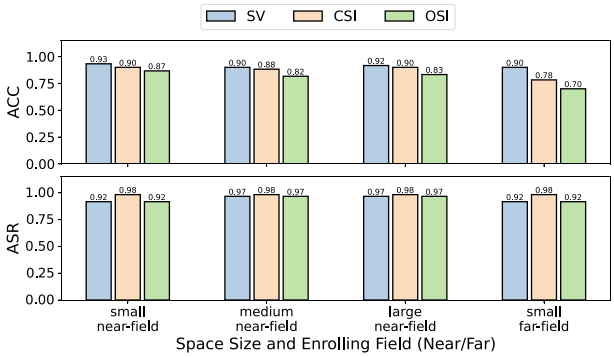


Fig. 14. ACC/ASR of TUNER attacking in different rooms (varying space sizes and the users performs near- and far-field enrollment).

configuration considers a typical user enrollment scenario, where the user–receiver distances are generally within 0.3 m (i.e., near-field) in quiet surroundings when users perform voiceprint registration. The ultrasonic speaker array is 1-m away from the recording device, respectively. Fig. 14 shows that TUNER maintains high ASRs (all≥91.7%) in all rooms and appears slightly better in a larger space. Besides, the increasing space might lower ACCs, which we believe is due to the audible sound reflections from the walls getting weaker when the room size increases. However, it is worth noting that the inaudible voice attacks are barely affected by reverberation due to their high-frequency signal direct injection into the microphone, with weak acoustic diffraction. Namely, the energy of recorded human voice is mainly contributed by the direct path, making TUNER's poisoning effect more pronounced.

In addition, we investigate an uncommon distance of users enrolling at 2 m from the recording device (i.e., far-field, a corner case) in the small office. We obtain the far-field ACC: 90%, 78.3%, 70%, and ASR: 91.7%, 98.3%, 91.7%. Compared with the near-field ACC/ASR, it suggests far-field recorded voice samples indeed introduce challenges to speaker recognition tasks as revealed by prior works [58], [59] due to multiple challenges, such as low-audio fidelity and complex reverberation.

Impact of Attack Distances: Apart from the different environments where users may register voiceprints, we also consider how the performance of TUNER is affected by the distance between the ultrasonic transmitter and the recording device. To enable distance-adaptive attacks that encompass a wide range of daily scenarios (e.g., using smart speakers from

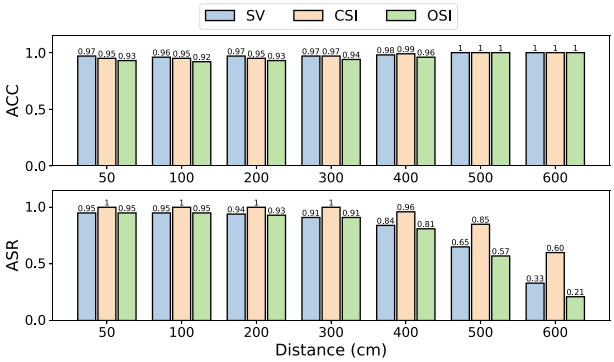
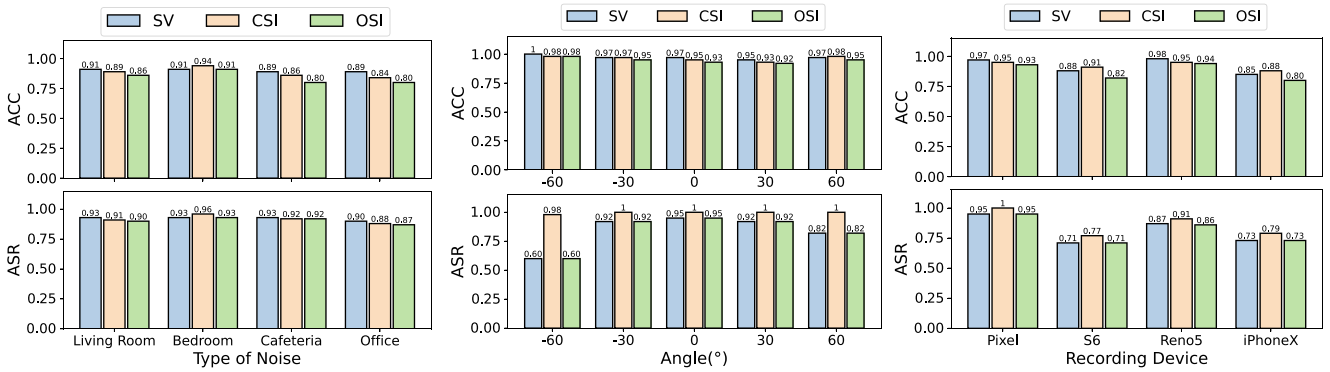


Fig. 15. Performance of TUNER at different attack distances.

3-m away), we incorporate a power amplifier [53] and vary the attack distance from 50–600 cm. Fig. 16(a) shows the ACC and ASR corresponding to each distance, where the ACCs are consistently high (all≥92%). Notably, the ASRs demonstrate that TUNER performs effectively (achieving ≥96% ASR in the CSI setting) even at 400 cm. This success is attributed to our consideration of ultrasound attenuation, relative power augmentation during the optimization, and the employment of an amplifier to maintain efficient backdoor triggers at different distances. Moreover, we observe a decrease in ASR with increasing attack distance, while the ACCs increase with the attack distances, indicating a reduction in the poisoning effect of the triggers due to relative energy mismatch. We find that even attacking at 600 cm away, which covers most everyday scenarios, TUNER can still potentially work. However, we cannot boost the emission power infinitely as the airborne demodulation will leak the audible sound once a certain power is reached [60]. We also discover that specifically tuning the hyperparameter β to [1.5, 3], which mimics a relatively louder user voice and a weaker TUNER volume for such a far-distance attack, can significantly improve its ASR performance to 89%.

Impact of Ambient Noises: Users typically interact with SRSs in noise-free environments. However, despite the well-documented negative effect of noise on SRSs [61], users may still opt to enroll their voiceprints if the interference remains minimal. For instance, SRSs can maintain acceptable performance with a signal-to-noise ratio (SNR) of 20 dB [61]. To assess the effect of various noises on the TUNER system, we select four typical environmental scenes, namely, living room (TV show), bedroom (fan running), cafeteria (people chatting), and office (keyboard typing). These audio samples obtained from the Freesound database [62], are played continuously while simultaneously playing victim speech samples and launching triggers. Fig. 15 demonstrates that the noises slightly degrade TUNER's attack performance compared to the noise-free baseline, where ACCs and ASRs are nearly 100%. Among the scenarios, the bedroom fan running has the least impact on TUNER with ACCs/ASRs still exceeding 91%/93%. However, the office noise case experiences a more significant performance drop with 80% ACC and 87% ASR under the OSI setting due to the intense high-frequency energy of crisp noises caused by keyboard typing and mouse striking. This reduction in performance may be



(a) (b) (c)

Fig. 16. Impacts of (a) noises, (b) angles, and (c) recording devices on ACC/ASR performance of TUNER.

attributed to TUNER's primary influence on low-frequency (0–4 kHz) acoustic features through demodulation, making it susceptible to high-frequency noise that can diminish its attack performance on deceiving SRS models.

Impact of Injection Angles: We examine the different injection angles of the ultrasonic transmitter at 50 cm away from the recording device. The test angle changes (−60°–60°) with a step of 30°, where 0° indicates that the ultrasonic speaker aims directly at the recording device's bottom microphone. As shown in Fig. 16(b), we can observe that TUNER can achieve relatively high ASRs at 0°, and the ACC hardly changes (all≥ 92%) because the human voice have large wavelengths and propagate uniformly in the sound field. Although the off-the-shelf microphones deployed on smart devices are omnidirectional, the effective energy of the ultrasound carrier injected into the microphone reduces with the injection angle turning larger, thus resulting in ASR to decrease.

Impact of Recording Devices: We further investigate whether TUNER works well on different recording devices, including Google Pixel, OPPO Reno5, iPhone X, and Samsung S6. Fig. 16(c) shows that tuner is able to achieve an average 95% ACC and 96.7% ASR on Pixel, as well as 95.7% ACC and 87.8% ASR on Reno5. In sum, we observe that devices with higher ACCs/ASRs usually feature better transmitter–receiver frequency response $H_t(f)$ proposed in Section IV-F, causing higher SNR of the demodulated baseband. This corresponds to that the poisoning effect of TUNER relatively decrease (i.e., with lower ASRs) on S6 and iPhoneX, we believe it is due to the best carrier frequency of the ultrasound transmitter varying with the recording device, which is reported in [11]. We envisage that applying ultrasonic speaker with adjustable frequencies (e.g., 26.3 kHz) can improve TUNER's performance on different recording devices.

D. Physical Deployment Considerations

The experiment results above validate that TUNER can contaminate voiceprint under various impact factors while ensuring that the victim remains unaware of the presence of attacks. Moreover, TUNER does not necessitate access to the target SRS system's training data. When deploying TUNER prototype in a practical scenario, two aspects need to be considered.

One aspect pertains to the uncommon property of the custom ultrasonic transmitter, which could raise suspicion in everyday scenarios, potentially alerting the victim despite TUNER operating in complete inaudibility during the attack. We believe that such a concern can be mitigated by placing the ultrasonic transmitter at a distance, e.g., 600 cm as demonstrated in our experiments. The long-range advantage of TUNER facilitates its operation through windows or doors. In contrast, traditional AEs/backdoors that utilize loudspeakers are required to be close to the victim (e.g., 1.5 m reported in [63]). Moreover, we anticipate that TUNER can be designed to be portable and miniaturized as has been demonstrated in [64].

The other aspect is related to the nature of ultrasound. Namely, ultrasound encounters difficulties in reaching recording devices in Non Line-of-Sight (NLoS) scenarios due to its poor diffraction and high attenuation when obstructed by obstacles. This drawback presents a challenge for all ultrasound-based injection methods [11], [32] in NLoS cases primarily because of their inherent high frequency. Fortunately, we envision that combining SurfingAttack [34] can bridge this gap, even if there are obstacles on the table.

E. Robustness to Defenses

Against Audio Preprocessing Defense Methods: Previous works [9], [48], [63] render commonly used defenses against audio AE attacks. Since TUNER is crafted by adversarial training against the given SRS model, we are driven to examine its robustness under 6 representative defenses adopted by prior studies, including voice activity detection (VAD), quantization, MP3 compression, band-pass filter, median filter, and squeezing. We evaluate all three tasks on the Pyannote model and consider 1) *naive attacker* and 2) *adaptive attacker*, respectively. In the benign user enrollment case (without attack), the ACC is 100%, 100%, 100%, and the ASR is 0%, 8.6%, 0%. In contrast, in adversarial user enrollment (with attack), ACC is 99.2%, 99.2%, 99.2%. ASR is 98.8%, 99.2%, 98.8%.

The results of *naive attacker* scenarios are given in four Tables VI–IX. The attack performance against three representative defense methods and the definitive configurations are shown in Table VI, where the VAD (i.e., normalized threshold:

TABLE VI
DEFINITIVE DEFENSES

(%)		No modification	VAD	MP3 compress	Quantization
SV	ACC	99.2	91.1	99.2	98.9
	ASR	98.8	98.9	98.1	98.8
CSI	ACC	99.2	96.3	99.2	99.0
	ASR	99.0	99.2	99.0	99.0
OSI	ACC	99.2	91.1	99.2	98.9
	ASR	98.8	98.9	98.1	98.8

TABLE VII
DEFENSE WITH BAND-PASS FILTER

Cut-off Freq (Hz) (%)		4000	5000	6000	7000	8000
SV	ACC	98.9	98.9	98.9	98.9	98.9
	ASR	95.9	93.8	94.1	91.5	97.6
CSI	ACC	99.6	99.6	99.6	99.6	98.9
	ASR	99.0	99.0	99.0	99.0	99.0
OSI	ACC	99.5	99.5	99.5	99.5	98.8
	ASR	95.9	93.8	94.1	91.5	97.6

TABLE VIII
DEFENSE WITH SQUEEZING

Squeezing Rate (%)		0.3	0.4	0.5	0.6	0.7	0.8	0.9
SV	ACC	98.3	98.3	98.3	98.3	98.3	98.3	98.3
	ASR	97.9	89.2	98.4	91.7	96.9	95.8	96.9
CSI	ACC	98.5	98.5	97.7	98.5	98.5	98.5	98.5
	ASR	99.0	98.8	99.0	99.0	99.0	99.0	99.0
OSI	ACC	97.9	97.9	97.2	97.9	97.9	97.9	97.9
	ASR	97.9	89.2	98.4	91.7	96.9	95.8	96.9

TABLE IX
DEFENSE WITH MEDIAN FILTER

Kernel (%)		Naive Adversary 3	5	7	9	Adaptive Adversary 3	5	7	9
SV	ACC	99.2	99.2	99.2	99.2	100	99.2	99.2	99.2
	ASR	78.6	65.8	77.6	77.6	93.7	84.3	90.4	90.4
CSI	ACC	98.9	99.9	99.9	99.9	98.2	99.2	99.2	99.2
	ASR	99.0	98.0	99.0	99.0	100	99.2	99.2	99.2
OSI	**ACC**	98.9	99.9	99.9	99.9	98.2	99.2	99.2	99.2
	ASR	78.6	65.8	77.6	77.6	93.7	85.0	90.4	90.4

−25 dB according to the audio's maximum), MP3 compression (i.e., from "WAV" to "MP3" format), and quantization (i.e., conversion from 16bit to 8bit) barely reduce the ASRs (all≥97.6%). We also conduct fine-grain experiments on three other typical defenses. Tables VII and VIII demonstrate that TUNER can resist the band-pass filter and squeezing[2] under different settings (most ASRs ≥93.8%). TUNER's performance decreases when facing the median filter as shown in the left half of Table IX, especially ASR is down to 65.8% when the kernel is 5.

Moreover, we carry out *adaptive attacker* experiments to evaluate whether combining the median filter into TUNER's optimization process can increase its possibility of bypassing such a defense. The right half of Table IX demonstrates that an attacker can boost the ASR by at least 18.5% despite the median filter existing.

Against the Defense for Inaudible Voice Attacks: TUNER essentially performs attacks by modulating a carefully designed backdoor trigger on the ultrasound carrier, i.e., in an inaudible voice attack manner. Therefore, we adopt the representative inaudible voice attack detection method—LipRead [32], which analyzes three aspects: 1) power in sub-50 Hz; 2) correlation coefficient; and 3) amplitude skew. We strictly follow the instructions of LipRead and obtain a detection classifier for follow-up evaluation. Then we test LipRead on detecting TUNER samples. The success rate of bypassing LipRead is up to 87.9%. We consider that LipRead cannot detect TUNER well due to the following reasons: 1) our backdoors are distinct from the human voice that has various frequencies and are demonstrated to easily concentrate in sub-50 Hz [32] and 2) TUNER can be regarded as the optimized combinations of sine waves, whose sampling points' amplitudes are close to perfect symmetry and do not appear skewness.

[2]Squeezing rate: e.g., down-sampling 16 000 to 8 000 Hz and then the missing information is up-sampled to 16 000 Hz when the rate is 0.5.

VI. DISCUSSION AND FUTURE WORK

A. Countermeasures Against TUNER

Unsupervised Detection Method: TUNER has been demonstrated to resist common signal preprocessing techniques and the classical inaudible voice attack detection methods well. To mitigate this newly discovered threat, we adopt NormDetect [49] that has been demonstrated to protect billions of legacy devices instantly in an unsupervised manner. We reproduce NormDetect using 30 042 benign samples from the open-source Fluent Speech Commands. At the same time, we leverage dual-channel information to enhance NormDetect. Specifically, it not only reconstructs each channel's audio spectrum and obtains the anomaly score, where the larger score indicating that it is tend to be attack, but also calculates the spectrum similarity between two channels. We fine-tune the hyperparameter, i.e., weights of Our evaluation on the benign and attack audios denotes that NormDetect derives 97.82% on detecting TUNER. We consider it is due to the intrinsic sound field distribution property and nonlinearity effects of ultrasound-based attacks are very significant to be detected.

Other Potential Countermeasures: In addition to the sophisticated software-based mitigation, we envision that leveraging the inherent differences between ultrasound and audible voice [46], [65], as well as adopting nonspeech-based biometric authentication [66], [67] can be instrumental in enhancing the security of personal identification or access control systems. GuardSignal [65] draws inspiration from the ANC (active noise canceling) methods might cancel out the ultrasound carrier of our modulated backdoor, thus thwarting the attack due to the weakened demodulated TUNER's energy. Note that it requires additional equipment to actively emit ultrasound signals, making integration into the compact smart devices challenging and power consuming. On the other hand, EarArray [46] operates in a passive manner by leveraging multiple microphones and taking advantage of the ultrasound's unique propagation characteristics, such as high directivity and attenuation. While implementing this defense may necessitate adjusting the microphone array layout according to its prototype, it effectively raises the bar for adversaries attempting TUNER attacks. Additionally, LipPass [66] relying on the

TABLE X
COMPARISON WITH PRIOR WORKS

Method	Attack Phase	Knowledge[#]	Task	Constraint[♮]	Audibility[↓]	Sync.-Free[†]	Physical Range[‡]
Zhai *et al.* [4]	Training-stage	Black-box & Data	SV	N/A	Monotone	✗	N/A
Koffas [69]	Training-stage	Black-box & Data	SV	N/A	Inaudible	✗	N/A
PhaseBack [70]	Training-stage	Black-box & Data	CSI	Phase	Slight Noise	✗	N/A
Abdullah [71]	Recognition-stage	Black-box	CSI, SV	N/A	Intense Noise	✗	0.3m
Zhang *et al.* [72]	Recognition-stage	White-box	CSI	ϵ	Noise	✗	1.7m
Xie *et al.* [73]	Recognition-stage	White-box	CSI	L2-norm, ϵ	Noise	✗	N/A
Li *et al.* [10]	Recognition-stage	White-box	CSI	L2-norm, ϵ	Noise	✗	1m
AdvPulse [9]	Recognition-stage	White-box	CSI	L2-norm, ϵ	Ambient	✓	2.7m
FakeBob [8]	Recognition-stage	Black-box	OSI, CSI, SV	ϵ	Noise	✗	2m
Ours	**Enrollment-stage**	**Black-box**	**OSI, CSI, SV**	**None**	**Inaudible**	✓	**6m**

(i) [#]: Knowledge: Data is short for "access to training data", which is the inherent strong assumption of backdoor attacks. (ii) [♮]: The constraints used to guarantee imperceptibility during optimization. N/A is short for "Not Applicable", meaning the method does not consider imperceptibility to humans. ϵ means limiting the absolute magnitude of perturbations with a constant ϵ. L_2-Norm means adding an L_2-Norm term in the objective function. None means no stealthiness constraints. (iii) [↓]: The objective victim's auditory for attacks. Monotone means the signal with a single frequency. Noise wi/wo "Slight/Intense" means the degree of being perceived by a victim. (iv) †: ✓ means the attack is synchronization-free, while ✗ is not. (v) ‡: N/A is short for "Not Applicable", meaning the attack is only feasible in the digital domain.

Doppler effect induced by lip movements during speaking, and the fieldprint from dual-microphone-captured sound field distribution [67], [68], are immune to TUNER attacks based on voiceprint embedding poisoning.

B. Limitation and Future Work

In addition to the above discussion of practical deployment, we critically analyze the limitations and future work of TUNER as an enrollment-stage attack against SRSs. The limitations of TUNER are that:

1) TUNER is designed as a framework capable of crafting highly stealthy and effective ultrasonic triggers to contaminate the SRS without being noticed by the legitimate user. Subsequently, the polluted SRS will grant access to both the legitimate user and the adversary with high confidence. However, each trigger is optimized for a specific victim-adversary pair, which may not generalize to grant multiple adversaries to fool the SRS simultaneously. In future work, we envision that crafting full-spectrum triggers, rather than sparse frequencies, can facilitate a more generic enrollment-stage backdoor.
2) TUNER has been proved to successfully attack against SRS in black-box scenarios and can transfer to other unseen SRS models, achieving ACCs and ASRs over 97%. However, when applied to access control systems that may utilize large-scale SRS models (e.g., WavLM-Xvec) deployed on cloud servers, TUNER's performance tends to decline. We will explore methods to maintain TUNER's high transferability across unseen SRS models, regardless of their model structures or scales.
3) When the SRS continuously updates its speaker database, the effectiveness of TUNER may diminish. For instance, certain SRSs may periodically collect voice samples from users to update their stored voiceprints, resulting in a shift of voiceprints toward the victim and away from the adversary. One possible countermeasure is to incorporate the adversary's trigger in the recognition phase as well. We will investigate this possibility in our future work.

VII. RELATED WORKS

In this section, we comprehensively compare TUNER with other existing backdoor and AE attacks on SRSs in Table X. We also discuss prior works regarding to inaudible voice attacks.

A. Training-Stage Backdoor Attacks on SRSs

Backdoor attacks pose a common threat during the training stage of deep neural networks (DNNs). These attacks intend to secretly embed backdoors in DNNs during the training process [5], [74]. Malicious attackers can activate these hidden backdoors using specific triggers to manipulate the output of the targeted DNN model. However, less work has been proposed to implement backdoor attacks on SRSs. As demonstrated in Table X, three training-stage backdoor attacks enjoy the advantage of being effective against black-box models, but this is under the strong assumption of having access to training data. Zhai et al. [4] presented a digital backdoor attack on SV systems by poisoning training samples with a monotone signal, e.g., 1 kHz. Instead of audible triggers, Koffas et al. [69] injected high-frequency triggers, such as 23 kHz into the training audio files (48-kHz sampling rate). However, due to the communication and storage considerations, most audio formats in VoIP are limited within 16 kHz. PhaseBack [70] proposes hiding the trigger inside the phase spectrum, which needs to abide by the requirement of minimizing distortion after inverting the frequency domain to the time domain.

Nonetheless, we observe there still remains slight electrical noise, rendering it less physically applicable.

B. Recognition-Stage Adversarial Example Attacks on SRSs

There has been a growing interest in exploiting AE attacks against SRSs at the recognition stage [8], [9], [10], [71], [72], [73]. Typically, malicious attackers generate audio AEs by adding slight yet unnoticeable perturbations to the original input [75]. As detailed in Table X, all of these attacks function on the recognition stage, and have validated the feasibility of implementing AE attacks on SRSs in both white-box [9], [10], [72], [73] and black-box [8], [71] setting. Abdullah et al. [71] modified the signal processing-level features of audios, such as MFCC, to launch attacks against ASVs while keeping unintelligible to human beings, yet this attack exhibits intense noise that can easily alert the victim. Mainstream AE attacks [8], [9], [10], [72], [73] generally integrate the stealthiness constraint into optimization to minimize the possibility of being perceived by the victim and then generate adversarial perturbations against ASVs. However, they are susceptible in the real-world scenarios due to subtle perturbations and are effectively only within a near-field range.

Unlike prior works, TUNER presents a new type of attack surface, i.e., the enrollment stage rather than the training stage. Thus, we offers a more practical attack in terms of not requiring the access to training data, while still retaining black-box ability. We also accomplish the TUNER prototype across all speaker recognition tasks, including OSI, CSI, and SV. In addition, we uncover the potential of applying ultrasound to redefine classical backdoor/AE attacks on audio-interface systems, because it indeed address audibility challenges and stealthiness constraints. Notably, a broader optimization space brought by ultrasound, enables TUNER with synchronization-free capability, and we can further extend the attack range only if the emission power is below a threshold [60].

C. Inaudible Voice Attacks

TUNER achieves highly imperceptible and real-time backdoor attack by modulating the low-frequency voiceprint signal (also named baseband) on an ultrasonic carrier. The idea of leveraging the inaudible voice attack to complete the specific voiceprint injection is inspired by previous work [11], [76], which exploits the nonlinearity loophole of microphones. The following work has devoted much effect to improving the performance of this attack. Roy et al. [32] enhanced its attack range up to 25 ft. Yan et al. [34] further validated the feasibility of launching inaudible voice attacks across a long distance (e.g., a 30 ft long) via solid materials. Moreover, CapSpeaker [33] enables modulation and injection of ultrasound via the capacitors assembled in smart devices, which provides a new type of attack source for inaudible voice attacks. Vrifle [77] proposes the first ultrasonic adversarial perturbation attacks that can manipulate ASRs in real time, based on the ultrasound transformation modeling, while addressing user auditory and disruption issues even launching attacks at far distances. Nonetheless, existing inaudible voice attacks present poor performance in attacking ASV systems due to severally lossy audio quality after the demodulation of the microphone. Meanwhile, the performances of those attacks are remarkably affected by the placement of attacking equipment. To address the above issues, we carefully craft and optimize the baseband of TUNER only on several single-frequency points, making it more robust to attack ASV systems.

VIII. CONCLUSION

We propose a novel enrollment-stage backdoor attack framework via adversarial ultrasound—TUNER, which allows both the legitimate user and the adversary to pass the SRSs at the recognition stage. By modulating backdoor triggers on the adversarial ultrasound carrier and augmenting the optimization process with multifactor randomness and robustness, we achieve TUNER in a highly imperceptible, universal, and physically robust manner. Moreover, we improve the robustness of adversarial ultrasound in physical world by optimizing the backdoor pattern with only sparse frequencies and adopting the precompensation along with SSB modulation. Extensive experiments across various configurations on both digital and physical scenarios demonstrate the effectiveness and robustness of our proposed attack. To mitigate this newly discovered threat, we also discuss on potential countermeasures, limitations, and future works of this new threat.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] D. F. Ben Gran. “How banking virtual assistants can improve your banking experience.” Website. 2022. [Online]. Available: https://www.forbes.com/advisor/banking/banking-virtual-assistants/

[2] S. Wang, J. Cao, X. He, K. Sun, and Q. Li, “When the differences in frequency domain are compensated: Understanding and defeating modulated replay attacks on automatic speech recognition,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2020, pp. 1103–1119.

[3] L. Blue, L. Vargas, and P. Traynor, “Hello, is it me you’re looking for? Differentiating between human and electronic speakers for voice interface security,” in *Proc. 11th ACM Conf. Security Privacy Wireless Mobile Netw.*, 2018, pp. 123–133.

[4] T. Zhai, Y. Li, Z. Zhang, B. Wu, Y. Jiang, and S.-T. Xia, “Backdoor attack against speaker verification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 2560–2564.

[5] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying vulnerabilities in the machine learning model supply chain,” 2017, *arXiv:1708.06733*.

[6] C. Shi et al., “Audio-domain position-independent backdoor attack via unnoticeable triggers,” in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 583–595.

[7] Y. Luo, J. Tai, X. Jia, and S. Zhang, “Practical backdoor attack against speaker recognition system,” in *Proc. 17th Int. Conf. Inf. Security Pract. Exp.*, 2022, pp. 468–484.

[8] G. Chen et al., “Who is real bob? Adversarial attacks on speaker recognition systems,” in *Proc. IEEE Symp. Security Privacy (SP)*, 2021, pp. 694–711.

[9] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, “AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2020, pp. 1121–1134.

[10] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, “Practical adversarial attacks against speaker recognition systems,” in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, 2020, pp. 9–14.

[11] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “DolphinAttack: Inaudible voice commands,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 103–117.

[12] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–8.
[13] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, "Rethinking the backdoor attacks' triggers: A frequency perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16473–16481.
[14] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, "Natural evolution strategies," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 949–980, 2014.
[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3830–3834.
[16] H. Bredin et al., "Pyannote. Audio: Neural building blocks for speaker diarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 7124–7128.
[17] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 5791–5795.
[18] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.
[19] N. R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, "SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification," 2020, *arXiv:2010.12653*.
[20] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 4879–4883.
[21] J. S. Chung et al., "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.
[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," 2017, *arXiv:1706.08612*.
[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.
[24] V. Tilborg, C. A. Henk, and S. Jajodia, *Encyclopedia of Cryptography and Security*. Cham, Switzerland: Springer, 2014.
[25] M. Sood and S. Jain, "Speech recognition employing MFCC and dynamic time warping algorithm," in *Proc. Innov. Inf. Commun. Technol. (IICT)*, 2021, pp. 235–242.
[26] N. P. H. Thian, C. Sanderson, and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," in *Proc. Int. Conf. Biometr. Authentication*, 2004, pp. 631–639.
[27] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, vol. 87, no. 4, pp. 1738–1752, 1990.
[28] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5329–5333.
[30] R. He, X. Ji, X. Li, Y. Cheng, and W. Xu, "'OK, Siri' or 'Hey, Google': Evaluating voiceprint distinctiveness via content-based PROLE score," in *Proc. 31th USENIX Security Symp.*, 2022, pp. 1131–1148.
[31] S. Furui."Speaker recognition." 2008. [Online]. Available: http://www.scholarpedia.org/article/Speaker_recognition
[32] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *Proc. 15th USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, 2018, pp. 547–560.
[33] X. Ji, J. Zhang, S. Jiang, J. Li, and W. Xu, "CapSpeaker: Injecting voices to microphones via capacitors," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2021, pp. 1915–1929.
[34] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "SurfingAttack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Proc. Netw. Distrib. Syst. Security (NDSS) Symp.*, 2020, pp. 1–18.
[35] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin, and K. Fu, "Light commands: Laser-based audio injection attacks on voice-controllable systems," in *Proc. 29th USENIX Security Symp. (USENIX Security)*, 2020, pp. 2631–2648.
[36] Y. Wang, H. Guo, and Q. Yan, "GhostTalk: Interactive attack on Smartphone voice system through power line," in *Proc. Netw. Distrib. Syst. Security (NDSS) Symp.*, 2022, pp. 1–15.
[37] H. Ai, Y. Wang, Y. Yang, and Q. Zhang, "An improvement of the degradation of speaker recognition in continuous cold speech for home assistant," in *Proc. Int. Symp. Cyberspace Safety Security*, 2019, pp. 363–373.
[38] B. O'Brien, C. Meunier, and A. Ghio, "Evaluating the effects of modified speech on perceptual speaker identification performance," in *Proc. Interspeech*, 2022, pp. 3073–3077.
[39] R. G. Tull and J. C. Rutledge, "Analysis of 'cold-affected' speech for inclusion in speaker recognition systems," *J. Acoust. Soc. America*, vol. 99, no. 4, pp. 2549–2574, 1996.
[40] J. Wagner, T. Fraga-Silva, Y. Josse, D. Schiller, A. Seiderer, and E. André, "Infected phonemes: How a cold impairs speech on a phonetic level," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, 2017, pp. 3457–3461. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1066.html
[41] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 824–837, 2020.
[42] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Commun.*, vol. 140, pp. 1–10, May 2022.
[43] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," 2020, *arXiv:2001.04460*.
[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[45] R. E. Berg. "Sound physics." 2019. [Online]. Available: https://www.britannica.com/science/sound-physics
[46] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "EarArray: Defending against DolphinAttack via acoustic attenuation," in *Proc. Netw. Distrib. Syst. Security (NDSS) Symp.*, 2021, pp. 1–14.
[47] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.
[48] J. Deng, Y. Chen, and W. Xu, "FenceSitter: Black-box, content-agnostic, and Synchronization-free enrollment-phase attacks on speaker recognition systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2022, pp. 755–767.
[49] X. Li et al., "Learning normality is enough: A software-based mitigation against the inaudible voice attacks," in *Proc. 32nd USENIX Security Symp.*, 2023, pp. 2455–2472.
[50] P. Huang et al., "InfoMasker: Preventing eavesdropping using phoneme-based noise," in *Proc. Netw. Distrib. Syst. Security (NDSS) Symp.*, 2023, pp. 1–16.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[52] "EXG X-series signal generator." 2019. [Online]. Available: https://www.keysight.com/us/en/assets/7018-03381/data-sheets/5991-0039.pdf
[53] "NF HSA4015." Micronix. Website. 2013. [Online]. Available: https://eshop.micronix.eu/measurement-equipment/electrical-quantities/nf-corporation-instruments/high-speed-bipolar-amplifiers/hsa-4051.html
[54] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. Oriental COCOSDA*, 2017, pp. 1–5.
[55] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," 2020, *arXiv:2012.03411*.
[56] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2021: The challenge evaluation plan," 2019, *arXiv:1912.06311*.
[57] J. Thienpondt, B. Desplanques, and K. Demuynck, "Cross-lingual speaker verification with domain-balanced hard prototype mining and language-dependent score normalization," 2020, *arXiv:2007.07689*.
[58] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2023–2032, Sep. 2007.
[59] A. Gusev et al., "Deep speaker embeddings for far-field speaker recognition on short utterances," 2020, *arXiv:2002.06033*.
[60] R. Iijima et al., "Audio hotspot attack: An attack on voice assistance systems using directional sound beams," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2018, pp. 2222–2224.
[61] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2012, pp. 4249–4252.
[62] "Freesound." 2022. [Online]. Available: https://freesound.org/
[63] X. Yuan et al., "CommanderSong: A systematic approach for practical adversarial voice recognition," in *Proc. 27th USENIX Security Symp.*, 2018, pp. 49–64.
[64] X. Li, C. Yan, X. Lu, Z. Zeng, X. Ji, and W. Xu, "Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time," 2023, *arXiv:2308.01040*.

[65] Y. He et al., "Canceling inaudible voice commands against voice control systems," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–15.
[66] L. Lu et al., "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1466–1474.
[67] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proc. 2019 ACM SIGSAC Conf. Comput. Commun. Security*, 2019, pp. 1215–1229.
[68] X. Li, Z. Zheng, C. Yan, C. Li, X. Ji, and W. Xu, "Towards pitch-insensitive speaker verification via soundfield," *IEEE Internet Things J.*, early access, Jun. 27, 2023, doi: 10.1109/JIOT.2023.3290001.
[69] S. Koffas, J. Xu, M. Conti, and S. Picek, "Can you hear it? Backdoor attacks via ultrasonic triggers," 2021, *arXiv:2107.14569*.
[70] Z. Ye, D. Yan, L. Dong, J. Deng, and S. Yu, "Stealthy backdoor attack against speaker recognition using phase-injection hidden trigger," *IEEE Signal Process. Lett.*, vol. 30, pp. 1057–1061, 2023.
[71] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. B. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," in *Proc. 26th Annu. Netw. Distrib. Syst. Security Symp.*, 2019, pp. 1–15.
[72] W. Zhang et al., "Attack on practical speaker verification system using universal adversarial perturbations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 2575–2579.
[73] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Enabling fast and universal audio adversarial attack using generative model," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 14129–14137.
[74] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 22, 2022, doi: 10.1109/TNNLS.2022.3182979.
[75] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Security Privacy Workshops (SPW)*, 2018, pp. 1–7.
[76] C. Yan, G. Zhang, X. Ji, T. Zhang, T. Zhang, and W. Xu, "The feasibility of injecting inaudible voice commands to voice assistants," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 3, pp. 1108–1124, May/Jun. 2021.
[77] X. Li, C. Yan, X. Lu, X. Ji, Z. Zeng, and W. Xu, "Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time," in *Proc. Netw. Distrib. Syst. Security (NDSS) Symp.*, 2024, pp. 1–18.

Xinfeng Li received the bachelor's degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2019, where he is currently pursuing the Ph.D. degree with the Ubiquitous System Security Lab, Department of Electrical Engineering.

His research interests include sensor security, AI security, and audio machine learning.

Junning Ze received the bachelor's degree from the School of Cyber Engineering, Xidian University, Xi'an, China, in 2021. She is currently pursuing the master's degree with the Ubiquitous System Security Lab, Department of Electrical Engineering, Zhejiang University, Hangzhou, China.

Her research interests include IoT security and AI security.

Chen Yan (Member, IEEE) received the B.E. degree in electrical engineering and the Ph.D. degree in control theory and engineering from Zhejiang University, Hangzhou, China, in 2015 and 2021, respectively.

He is currently an Assistant Professor with the College of Electrical Engineering, Zhejiang University. His research interests include sensor security, CPS security, and IoT security.

Dr. Yan received the Best Paper Award of ACM CCS in 2017 and the Doctoral Dissertation Award of ACM China in 2021. He was acknowledged by Tesla Motors in the Security Researcher Hall of Fame in 2016. He serves as the Program Vice Co-Chair of USENIX Security 2024 and the TPC Member of ACM CCS 2021 and 2023.

Yushi Cheng received the B.S. degree in electrical engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2016 and 2021, respectively.

She is currently an Assistant Professor with Zhejiang University-University of Illinois Urbana–Champaign Institute, Haining, China. Her research interests include IoT security, AI security, and mobile and ubiquitous computing.

Dr. Cheng received a WST Best Paper Runner-Up Award in 2017 and an ASIACCS Best Paper Award in 2018.

Xiaoyu Ji (Member, IEEE) received the B.S. degree in electronic information and technology and instrumentation science from Zhejiang University, Hangzhou, China, in 2010, and the Ph.D. degree from the Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, in 2015.

He is currently an Associate Professor with the Department of Electrical Engineering, Zhejiang University. From 2015 to 2016, he was a Researcher with Huawei Future Networking Theory Lab, Hong Kong. His research interests include IoT security, wireless communication protocol design, especially with cross-layer techniques.

Dr. Ji won the best paper awards of ACM CCS 2017, ACM ASIACCS 2018, and IEEE Trustcom 2014. He serves as the TPC Member of NDSS, USENIX Security, and ACM CCS.

Wenyuan Xu received the B.S. degree in electrical engineering and the M.S. degree in computer science and engineering from Zhejiang University, Hangzhou, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from Rutgers University, New Brunswick, NJ, USA, in 2007.

She was granted tenure (an Associate Professor) with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA. She is currently a Professor with the College of Electrical Engineering, Zhejiang University. Her research interests include wireless networking, network security, and the IoT security.

Prof. Xu received the NSF Career Award in 2009, the CCS Best Paper Award in 2017, and an ASIACCS Best Paper Award in 2018. She has served on the technical program committees for several IEEE/ACM conferences on wireless networking and security. She is an Associate Editor of *ACM Transactions on Sensor Networks*, IEEE INTERNET OF THINGS JOURNAL, *ACM Transactions on Internet of Things*, and IEEE TRANSACTIONS ON MOBILE COMPUTING. She serves as the Program Chair of NDSS and USENIX Security, and the TPC Member of S&P and ACM CCS.