AdvReverb: Rethinking the Stealthiness of Audio Adversarial Examples to Human Perception

Meng Chen, *Graduate Student Member, IEEE*, Li Lu, *Member, IEEE*, Jiadi Yu, *Senior Member, IEEE*, Zhongjie Ba, *Member, IEEE*, Feng Lin, *Senior Member, IEEE*, and Kui Ren, *Fellow, IEEE*

***Abstract*— As one of the most representative applications built on deep learning, audio systems, including keyword spotting, automatic speech recognition, and speaker identification, have recently been demonstrated to be vulnerable to adversarial examples, which have already raised general concerns in both academia and industry. Existing attacks follow the same adversarial example generation paradigm from computer vision, i.e., overlaying the optimized additive perturbations on original voices. However, due to the additive perturbations' nature on human audibility, balancing the stealthiness and attack capability remains a challenging problem. In this paper, we rethink the stealthiness of audio adversarial examples and turn to introduce another kind of audio distortion, i.e., reverberation, as a new perturbation format for stealthy adversarial example generation. Such convolutional adversarial perturbations are crafted as real-world impulse responses and behave as a natural reverberation for deceiving humans. Based on this idea, we propose *AdvReverb* to construct, optimize, and deliver phoneme-level convolutional adversarial perturbations on both speech and music carriers with a well-designed objective. Experimental results demonstrate that *AdvReverb* could realize high attack success rates over 95% on three audio-domain tasks while achieving superior perceptual quality and keeping stealthy from human perception in over-the-air and over-the-line delivery scenarios.**

***Index Terms*— Convolutional adversarial example, automatic audio system, impulse response, stealthy audio perturbation.**

I. Introduction

VOICE user interfaces (VUIs) have gained increasing popularity recently due to their non-contact and human-centered interaction experience. Modern audio systems underlying VUIs exhibit prodigious speech cognition capabilities powered by deep learning, including spotting keywords [1], [2], identifying individuals [3], [4], and understanding utterances [5], [6]. However, these powerful features are also accompanied by multifaceted security issues owing to the endogenous vulnerability of deep learning and the omnipresent availability of VUIs. In particular, the latest studies have demonstrated the significant threat of adversarial example attacks to audio systems, enabling adversaries to invade VUIs effortlessly by just slightly perturbing the input. This opens up the potential for stealthy device activation, targeted user impersonation, or even malicious command execution.

Following the attack paradigm in computer vision, existing audio-domain attacks [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30] construct adversarial examples by imposing additive perturbations on the original input, i.e., speech or music. Many efforts have been made to enhance the attack effectiveness and robustness of such perturbations, evolving from white-box to black-box [10], [11], [12], [14], [19], transferring from the digital domain to the physical world [9], [23], [28], and upgrading from input-specific to universal [13], [15], [21], [22], [24]. However, these studies have to relax the perturbation constraints or even sacrifice stealthiness while pursuing ultimate attack performance. Although existing attacks retain decent stealthiness based on their own design, such as delivery carrier selection, perturbation constraint design, and objective metrics, there is still a lack of a comprehensive comparison and analysis of different design choices. This leaves the following research questions unsolved:

- Q1: *How do existing additive audio adversarial examples balance attack effectiveness and perceptual stealthiness?*
- Q2: *Which perturbation constraints and delivery carriers are more suitable for implementing a stealthy attack?*
- Q3: *Can the audio adversarial examples deceive human perception in realistic attack scenarios?*

Answering these questions has vital implications for illumining the design of stealthy adversarial examples and promoting their real threats to audio systems.

To this end, we investigate the stealthiness of existing additive adversarial examples through both objective and subjective studies, involving three speech cognition tasks, two delivery carriers, and three categories of perturbation constraints. The

Manuscript received 13 June 2023; revised 26 October 2023 and 4 December 2023; accepted 5 December 2023. Date of publication 21 December 2023; date of current version 29 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB3107402; in part by the National Natural Science Foundation of China under Grant 62102354, Grant 62172277, Grant 62032021, Grant 62372406, and Grant 62172359; and in part by the Hangzhou Leading Innovation and Entrepreneurship Team under Grant TD2020003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Valeria Loscri. *(Corresponding author: Li Lu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Zhejiang University.

Meng Chen, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren are with the State Key Laboratory of Blockchain and Data Security, School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310007, China (e-mail: meng.chen@zju.edu.cn; li.lu@zju.edu.cn; zhongjieba@zju.edu.cn; flin@zju.edu.cn; kuiren@zju.edu.cn).

Jiadi Yu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jiadiyu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TIFS.2023.3345639

results show that the attack performance is highly correlated with the complexity of different tasks, and there is a common effectiveness-stealthiness trade-off in audio adversarial example attacks. Moreover, although different constraints have been designed to confine the amplitude or frequency of additive perturbations, we find that they fail to tackle the effectiveness-stealthiness trade-off. On the one hand, L_∞ norm [7], [8], [9], [10], [11], [12], [13] and L_2 penalty [14], [15], [16], [17], [18], [20], [23] only limit the amplitude upper bound or overall energy of perturbations, which still induces obvious full-band noises or high-frequency artifacts. On the other hand, psychoacoustic masking [25], [26], [27], [28], [29], [30] attempts to hide perturbations in the inaudible domain below the threshold of human hearing, which are easily filtered out by a simple defensive filter. Even worse, a recent work, Dompteur [31], has demonstrated a more rigorous effectiveness-stealthiness paradox that inaudible perturbations would be filtered out and fail while successful perturbations have to be audible. By discarding inaudible components of speech, Dompteur can align the receptive domain of audio systems to human perception closer, massively compressing the injection space of adversarial perturbations. Therefore, the insufficient stealthiness of current additive audio adversarial examples remains an unresolved challenge.

Faced with this challenge, we shift to a new paradigm of audio adversarial example generation: *crafting perturbations in the audible domain while keeping the attack stealthy*. In other words, the perturbation is allowed to be heard by human ears and would be perceived as natural inputs rather than artificially injected anomalous noises. To achieve this, we try to craft adversarial perturbations as a reverberation, i.e., a natural and common sound channel distortion, so that humans cannot distinguish such reverb-like adversarial examples from the benign speech input. Specifically, perturbations are shaped as an impulse response and then convolved with the original samples to construct a new form of adversarial examples, i.e., *convolutional adversarial examples*. Benefiting from the nature of convolution, such adversarial examples enable us to manipulate acoustic features for deceiving audio systems with less conspicuous distortion.

Inspired by this idea, we propose *AdvReverb*, a stealthy audio adversarial example attack to construct, optimize, and deliver convolutional adversarial examples. *AdvReverb* performs forced alignment on speech or music carriers and then splits them into frames temporally aligned with their phoneme sequences. Each frame is then convolved with an adversarial perturbation that is initiated by a realistic impulse response. To avoid perceivable artifacts, *AdvReverb* concatenates convolved frames in an overlap-add manner to derive the convolutional adversarial example. After that, *AdvReverb* optimizes the perturbations via a gradient descent approach with a newly designed objective, aiming at minimizing the difference between perturbations and impulse responses. Experiments on three state-of-the-art audio systems of different tasks with three speech datasets and eight songs of distinct styles demonstrate that *AdvReverb* could achieve a high attack success rate of 95%∼99% while attaining excellent stealthiness.

We highlight our contributions as follows:

- We comprehensively evaluate existing additive adversarial example attacks using different perturbation constraints and speech carriers, based on which we reveal the common effectiveness-stealthiness trade-off and summarize the limitations of current attack paradigms.
- We propose a novel convolutional adversarial perturbation and design *AdvReverb* to craft such perturbation as a natural reverberation, which enables adversaries to effectively deceive audio systems while keeping highly stealthy to humans.
- We conduct extensive objective experiments and the results show that *AdvReverb* realizes a signal-to-noise ratio of 18.7dB∼30.3dB, a perceptual quality of 3.11∼3.82, and a Mel cepstral distortion of 1.76dB∼3.20dB. The subjective test also shows an equal error rate of 0.33 against human perception, confirming its superior stealthiness. We release our source code and audio samples at **https://zju-muslab.github.io/projects/advreverb**.

II. Background and Related Work

A. Automatic Audio System

Fig. 1 illustrates the interaction between a user and a VUI device as well as its internal processing logic. Generally, the device keeps listening to surrounding sounds in a low-power state and detects human speech signals through a voice activity detection module. Then the detected speech signal is sent to the following automatic audio systems.

1) Keyword Spotting (KWS): The on-device KWS system spots predefined wake-up words, e.g., "OK Google", from the streaming speech signal, and then the device will be activated. Specifically, the KWS system first splits the input speech x into consecutive frames and then converts them to the frequency domain to extract acoustic features $h(x)$ such as Mel frequency cepstral coefficients, filter banks, and log Mel spectrum. These features are fed to a deep neural network (DNN) $f(\cdot)$ for classification, i.e., determining whether the input belongs to one of the predefined keywords Y or other unknown ones. Such a process is in essence a multi-class classification task, which can be formulated as follows:

$$F(x) = \underset{y \in \{Y, <unknown>\}}{\arg\max} f(h(x)), \tag{1}$$

where $F(\cdot)$ represents the entire audio system that consumes the input speech and outputs a decision result.

2) Automatic Speech Recognition (ASR): After being activated by KWS, the device awaits the following commands, which are usually uploaded to a cloud server for ASR due to the local limited resources. Similar to KWS, a typical ASR system first performs input pre-processing and feature extraction. Then speech decoding $f(\cdot)$ is applied to learn a sequence-to-sequence mapping from frame-wise features $h(x)$ to phoneme tokens y.

State-of-the-art systems apply end-to-end decoding based on Connectionist Temporal Classification (CTC) and ASR can be

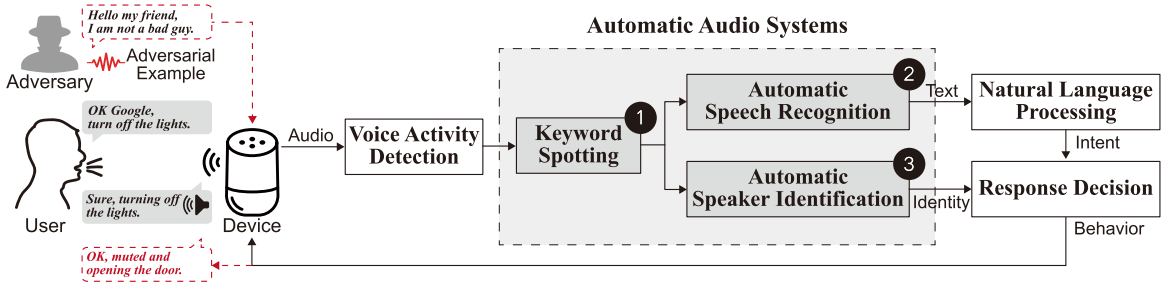


Fig. 1. Illustration of automatic audio system and audio adversarial example attack.

formulated as follows:

$$F(x) = \arg\max_{y \in Y} f(h(x)), \tag{2}$$

where Y denotes the potential paths derived by decoding.

3) Automatic Speaker Identification (ASI): Apart from ASR, the device also identifies speakers through ASI for user authentication or personalized interaction. In particular, the ASI system needs to identify legitimate users Y from other unenrolled speakers Y' for access control. To this end, ASI is defined as an open-set identification task involving training, enrollment, and testing. In the training stage, a DNN model $f(\cdot)$ is trained to learn distinguishable speaker embedding from the acoustic features $h(x)$. In the enrollment stage, each user's utterances are fed to the DNN model to extract an averaged embedding, which is stored as the user profile p. During the testing, the system extracts an embedding for each input speech and compares it with the user profiles P via a back-end scorer $s(\cdot)$, e.g., PLDA [32], cosine similarity. The access is accepted only if the similarity score exceeds a predefined threshold θ, otherwise rejected:

$$F(x) = \begin{cases} \arg\max_{y \in Y} s(f(h(x)), P) & \max s(f(h(x)), P) > \theta; \\ \text{reject} & \text{otherwise.} \end{cases} \tag{3}$$

After ASR, the transcribed text is translated into more specific intents by the natural language processing module. Combined with the identification result from ASI, the device decides how to respond to the user's request.

B. Audio Adversarial Example Attack

Behind the powerful capability of automatic audio systems, their vulnerability to adversarial example attacks raises increasing security concerns. Extensive research has demonstrated the feasibility of deceiving automatic audio systems $F(\cdot)$ by imposing subtle additive perturbations δ on benign samples x: $x' = x + \delta$ so that $F(x') \neq F(x)$ for an untargeted attack or $F(x') = y_t$ for a targeted attack. We focus on targeted attacks in this paper due to their larger practical threat. To ensure attack stealthiness, existing studies construct and deliver audio adversarial examples with different designs and adopt various metrics for evaluation, as summarized in Table I.

1) Perturbation Constraint: Typical adversarial example attacks query the target system and employ its gradient information to optimize additive perturbations iteratively. The generated perturbations are in general stochastic and irregular noises, significantly distorting the original audio. Hence, previous attacks confine the perturbation audibility in terms of amplitude and frequency. L_∞ norm [7], [8], [9], [10], [11], [12], [13] is the most straightforward and effective way to restrict the perturbation amplitude, with which the objective function of a typical targeted attack can be defined as:

$$\arg\min_{\delta} \mathcal{L}(F(x+\delta), y_t), \quad s.t.\ \|\delta\|_\infty \leq \epsilon, \tag{4}$$

where $\mathcal{L}(\cdot)$ refers to the system-specific loss function that measures the difference between the system output and desired target y_t, and ϵ denotes the perturbation budget size. This constrains the perturbation within a small range $[-\epsilon, \epsilon]$ so that it would not be noticed by humans. Besides, other studies [14], [15], [16], [17], [18], [19] propose to incorporate L_2 norm as a penalty in the objective function:

$$\arg\min_{\delta} \mathcal{L}(F(x+\delta), y_t) + \alpha \|\delta\|_2, \tag{5}$$

where α is used to tune the penalty weight. Moreover, some studies [20], [21], [22], [23], [24] also combine L_∞ norm and L_2 penalty together for stronger perturbation suppression. These amplitude suppression methods limit the perturbation audibility to some extent but still induce perceivable artifacts, particularly in the high-frequency bands. Towards this, the following studies [25], [26], [27], [28], [29], [30] turn to apply Psychoacoustic Masking (PsychoMask) to confine perturbation frequency:

$$\arg\min_{\delta} \mathcal{L}(F(x+\delta), y_t) + \alpha \sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \max\{S_\delta(k) - H_x(k), 0\}, \tag{6}$$

where S_δ is the normalized power spectral density of the perturbation, H_x is the hearing threshold derived from the original audio, and N is the frame window size. Under this constraint, the perturbations are forced to fall below the hearing threshold and masked by the original audio, thus becoming inaudible to humans.

2) Delivery Carrier: The original audio serves as a carrier to deliver the optimized perturbations. Some studies [9] suggest that the choice of delivery carrier has a large impact on the quality of adversarial examples. As shown in Table I, existing

TABLE I
STATE-OF-THE-ART ADVERSARIAL EXAMPLE ATTACKS AGAINST AUTOMATIC AUDIO SYSTEMS

Attack	Task	Delivery Carrier	Perturbation Constraint	Objective Metric	Subjective Test
Houdini [7]	ASR	Speech	L_∞ Norm	–	✓
CommanderSong [8]	ASR	Music		SNR	✓
Devil's Whisper [9]	ASR	Music		SNR	✓
Occam [10]	ASR,ASI	Music		SNR	✓
SEGA [11]	ASR	Speech		SNR, PCC	✗
FakeBob [12]	ASI	Speech		SNR	✓
PhoneyTalker [13]	ASI	Speech		SNR	✗
MGSA [14]	ASR	Speech	L_2 Penalty	SNR, PCC	✗
UAP(ASI) [15]	ASI	Speech		SNR, PESQ	✗
AS2T [16]	ASI	Speech		L_2 Distance, SNR, PESQ	✓
SirenAttack [17]	ASR,ASI,KWS	Speech		SNR	✓
Yakura et al. [18]	ASR	Music		SNR	✓
KENKU [19]	ASR	Music		SNR	✓
C&W [20]	ASR	Speech	L_∞ Norm+L_2 Penalty	SNR	✗
UAP(ASR) [21]	ASR	Speech		SNR	✗
AdvPulse [22]	ASI,KWS	Speech (pulse)		SNR	✗
Metamorph [23]	ASR	Music,Speech		MCD	✓
FAPG,UAPG [24]	ASI,KWS	Speech		SNR	✗
Schönherr et al. [25]	ASR	Music,Speech	Psychoacoustic Masking	SNR	✓
Qin et al. [26]	ASR	Speech		–	✓
Adversarial Music [27]	KWS	Music		–	✗
Imperio [28]	ASR	Speech		Segmental SNR	✗
VMask [29]	ASI	Speech		SNR	✗
SpecPatch [30]	ASR	Speech (patch)		L_2 Distance	✓

attacks select speech or music as the perturbation carrier, but a detailed comparison of their stealthiness is missing.

3) Objective Metric: Most previous studies treat the additive perturbations as noises and evaluate the attack stealthiness using the Signal-to-Noise Ratio (SNR), i.e., a higher SNR indicates less distortion and better stealthiness. Besides, Pearson Correlation Coefficient (PCC) and L_2 distance are also used to quantify the similarity and difference between the original audio and adversarial example. In addition, some studies also adopt perceptual metrics, e.g., Mel Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ), to approximately assess the quality of adversarial examples under human perception.

4) Subjective Test: Since the objective metrics cannot fully reflect the subjective perception of human listeners, many studies [7], [8], [9], [10], [12], [16], [17], [18], [23], [25], [26], [30] further conduct a listening test to evaluate the attack stealthiness. Volunteers of different genders and ages are recruited and asked to assess the quality of adversarial examples or distinguish them from original samples.

III. THREAT MODEL

Existing attacks pose substantial threats to automatic audio systems by pursuing extreme attack performance while sacrificing partial stealthiness, enlarging the risk of attack exposure. In this work, we focus on a stealthy adversarial example attack that effectively deceives audio systems without exposure.

We assume a targeted attack on a KWS, ASR, or ASI system, i.e., the adversary exploits adversarial examples to inject a specific speech keyword/command or impersonate a victim user. Considering that the audio systems are usually connected to smartphones or smart homes and store much personal information, this attack may induce harmful consequences such as property loss and privacy disclosure. In this attack, we focus on investigating the attack's stealthiness and assume a white-box setting, where the adversary is aware of the audio system implementation for adversarial example generation. Since the stealthiness design does not rely on prior knowledge about the system, further enhancement [11], [14] can be performed to realize a black-box attack. We demonstrate the black-box feasibility in Section VI-E. As shown in Fig. 2, we consider both digital and physical attack scenarios where the adversary delivers the adversarial example over the air (OTA) or over the line (OTL). In the OTA delivery, the adversary physically approaches the audio system and exploits a loudspeaker device to emit adversarial examples. In this case, the adversary suffers from exposure by human listeners in the vicinity of the audio system. In the OTL delivery, the adversary directly injects adversarial examples into audio systems through digital access, e.g., online API query. Such an attack suffers from the risk of exposure by human inspectors who conduct sampling and inspection in the backend. In both scenarios, the adversary must ensure the perceptual stealthiness of adversarial examples.

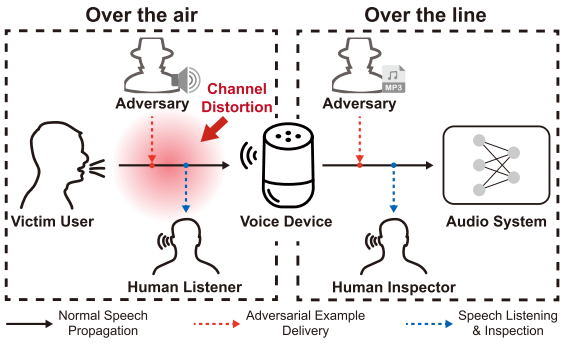


Fig. 2. Threat model.

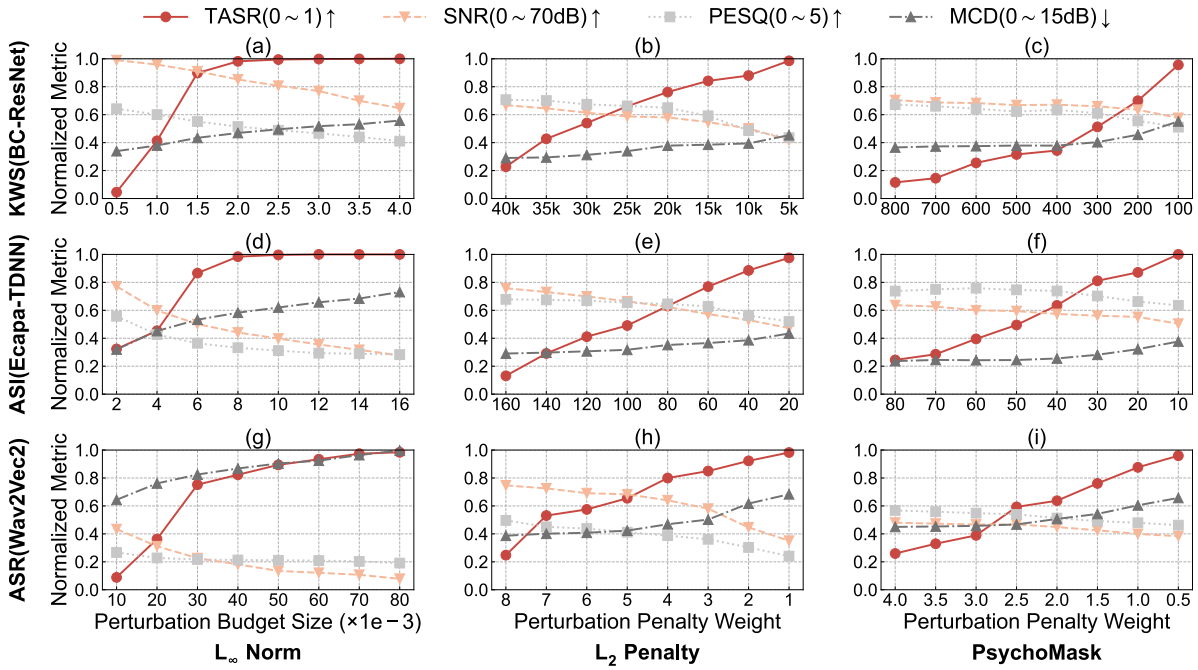


Fig. 3. Effectiveness and stealthiness of adversarial examples against audio systems with different perturbation constraints. The metrics are normalized within [0,1], i.e., the y-axis range represents 0∼1 for TASR, 0∼35dB for SNR, 0∼5 for PESQ, and 0∼15dB for MCD, respectively.

IV. Limitations of Additive Perturbation

According to the threat model, both effectiveness and stealthiness are indispensable for a successful adversarial example attack. We first evaluate existing additive audio adversarial example attacks and understand their limitations by answering **Q1**∼**Q3**.

A. Objective Evaluation

To answer these questions, we conduct experiments to attack state-of-the-art (SOTA) automatic audio systems including BC-ResNet [2] for KWS, Ecapa-TDNN [4] for ASI, and Wav2Vec2 [5] for ASR, respectively, whose details are introduced in Section VI-A. We select 1,750 keywords in Google Speech Command [33], 1,200 utterances in VoxCeleb [34], and 120 commands in LibriSpeech [35] as speech delivery carrier. For each trial, we employ a typical iterative gradient-based approach to generate adversarial examples under the perturbation constraint of L_∞ norm, L_2 penalty, and PsychoMask, respectively, according to Equation (4) to (6). Note that there are several variants of PsychoMask [25], [26], [27], [28], [29], [30] with different technical details. We follow the implementation of Qin et al. [26] considering it is a pioneering and representative attack with better code availability. We use an Adam optimizer with a learning rate of 0.001 and set the iteration steps to 200, 1,000, and 3,000 for KWS, ASI, and ASR respectively, due to their different levels of model complexity. We vary the perturbation budge size ϵ and penalty weight α to study their impact.

Then we feed the generated adversarial examples to the audio systems, count the targeted attack success rate (TASR), and assess their stealthiness using three common metrics, i.e., SNR, PESQ, and MCD, which are defined in Section VI-A. As shown in Fig. 3, for all the 9 attack combinations of audio systems and perturbation constraints, we can see that the TASR gradually grows to nearly 1.0 as the perturbation budget size increases or penalty weight decreases. Meanwhile, the SNR and PESQ largely degrade, and the MCD increases. This indicates a trade-off that relaxing the perturbation constraint boosts the attack performance but also induces larger distortion, which can answer **Q1**. Note that the x-axis range of perturbation budget size and penalty weights varies with tasks and constraints, due to their different levels of task complexity and distinct constraint spaces. Under the same constraint, the KWS system takes the least perturbation budget or the largest penalty weight to achieve high TASRs, followed by the ASI system, and the ASR system relaxes the constraint the most to reach similar TASRs. In addition, for the same audio system, different perturbation constraints also lead to distinct performances.

> **Insight 1.** The attack performance varies on different speech tasks and perturbation constraints, and there is a common trade-off between the effectiveness and stealthiness of audio adversarial example attacks.

Based on this observation, we further compare different perturbation constraints and delivery carriers for answering **Q2**. Specifically, we select the "turning point" for each combination in Fig. 3, where the adversarial examples achieve a TASR over 0.95 while causing distortion as little as possible so that we can compare their stealthiness in terms of SNR, PESQ, and MCD with similar effectiveness. Moreover, we also select music samples from 8 songs of distinct styles as music carriers, and each of them is split into segments of the same length as

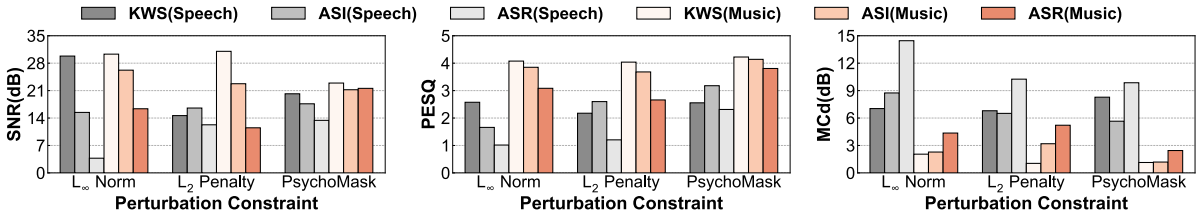


Fig. 4. Stealthiness comparison among adversarial examples based on speech and music carriers with different perturbation constraints.

speech samples, as introduced in Section VI-A. We also adopt the turning point of music carriers for comparison.

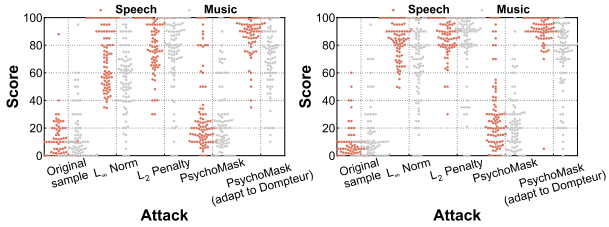
As shown in Fig. 4, for the speech carrier, L_∞ norm achieves the highest SNR and PESQ as well as the least MCD against the simplest KWS system but performs poor against the ASI and ASR systems that are more complex. By contrast, L_2 penalty and PsychoMask realize better SNR and PESQ as well as less MCD against ASI and ASR. We also find that although PsychoMask achieves sub-optimal SNR, it produces relatively excellent PESQ and MCD than L_∞ norm and L_2 penalty in most cases. This suggests that such a frequency masking scheme outperforms amplitude suppression approaches in reducing distortion. Furthermore, adversarial examples based on music carriers show a significant improvement in stealthiness. Especially with PsychoMask, the attack simultaneously achieves an SNR over 21dB, a PESQ around 4.0, and an MCD below 3dB, largely promoting the attack's stealthiness. This is probably because the music samples are filled with high-energy melodies that contribute to mask perturbations, while the speech samples contain more silence pieces that tend to make perturbations more obvious.

Insight 2. The frequency masking scheme produces superior stealthiness than amplitude suppression constraints, and music carriers contribute to a higher quality than speech carriers in most cases.

B. Subjective Evaluation

In addition, we further conduct a subjective listening test to evaluate the stealthiness to human perception for answering **Q3**. Specifically, we recruited 30 volunteers to participate in both OTA perception (audios are played by a loudspeaker 0.5m away) and OTL inspection (audios are played by a laptop and volunteers wear earphones). For each test, we play both original samples and adversarial examples in random order (96 test samples per volunteer). The volunteers need to assess each sample and give a score ranging 0∼100, representing how likely they think it is a noisy adversarial example but not a clean speech. Detailed settings are introduced in Section VI-A.

As shown in Fig. 5, we can observe an obvious separation of the score distribution between original samples and adversarial examples generated by L_∞ norm and L_2 penalty, and volunteers can distinguish these two kinds of attack attempts from benign inputs well with a low equal error rate (EER) of 0.07 and 0.06 on average. Instead, the scores of adversarial examples generated by PsychoMask distribute



(a) OTA perception. (b) OTL inspection.

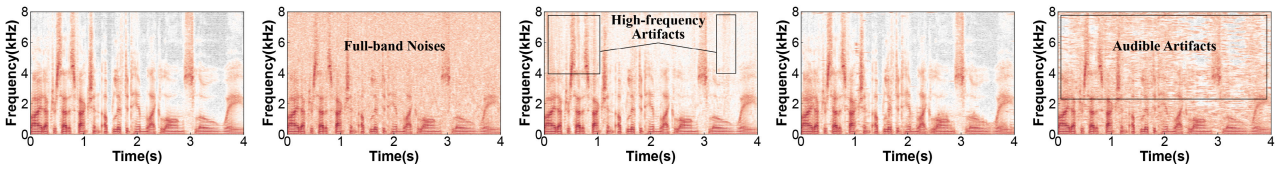
Fig. 5. Score distribution of original samples and adversarial examples based on speech and music carriers with different perturbation constraints.

rather close to those of original samples with a high EER of 0.28, validating its excellent stealthiness to human perception. However, a recent work Dompteur [31] has demonstrated that PsychoMask can be easily tamed by a simple filter. This filter removes inaudible components in the speech by multiplying the spectrum with a mask $M(n,k) = \mathbb{I}(S(n,k) > H(n,k) + \Phi)$, where S and H are the speech spectrum and its hearing threshold, $\mathbb{I}(\cdot)$ denotes the indicator function, and Φ is a parameter to tune the masking scale. Hence, we further apply Dompteur to study its impact on PsychoMask-based adversarial examples. As shown in Table II, the TASR of PsychoMask remarkably drops and even degrades to 0 as the filter gets more aggressive with Φ=0, 2, 4. The inaudible domain is totally discarded so that PsychoMask fails to hide perturbations. In this case, adaptive adversaries have to craft perturbations over the hearing threshold, i.e., audible to humans. To validate this, we enhance PsychoMask by considering Dompteur during the perturbation optimization. Table II shows that the TASR of PsychoMask (adaptive) restores to 0.92∼1.00, but as shown in Fig. 5, we find that the adversarial examples' scores are clearly separated from those of original samples with a low EER of 0.07.

Insight 3. Current additive adversarial examples under the constraint of amplitude suppression can be easily detected by humans, while frequency masking also fails to remain stealthy under the impact of Dompteur.

C. Limitation Analysis

To explore causes underlying the insufficient stealthiness of additive adversarial examples, we visualize the spectrum of an original sample and the corresponding adversarial examples. Compared with the original sample in Fig. 6(a), we can observe obvious full-band noises in Fig. 6(b) of L_∞ norm, which severely buries the original pitches and harmonics.



(a) Original sample. (b) L_∞ Norm. (c) L_2 Penalty. (d) PsychoMask. (e) PsychoMask (adaptive).

Fig. 6. Example of original sample and adversarial examples with different perturbation constraints.

TABLE II
TASR OF PSYCHOMASK WITH AND WITHOUT DOMPTEUR

Parameter	Speech Carrier			Music Carrier		
	KWS	ASI	ASR	KWS	ASI	ASR
w/o	0.956	1.000	0.959	0.993	1.000	1.000
Φ=0dB	0.217	0.012	0.000	0.396	0.464	0.000
Φ=2dB	0.080	0.000	0.000	0.133	0.081	0.000
Φ=4dB	0.071	0.000	0.000	0.022	0.000	0.000
Φ=4dB (adaptive)	0.946	1.000	0.922	0.995	1.000	1.000

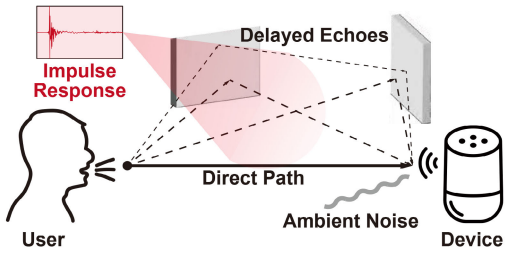


Fig. 7. Sound channel distortion over the air.

Although the noises get smaller in Fig. 6(c) of L_2 penalty, there are many irregular artifacts at the high-frequency band. These noises and artifacts can be easily perceived by human ears, making the victim aware of the attack attempt. Besides, PsychoMask yields a clean spectrum in Fig. 6(d) since perturbations are hidden below the hearing threshold. However, there are also audible artifacts in Fig. 6(e) when PsychoMask adapts to Dompteur. Therefore, current additive adversarial perturbations cannot address the effectiveness-stealthiness trade-off and a new solution is highly desired for successful attacks.

V. STEALTHY ADVERSARIAL EXAMPLE

A. Basic Idea

Faced with the effectiveness-stealthiness trade-off, we propose a paradigm shift: instead of suppressing the amplitude or frequency of perturbations to make them inaudible, the perturbations are allowed to be heard by human ears and would be perceived as natural inputs rather than artificially injected anomalous noises. To achieve this, we craft adversarial perturbations as a natural sound distortion to confuse humans. As shown in Fig. 2, normal speech signals from the victim user need to travel through the physical world so as to be recorded by the VUI device, where channel distortion is involved inevitably. The adversary can exploit this to deliver distortion-like adversarial perturbations so that human listeners and inspectors could hardly distinguish the adversarial examples from normal distorted speech.

Inspired by this, we propose a novel **convolutional adversarial perturbation** to approximate a natural reverberation effect. As shown in Fig. 7, reverberation is a common channel distortion due to the multi-path effect. The user's sound waves propagate omnidirectionally so that part of them propagates along the direct path and others are delayed due to absorption or reflection by obstacles. All these sounds arrive at the device and overlay with each other to form a reverberation effect. Theoretically, the reverberation process can be quantified as a convolution with impulse response r: $\hat{x} = x * r$, where x and $\hat{x}$ are the raw and reverberated signals respectively. Different from the existing additive adversarial example, i.e., $x' = x + \delta$, we turn to derive the adversarial example by $x' = x * r'$, here r' is called convolutional adversarial perturbation. Considering that temporal convolution is equivalent to the multiplication in the frequency domain, such adversarial examples in essence are derived by frequency filtering over original samples, enabling us to manipulate acoustic features for deceiving audio systems with less conspicuous distortion.

B. Attack Design

Based on the initial idea, we propose *AdvReverb* to implement the construction, optimization, and delivery of convolutional adversarial perturbations. Fig. 8 shows the overall design of *AdvReverb*, including carrier selection, forced alignment, phoneme-wise convolution, overlap-add concatenation, and gradient-based optimization.

1) Carrier Selection: Similar to existing attacks, *AdvReverb* employs both speech and music as the delivery carrier to support richer attack scenarios. For instance, speech carriers are less conspicuous in indoor multi-person conversations, while music carriers are more suitable for individual entertainment and relaxation spaces. As revealed in Section IV, adopting music samples to deliver additive adversarial perturbations yields higher quality than speech samples. Hence, we also compare the adaptability of these two kinds of carriers for convolutional adversarial perturbations.

2) Forced Alignment: Considering the different levels of scale and complexity among audio systems, it is difficult to accurately manipulate the local acoustic features by applying a single global convolutional adversarial perturbation. Especially for a targeted attack, perturbations need to approximate acoustic features of the target keyword, speaker identity, or speech text as possible. Towards this, we propose to divide speech into short segments and inject independent perturbations for each segment. Instead of simply dividing with a fixed length,

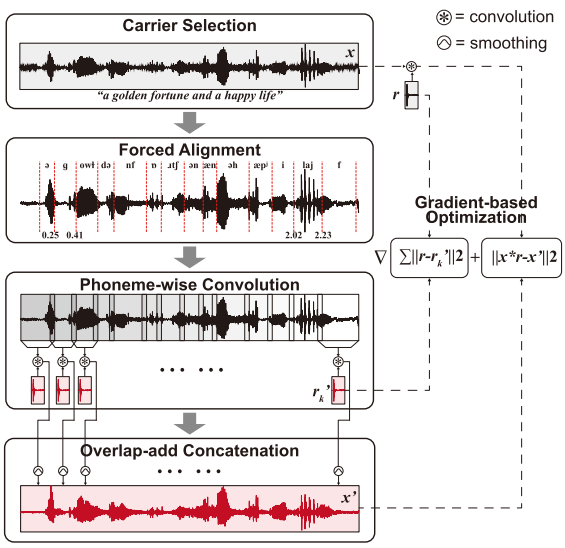


Fig. 8. System overview of *AdvReverb*.

Algorithm 1 Stealthy Adversarial Example Construction

Input: audio system $F(\cdot)$, delivery carrier x and its speech transcript t, target output y_t, real-world impulse response r, penalty weight α and β, learning rate η.
Output: stealthy adversarial example x'.

$\{(x_k, p_k)|T(x_k) = p_k\} \leftarrow FA(x, t)$
$r \leftarrow \frac{r}{\|r\|}$, $\{r'_k\} \leftarrow r$
for each step **do**
  for $k = 1, 2, \ldots, K$ **do**
    $x'_k = x_k * r'_k$
  end for
  $x' \leftarrow \sum_{k=0}^{K}(x'_k \otimes w)$
  $\mathcal{L}_{adv} \leftarrow \mathcal{L}(F(x'), y_t)$
  $\mathcal{L}_{pen\_ir} \leftarrow \sum_{k=1}^{K} \|r - r'_k\|_2$
  $\mathcal{L}_{pen\_ae} \leftarrow \|x * r - x'\|_2$
  $\mathcal{L}(r'_k) = \mathcal{L}_{adv} + \alpha(\mathcal{L}_{pen\_ir} + \mathcal{L}_{pen\_ae})$
  $r'_k \leftarrow r'_k - \eta \nabla_{r'_k} \mathcal{L}(r'_k)$
end for

we perturb the original speech at the phoneme level to avoid temporal inconsistency between segments.

During the offline adversarial example generation, both the speech signal and its orthographic transcript are known, e.g., the waveform and transcript "*a golden future and a happy life*" in Fig. 8. We exploit the forced alignment technique to align the speech x and transcript t and then derive the time duration of phonemes: $\{(x_k, p_k)|T(x_k) = p_k\}$, where x_k and p_k represent the speech segment and the corresponding phoneme label respectively, and $T(\cdot)$ stands for phoneme recognition. For example, the phoneme */g/* corresponds to [0.25s, 0.41s] of the waveform in Fig. 8. Next, we can perturb these temporally aligned phoneme segments with convolutional adversarial perturbations. Note that the minimum length of phoneme segments needs to be longer than the perturbation for convolution, so we cluster consecutive short phonemes together to form longer segments, e.g., */laj/* at [2.02s, 2.23s] in Fig. 8.

3) Phoneme-Wise Convolution: For each phoneme segment x_k, where $k \in \{1, \ldots, K\}$, *AdvReverb* initializes a convolutional adversarial perturbation r'_k based on a real-world impulse response r. Such impulse responses can be easily measured in the physical space or directly retrieved from some open-source datasets [36]. Although the phoneme segments vary in length, the duration of convolutional adversarial perturbation is fixed. Considering that the energy of impulse response concentrates at the beginning, we truncate r'_k to 0.1s to avoid long reverberation. Then we normalize the perturbation, i.e., $r'_k = \frac{r'_k}{\|r'_k\|}$, and convolve it with the corresponding segment in a phoneme-wise manner: $x'_k = x_k * r'_k$.

4) Overlap-Add Concatenation: Simply concatenating the perturbed segments results in noticeable artifacts at the segment boundaries, i.e., obvious "click" sounds, due to the sudden discontinuity of the waveform at the concatenating point. To address this, we employ an overlap-add concatenation to merge perturbed segments. In particular, we expand the time duration of each phoneme segment by 0.01s at the beginning so that the perturbed segments would overlap each other. For each concatenating point between two segments, we apply a Hanning window w of 0.02s to smooth the end 0.01s of the last segment from 1 to 0 and the start 0.01s of the next segment from 0 to 1. Then we can add the smoothed segments to form a complete adversarial example with continuous waveform: $x' = \sum_{k=0}^{K}(x'_k \otimes w)$, where $\otimes$ denotes the boundary smoothing operation.

5) Gradient-Based Optimization: The adversarial examples are fed to audio systems for querying output and retrieving gradient information, which is used to optimize the convolutional adversarial perturbations. For a targeted attack, *AdvReverb* optimizes the following adversarial objective:

$$\mathcal{L}_{adv} = \mathcal{L}(F(x'), y_t), \tag{7}$$

where $F(\cdot)$ refers to the audio system and $\mathcal{L}(\cdot)$ denotes its loss function, i.e., cross-entropy loss for KWS, cosine similarity scorer for ASI, and CTC loss for ASR. As for perturbation constraint, *AdvReverb* minimizes the difference between the adversarial example and the over-the-air sample that is convolved with the clean impulse response:

$$\mathcal{L}_{pen\_ae} = \|x * r - x'\|_2. \tag{8}$$

In addition, *AdvReverb* penalizes the difference between the adversarial perturbations and impulse response:

$$\mathcal{L}_{pen\_ir} = \frac{1}{K}\sum_{k=1}^{K} \|r - r'_k\|_2. \tag{9}$$

This enables *AdvReverb* to reshape the convolutional adversarial perturbations into real-world impulse responses to yield a natural reverberation. Based on this hybrid constraint, the objective function of *AdvReverb* is formulated as:

$$\arg\min_{r'_k} \mathcal{L}_{adv} + \alpha(\mathcal{L}_{pen\_ir} + \mathcal{L}_{pen\_ae}), \tag{10}$$

where α is a hyper-parameter for tuning the penalty weight. We summarize the whole procedure of stealthy adversarial example construction in Algorithm 1.

TABLE III
SOTA AUDIO SYSTEMS

Task	System	#Parameter	Source	Performance
KWS	BC-ResNet	0.4M	Reproduction	ACC=0.953
ASI	Ecapa-TDNN	22.1M	SpeechBrain	EER=0.008
ASR	Wav2Vec2	317.6M	SpeechBrain	WER=0.019

TABLE IV
SPEECH AND MUSIC DELIVERY CARRIERS

Dataset	#Class	#Utterance	Duration(s)
Google Speech Command	35	1,750	1.0
VoxCeleb1	40	1,200	3.5∼15.1
LibriSpeech	–	120	1.9∼10.7
Title	**Artist**	**Type**	**Duration(s)**
–	Bach	Violin	248
–	Van Halen	Guitar	736
–	Chopin	Piano	107
To The Sky	Owl City	Pop	220
Hello Seattle	Owl City	Pop	187
Rap God	Eminem	Rap	364
I Feel The Earth Move	Carole King	Pop rock	179
The Pretty Reckless	My Medicine	Hard rock	194

VI. EVALUATION

A. Experimental Setup

1) Audio Systems: We adopt three SOTA audio systems as the attack target, including BC-ResNet [2] for KWS, Ecapa-TDNN [4] for ASI, and Wav2Vec2 [5] for ASR. As shown in Table III, BC-ResNet is reproduced following the original paper and achieves high recognition accuracy (ACC) of 0.953 on Google Speech Command [33]. Ecapa-TDNN and Wav2Vec2 are pre-trained by SpeechBrain [37], and realize an Equal Error Rate (EER) of 0.008 on VoxCeleb1 [34] and a Word Error Rate (WER) of 0.019 on LibriSpeech [35] respectively. Note that these systems differ largely in the number of parameters, due to the different levels of task complexity.

2) Dataset Statistics: As shown in Table IV, we employ samples in Google Speech Command [33], VoxCeleb1 [34], and LibriSpeech [35] as speech carrier. Specifically, there are 35 classes of 1,750 keywords in Google Speech Command for KWS and 1,200 utterances from 40 speakers in VoxCeleb1 for ASI, and 120 transcribed utterances are selected from LibriSpeech for ASR. We also choose 8 songs of distinct styles as music carrier, which involves pure music of different instruments (e.g., violin, guitar, and music) and English songs of different genres (e.g., pop, rap, and hard rock). Each song is split into segments with a fixed length in different tasks, e.g., 1s for KWS, 3s for ASI, and 10s for ASR. Note that all these original samples are out of the training set and unseen for the audio systems.

3) Implementation Details: In each attack trial, we generate adversarial examples according to Algorithm 1. We adopt Montreal Forced Aligner [38] to perform forced alignment on the original speech, and divide the non-speech pure music segments into phoneme-level pieces of equal length (e.g., 0.2s). Several real-world impulse responses are selected from RVB2014 [36] as the initialization template of our perturbations. We set the length of impulse responses as 0.2s by default and study the impact of their different types and lengths in Section VI-D. Besides, we also tune the penalty weight α in Equation 10 to investigate the effectiveness-stealthiness trade-off. We employ an Adam optimizer with a learning rate of 0.001 and iterate 200, 1,000, and 3,000 steps for attacking KWS, ASI, and ASR, respectively.

4) Evaluation Metrics: We adopt the Targeted Attack Success Rate (TASR) to evaluate the attack effectiveness, which is defined as the ratio between the number of samples that are successfully recognized as the target output and the number of total attack attempts. Besides, we follow previous studies to evaluate the attack stealthiness using the following objective metrics: (1) Signal-to-Noise Ratio (SNR): $10\log_{10}\frac{P(x)}{P(x'-x)}$, where $P(x)$ and $P(x'-x)$ are the powers of the original speech and injected noise, i.e., the difference between the adversarial example and original sample. A higher SNR indicates less distortion and better stealthiness. (2) Mel Cepstral Distortion (MCD): $\frac{10}{\ln 10}\sqrt{2\sum\|mc_o - mc_a\|^2}$, where mc_o and mc_o are the Mel frequency coefficients of the original speech and adversarial example. The lower the MCD, the less the speech distortion. (3) Perceptual Evaluation of Speech Quality (PESQ): PESQ aligns the original speech with the adversarial example and predicts a quality score ranging from -0.5 to 4.5, indicating the quality from bad to excellent, which is further mapped to 1.02∼4.56 based on PESQ MOS-LQO to approximate the perceptual quality under human listening. Note that PESQ was originally designed for assessing human speech and we also extend it to evaluating high-quality songs, due to the lack of widely recognized metrics tailored to music-specific attributes. As it may not encompass the full spectrum of factors for music quality, we encourage readers to interpret our findings with an awareness of this metric choice.

5) Listening Test Settings: We recruit 30 volunteers (15 males and 15 females) aged 18∼48 to participate in a subjective listening test. The volunteers including graduate students and faculties are all good at listening, speaking, and writing in English without any hearing issues. Note that we obtain informed consent from each volunteer in advance, collect no personal information during the test, and compensate each volunteer with $5 after the test. All these subjective experiments on volunteers are validated by the institutional review board at our university. Consistent with the digital and physical delivery scenarios mentioned in the threat model, this test includes both over-the-air (OTA) and over-the-line (OTL) settings. In the OTA setting, each volunteer behaves as a human listener who listens to audio samples played by a loudspeaker 0.5m away. In the OTL setting, the volunteers act as human inspector who is asked to wear earphones for listening and checking audio samples. Before the test, we first have each volunteer listen to several original speech and music samples to refresh their perception and then play both original samples and adversarial examples in random order for testing (96 test samples per volunteer). The volunteers need to assess each test sample and give a score ranging 0∼100, which represents how likely they think it is a noisy adversarial example but not a clean sample.

TABLE V

OVERALL COMPARISON BETWEEN *AdvReverb* AND ADDITIVE ADVERSARIAL EXAMPLE ATTACKS AGAINST DIFFERENT AUDIO SYSTEMS

Task		KWS (BC-ResNet)							ASI (Ecapa-TDNN)							ASR (Wav2Vec2)						
L_∞ Norm	Budget($\times 10^{-3}$)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	2	4	6	8	10	12	14	10	20	30	40	50	60	70
	TASR↑	0.05	0.41	0.90	0.98	0.99	1.00	1.00	0.32	0.45	0.87	0.98	1.00	1.00	1.00	0.09	0.36	0.75	0.82	0.89	0.93	0.97
ADV vs. ORI	SNR(dB)↑	34.7	33.6	31.8	29.8	28.2	26.9	24.5	27.0	20.9	17.5	15.4	13.9	12.4	11.1	15.2	10.9	7.9	6.4	4.72	4.25	3.75
	PESQ↑	3.21	3.00	2.75	2.57	2.44	2.33	2.20	2.79	2.13	1.82	1.66	1.56	1.47	1.45	1.33	1.14	1.09	1.07	1.06	1.05	1.02
	MCD(dB)↓	5.09	5.69	6.51	7.04	7.43	7.77	7.98	4.80	6.78	7.99	8.74	9.30	9.86	10.2	9.66	11.4	12.3	13.0	13.5	13.8	14.4
L_2 Penalty	Penalty	35k	30k	25k	20k	15k	10k	5k	140	120	100	80	60	40	20	7	6	5	4	3	2	1
	TASR↑	0.43	0.54	0.66	0.76	0.84	0.88	0.99	0.29	0.41	0.49	0.63	0.77	0.89	0.98	0.53	0.57	0.65	0.80	0.85	0.92	0.98
ADV vs. ORI	SNR(dB)↑	22.6	21.5	20.6	20.3	19.1	17.5	14.6	25.6	24.5	23.3	21.8	20.0	18.6	16.6	25.4	24.2	23.9	22.4	20.3	15.6	12.3
	PESQ↑	3.51	3.37	3.31	3.25	2.96	2.43	2.18	3.37	3.34	3.27	3.23	3.14	2.81	2.60	2.25	2.19	2.06	1.94	1.81	1.51	1.20
	MCD(dB)↓	4.41	4.68	5.09	5.68	5.78	5.92	6.80	4.46	4.60	4.76	5.28	5.49	5.79	6.51	6.02	6.12	6.35	7.03	7.53	9.25	10.2
Psycho Mask	Penalty	700	600	500	400	300	200	100	70	60	50	40	30	20	10	3.50	3.00	2.50	2.00	1.50	1.00	0.50
	TASR↑	0.15	0.26	0.32	0.34	0.51	0.70	0.96	0.29	0.40	0.50	0.64	0.81	0.87	1.00	0.33	0.39	0.59	0.64	0.76	0.88	0.96
ADV vs. ORI	SNR(dB)↑	24.1	23.9	23.4	23.5	23.1	22.2	20.2	21.9	21.0	20.8	20.1	19.7	19.4	17.7	16.6	16.4	16.3	15.7	14.9	14.0	13.4
	PESQ↑	3.30	3.21	3.11	3.17	3.05	2.78	2.55	3.75	3.79	3.73	3.69	3.51	3.31	3.18	2.79	2.74	2.70	2.57	2.46	2.40	2.31
	MCD(dB)↓	5.59	5.62	5.68	5.68	6.05	6.85	8.28	3.67	3.64	3.66	3.83	4.23	4.83	5.65	6.79	6.86	7.01	7.60	8.14	9.07	9.86
Adv Reverb	Penalty	35	30	25	20	15	10	5	35	30	25	20	15	10	5	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	TASR↑	0.42	0.57	0.63	0.75	0.83	0.92	0.99	0.19	0.33	0.42	0.52	0.64	0.77	0.97	0.48	0.60	0.68	0.75	0.79	0.89	0.95
ADV vs. ORI	SNR(dB)↑	-4.3	-4.4	-4.4	-4.4	-4.4	-4.4	-4.4	-4.0	-4.1	-4.1	-4.1	-4.1	-4.1	-4.1	-4.2	-4.3	-4.3	-4.5	-4.7	-4.8	-5.0
	PESQ↑	1.91	1.92	1.91	1.90	1.90	1.90	1.90	1.64	1.63	1.63	1.63	1.62	1.62	1.62	1.44	1.40	1.38	1.38	1.37	1.37	1.36
	MCD(dB)↓	9.64	9.69	9.73	9.76	9.75	9.77	9.78	6.37	6.41	6.43	6.43	6.46	6.49	6.51	8.19	8.31	8.57	8.63	8.63	8.73	8.90
ADV vs. OTA	SNR(dB)↑	32.2	32.0	31.9	31.1	30.6	30.5	**30.3**	31.4	31.2	30.8	30.5	29.9	29.8	**29.1**	24.7	24.2	22.7	21.8	21.1	20.1	**18.7**
	PESQ↑	3.70	3.68	3.64	3.61	3.58	3.55	**3.57**	4.16	4.13	4.07	3.98	3.91	3.87	**3.82**	3.84	3.79	3.74	3.57	3.44	3.31	**3.11**
	MCD(dB)↓	2.34	2.39	2.45	2.52	2.55	2.63	**2.79**	1.43	1.46	1.49	1.57	1.60	1.66	**1.76**	2.19	2.31	2.39	2.55	2.65	2.85	**3.20**

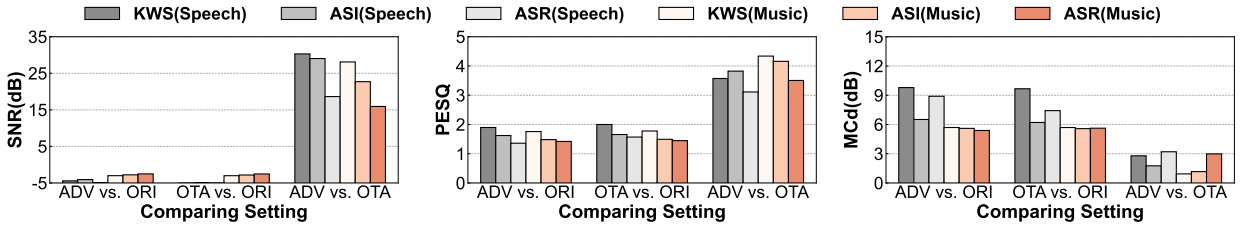


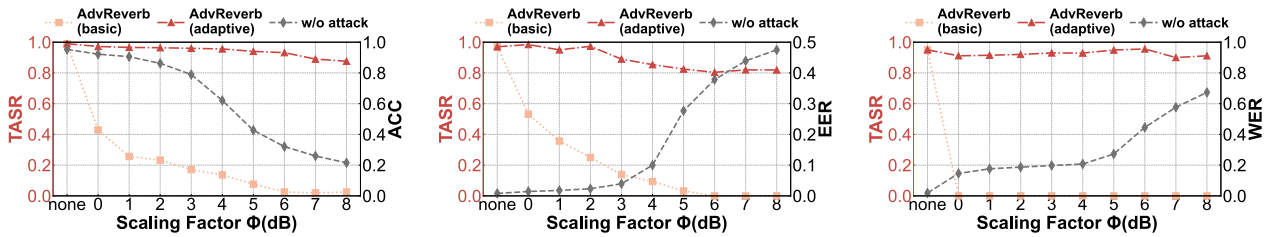
Fig. 9. Stealthiness of adversarial examples generated by *AdvReverb* based on speech and music carriers.

B. Overall Effectiveness

We first evaluate the overall effectiveness of *AdvReverb* against different audio systems in terms of attack performance and objective speech quality.

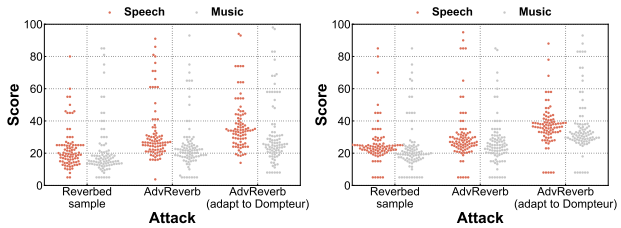
Similar to the methodology of evaluating baseline attacks in Section IV, we vary the penalty weight to find a "turning point" where the TASR exceeds 0.95 while minimizing the speech quality degradation. As summarized in Table V, we use a gray background to highlight the results at the "turning points" of the three baselines and *AdvReverb*. We can see that these additive adversarial example attacks attain high TASRs over 0.95, but at the cost of insufficient stealthiness with low SNR of 3.8dB∼29.8dB, low PESQ of 1.02∼3.18, and high MCD of 5.6dB∼14.4dB. As for *AdvReverb*, the TASRs against three audio systems increase to 0.99, 0.97, and 0.95 as the penalty weight gets smaller. Comparing the convolutional adversarial examples with original samples, i.e., ADV vs. ORI, we can observe that the speech quality significantly declines even with SNRs less than -4.00dB, PESQs below 1.90 and high MCDs ranging 6.51dB∼9.78dB. This is caused by the nature of convolution operation that largely changes the waveform, while these objective metrics are designed for the additive operation that subtly perturbs the waveform, and thus cannot directly reflect the perceptual quality of *AdvReverb*. For a fair comparison, we need to compare the convolutional adversarial examples to normal over-the-air samples that have convolved with clean impulse responses, i.e., ADV vs. OTA. As shown in Table V, *AdvReverb* achieves rather high SNRs of 30.3dB, 29.1dB, and 18.7dB and PESQs of 3.57, 3.82, and 3.11 as well as extremely low MCDs of 2.79, 1.76, and 3.20, validating the excellent speech quality of convolutional adversarial examples and its high similarity to over-the-air samples with natural reverberation.

In addition to the aforementioned study on speech carriers, we also adopt music carriers to evaluate their overall effectiveness. Here we use the same penalty weight as speech carriers at the "turning points" and present the objective metrics of ADV vs. ORI, OTA vs. ORI, and ADV vs. OTA settings together, as shown in Fig 4. On the one hand, for all three audio systems and both two carriers, the SNR, PSEQ, and MCD of ADV vs. ORI and OTA vs. ORI trials are obviously poor and quite similar, indicating that the quality degradation mostly results from convolution operation itself but not our adversarial perturbation. On the other hand, for ADV vs. OTA setting, *AdvReverb* reaches excellent audio quality on both speech and music carriers with high SNRs at least 15dB, excellent PESQs over 3, and minute MCDs below 3dB, outperforming existing additive attacks. This reveals that our convolutional adversarial perturbations approximate realistic



(a) KWS (BC-ResNet). (b) ASI (Ecapa-TDNN). (c) ASR (Wav2Vec2).

Fig. 10. Impact of Dompteur on the attack effectiveness of *AdvReverb* and the normal performance of audio systems.



(a) OTA perception. (b) OTL inspection.

Fig. 11. Score distribution of original samples and adversarial examples based on speech and music carriers with different perturbation constraints.

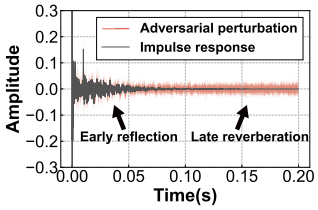


Fig. 12. Waveform of an impulse response and convolutional adversarial perturbations.

impulse responses well, which contributes to disguising the attack as a natural reverberation.

We also apply Dompteur with different scaling factors on basic and adaptive versions of *AdvReverb*. As shown in Fig. 10, compared to the performance without Dompteur (denoted as none on the x-axis), the TASRs drop below 0.5 and even gradually decrease to 0 as the scaling factor grows. Meanwhile, the normal performance of audio systems also degrades, i.e., the ACC of KWS drops below 0.2 and the EER of ASI and WER of ASR rise over 0.5. This is due to the significant distortion of Dompteur that undermines both the original samples and adversarial examples. Despite this, we found that *AdvReverb* still exhibits better robustness than PsychoMask when compared to Table II. And *AdvReverb* (adaptive) retains high TASRs over 0.8 even under very aggressive filtering with ϕ=8, suggesting the robustness of our convolutional adversarial perturbations.

C. Perceptual Stealthiness

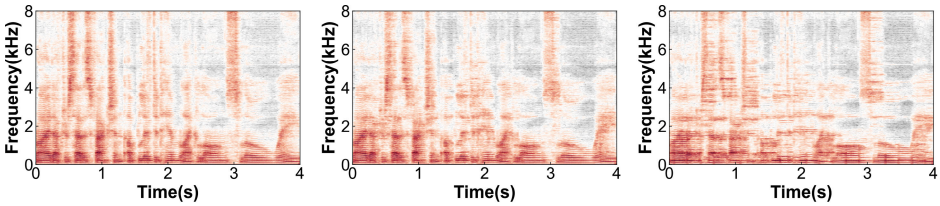
To validate the perceptual stealthiness of *AdvReverb*, we follow the listening test settings described in Section VI-A. The test samples include over-the-air samples that have been convolved with clean impulse responses and adversarial examples generated by *AdvReverb*. Similar to the test on PsychoMask, we also consider the enhanced version of *AdvReverb* (adaptive) to investigate whether it can still remain stealthy under the impact of Dompteur. Here we adopt a scaling factor ϕ=4, which indicates the most aggressive filtering without impairing the normal performance of audio systems too much, according to Fig. 10. As shown in Fig. 11, we can observe that the score distributions of over-the-air samples are very close to those of adversarial examples of *AdvReverb* even under the impact of Dompteur. This leads to high EERs of 0.37 and 0.22, showing that it is difficult for humans to distinguish the convolutional adversarial examples from natural reverberations.

To better understand the stealthiness of *AdvReverb*, we also visualize the convolutional adversarial perturbations and examples. As shown in Fig. 12, we can see negligible differences between the waveforms of perturbations and impulse responses at the early reflection range that has large amplitude and energy. This indicates that our convolutional adversarial perturbations learn the general structure of impulse responses well, thus providing superior perceptual stealthiness. Moreover, their main differences concentrate on the late reverberation range with small amplitude and energy, which contributes most to the attack success. Fig. 13 shows the spectrum of an over-the-air sample and convolutional adversarial examples with and without adapting to Dompteur. Compared to the original sample in Fig. 6(a), there are slight blurs in the over-the-air sample and adversarial examples but it is difficult to distinguish them, i.e., whether they are caused by natural reverberations or convolutional perturbations. This is consistent with *AdvReverb*'s basic idea of disguising the perturbation as normal channel distortion for stealthy attacks. Unlike PsychoMask which has to sacrifice stealthiness to achieve a successful attack, the enhanced adversarial example also remains a clear spectrum without obvious artifacts. This confirms that *AdvReverb* could overcome the effectiveness-stealthiness trade-off.

D. Micro-Benchmarks

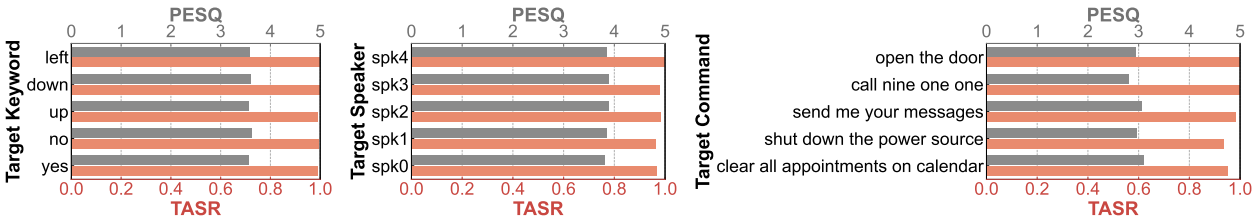
In this section, we study the impact of several micro-benchmarks on *AdvReverb*, including attack target class, impulse response template, and adversarial perturbation length. For simplicity, we adopt speech as the default delivery carrier and evaluate the objective speech quality by PESQ.

We first conduct experiments on different attack target classes, i.e., 5 keywords for KWS, 5 speakers for ASI, and 5 commands for ASR, while keeping all other conditions the same. Fig. 14 describes the TASR and PESQ of *AdvReverb*,



(a) Over-the-air sample. (b) Adversarial example. (c) Adversarial example (Φ=4).

Fig. 13. Spectrum of an over-the-air sample and convolutional adversarial examples with and without enhancement against Dompteur.



(a) KWS (BC-ResNet). (b) ASI (Ecapa-TDNN). (c) ASR (Wav2Vec2).

Fig. 14. TASR and PESQ of *AdvReverb* on different target classes.

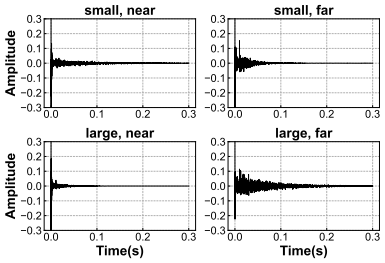


Fig. 15. Waveform of impulse responses in small or large rooms at near or far distances.

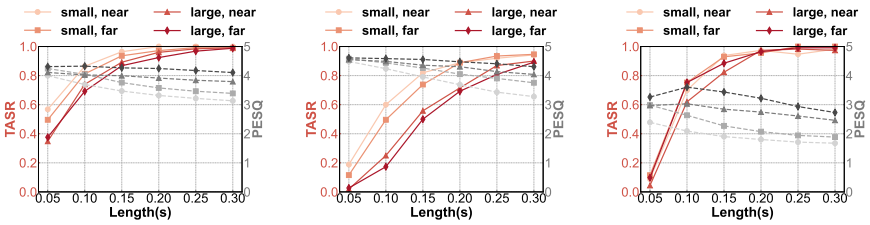
from which we can see that the TASRs approach 1.0 and the PESQs also remain steady over different targets, validating the excellent generalization ability of *AdvReverb*. Then we select four different impulse responses as the initialization template, which are measured in a small or large room at near or far distances. As shown in Fig. 15, We can see stronger reverberation in larger rooms and at far distances. We also vary the perturbation length from 0.05s to 0.30s with different templates to investigate their impact together. As shown in Fig. 16, the TASRs grow to nearly 1.0 while the PESQs degrade as the perturbation length increases. This is because longer perturbations provide a larger space to search for a successful adversarial example and meanwhile induce more distortion as well. Besides, we also find that impulse responses with weaker reverberation lead to relatively better TASRs but yield lower PESQs. This may be due to their longer and weaker late reverberation that can be easily perturbed for improving the success probability and also causing larger distortion. Based on this, the adversary can select appropriate impulse responses as the initialization template and configure a suitable perturbation length for the attack.

E. Black-Box Feasibility

Benefiting from the system-independent design, *AdvReverb* can be easily extended to black-box scenarios where the implementation details of the target system are agnostic to the adversary. Considering that *AdvReverb* focuses on improving perturbation stealthiness rather than proposing a new black-box attack, we directly implement a query-based black-box attack based on the gradient estimation method in FakeBob [12], which can also be realized using any other black-box strategy, such as Occam [10], SEGA [11], and MGSA [14].

Particularly, we leverage the natural evolution strategy (NES) to estimate gradients from the system's output: $\nabla_{x'}F(x') = \frac{1}{m\times\sigma}\sum_{i=1}^{m}F(x'_i)\times u_i$, where $x'_i = x' + \sigma \times u_i$ and u_i represents a Gaussian noise. However, the estimated gradients cannot be applied to construct adversarial examples in *AdvReverb*, due to the convolution operation. To this end, we further estimate the gradient related to the convolutional perturbation: $\nabla_{r'}F(x') = \nabla_{x'}F(x') * r'$, and then update the perturbation through gradient descent: $r' \leftarrow r' - \eta \times \nabla_{r'}F(x')$. Then we can iteratively optimize the perturbation with the estimated gradient and finally derive the adversarial example: $x' = x * r'$. We adopt the default parameters in FakeBob for NES, i.e., m=50, σ=0.001, η=0.001, and set ϵ as 0.01 to bound the perturbation scale. We implement this NES-enhanced *AdvReverb* against KWS, ASI, and ASR and compare it with KENKU [19], an efficient and stealthy black-box adversarial example attack against ASR. Unlike the gradient estimation scheme, KENKU crafts perturbations to directly manipulate MFCC features without querying the target system, by optimizing an acoustic feature loss and a perturbation loss: $\|MFCC(x') - MFCC(x)\|_2 + \lambda \times \|\delta\|_2$. Following the original paper, we query the ASR system and perform a binary search to tune λ on a small dataset. Then we use the tuned λ to generate each adversarial example through 10,000 iterations. Considering KENKU adopts music clips as delivery carriers, we also choose music carriers for a fair comparison.

As shown in Table VI, NES-enhanced *AdvReverb* achieves high TASRs of 0.93, 1.00, and 0.73 against black-box systems. However, this is at the cost of degradation in stealthiness and high time complexity, e.g., it requires 2,300∼63,103 system



(a) KWS (BC-ResNet). (b) ASI (Ecapa-TDNN). (c) ASR (Wav2Vec2).

Fig. 16. TASR and PESQ of *AdvReverb* using different impulse responses as initialization template with different perturbation lengths.

TABLE VI

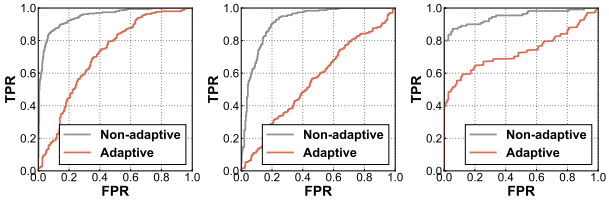
BLACK-BOX PERFORMANCE OF NES-ENHANCED *AdvReverb* AND KENKU WITH MUSIC CARRIERS

Metric	AdvReverb			KENKU
	KWS	ASI	ASR	ASR
TASR↑	0.93	1.00	0.73	0.75
SNR(dB)↑	8.25	8.04	7.27	7.92
PESQ↑	2.81	2.54	1.98	1.12
MCD(dB)↓	2.19	2.30	3.78	14.56
#System Queries↓	16,349	2,352	63,103	∼10
Optimization Time(s)↓	29.58	107.21	409.18	13.17

queries and takes about 30s∼400s for our server (Ubuntu 21.04 and Intel Xeon Gold 6226R 2.9GHz with 256G RAM and 16 cores) to optimize a single adversarial example. This indicates a challenging problem of balancing effectiveness and stealthiness under the black-box setting. By contrast, KENKU requires only about 10 queries to tune λ and spends 13.17s in optimization, owing to the acoustic feature-level generation. However, we find that at a similar TASR, KENKU shows a higher SNR of 7.92dB, a lower PESQ of 1.12, and a higher MCD of 14.56dB. We think this is due to the difference between additive and convolutional perturbations. More specifically, *AdvReverb*'s convolution operation changes the original waveform a lot but retains most of the acoustic features, while KENKU injects additive perturbations that preserve the original waveform but largely distort the MFCC features. According to our evaluation in Section IV, such additive adversarial example attacks that aim to preserve the waveform result in insufficient stealthiness. In addition, considering that MFCC follows the principle of the human auditory system and reflects the perceptual audio quality, we think directly distorting MFCC features also goes against the design goal of attack stealthiness.

F. Detection Evasion

Various methods are proposed to detect additive adversarial examples exploiting their instability, such as distorting perturbations through signal pre-processing [39], [40], [41], corrupting temporal dependencies by speech segmentation [42], etc. A recent defense [43] further proposes noise padding and sound reverberation to distort perturbations and corrupt their temporal dependencies simultaneously. Hence, we implement this detection method to investigate its performance in identifying our convolutional adversarial examples.



(a) KWS (b) ASI (c) ASR

Fig. 17. ROC of work [43] on detecting convolutional adversarial examples.

Following the original paper, we convolve each benign or adversarial sample with 4 random room impulse responses to perform reverberation and plant Gaussian noises in their silence parts to destroy the perturbation continuity. We feed the processed samples to audio systems and extract logit vectors to calculate the cosine similarity for detection. As shown in Fig. 17, the defense [43] can detect non-adaptive *AdvReverb* well with AUCs of 0.94, 0.91, and 0.95 on KWS, ASI, and ASR, respectively. However, the AUCs degrade to 0.71, 0.55, and 0.73 against adaptive *AdvReverb* where the adversary also performs the same noise padding and reverberation process to minimize the similarity variation. In this case, *AdvReverb* also realizes high TASRs of 0.85, 0.79, and 0.98, indicating that the adversary still has a high probability of bypassing the detection to successfully deceive the target system.

VII. DISCUSSION

We discuss several potential countermeasures against *AdvReverb* at different levels.

A. Source-Level Liveness Detection

The adversarial examples need to be played over the air to launch the attack, thus could be rejected by liveness detection, including active [44], [45] and passive [46], [47] schemes that depend on an emitted sensing signal or only the received speech. We think this would be an effective countermeasure against *AdvReverb*, but additional device is required, e.g., a multi-microphone array or other signal transmission devices.

B. Data-Level Perturbation Purification

Various purification methods based on signal processing techniques are proposed to remove adversarial perturbations, such as smoothing, filtering, squeezing, and quantization. However, such aggressive defenses impair both perturbations and benign speech. Defenders need to be careful about the trade-off between defense performance and the impact on audio systems.

C. Model-Level Adversarial Training

By taking adversarial examples into the training dataset, adversarial training [48] could enhance the audio systems to learn fine-grained differences between real-world reverberations and convolutional adversarial perturbations. This is a promising approach to defend *AdvReverb*. However, we don't think this is an ideal defense due to its requirement for adversarial perturbation generation and model re-training. Moreover, adaptive adversaries can also bypass adversarial training-enhanced systems, leading to an ever-lasting cat-and-mouse game.

VIII. Conclusion

This paper rethinks the stealthiness of audio adversarial examples and reveals the common effectiveness-stealthiness trade-off, which exposes the insufficient stealthiness of current attack paradigms. Towards this, we craft a novel convolutional adversarial perturbation as natural reverberation for deceiving humans. Then we design *AdvReverb*, an effective and stealthy attack to construct, optimize, and deliver convolutional adversarial examples. Both objective and subjective experimental results validate that *AdvReverb* could achieve a superior attack success rate on three audio tasks without being detected by humans, overcoming the effectiveness-stealthiness trade-off.

References

[1] S. Choi et al., "Temporal convolution for real-time keyword spotting on mobile devices," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 3372–3376.

[2] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 4538–4542.

[3] 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)*, Alberta, AB, Canada, Apr. 2018, pp. 5329–5333.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, Oct. 2020, pp. 3830–3834.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[6] A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022, *arXiv:2212.04356*.

[7] M. Cissé, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 6977–6987.

[8] X. Yuan et al., "CommanderSong: A systematic approach for practical adversarial voice recognition," in *Proc. USENIX Secur.*, Baltimore, MD, USA, 2018, pp. 49–64.

[9] Y. Chen et al., "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proc. USENIX Secur.*, Boston, MA, USA, 2020, pp. 2667–2684.

[10] B. Zheng et al., "Black-box adversarial attacks on commercial speech platforms with minimal information," in *Proc. ACM CCS*, 2021, pp. 86–107.

[11] Q. Wang, B. Zheng, Q. Li, C. Shen, and Z. Ba, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 896–908, 2021.

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

[13] M. Chen, L. Lu, Z. Ba, and K. Ren, "PhoneyTalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition," in *Proc. IEEE Conf. Comput. Commun.*, May 2022, pp. 1419–1428.

[14] S. Wang, Z. Zhang, G. Zhu, X. Zhang, Y. Zhou, and J. Huang, "Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 351–364, 2023.

[15] J. Li et al., "Universal adversarial perturbations generative network for speaker recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, U.K., Jul. 2020, pp. 1–6.

[16] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems," *IEEE Trans. Dependable Secure Comput.*, early access, pp. 1–17, 2022.

[17] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "SirenAttack: Generating adversarial audio for end-to-end acoustic systems," in *Proc. 15th ACM Asia Conf. Comput. Commun. Secur.*, New York, NY, USA, Oct. 2020, pp. 357–369.

[18] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Macao, China, Aug. 2019, pp. 5334–5341.

[19] X. Wu et al., "KENKU: Towards efficient and stealthy black-box adversarial attacks against ASR systems," in *Proc. USENIX Secur.*, Anaheim, CA, USA, 2023, pp. 247–264.

[20] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, San Francisco, CA, USA, May 2018, pp. 1–7.

[21] P. Neekhara, S. Hussain, P. Pandey, S. Dubnov, J. McAuley, and F. Koushanfar, "Universal adversarial perturbations for speech recognition systems," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 481–485.

[22] 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. Secur.*, New York, NY, USA, Oct. 2020, pp. 1121–1134.

[23] T. Chen, L. Shangguan, Z. Li, and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2020.

[24] 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*, 2021, vol. 35, no. 16, pp. 14129–14137.

[25] L. Schonherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2019, pp. 1–15.

[26] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proc. ICML*, Long Beach, CA, USA, 2019, pp. 5231–5240.

[27] J. Li, S. Qu, X. Li, J. Szurley, J. Z. Kolter, and F. Metze, "Adversarial music: Real world audio adversary against wake-word detection system," in *Proc. NeurIPS*, Vancouver, BC, Canada, 2019, pp. 11908–11918.

[28] L. Schönherr, T. Eisenhofer, S. Zeiler, T. Holz, and D. Kolossa, "Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems," in *Proc. Annu. Comput. Secur. Appl. Conf.*, New York, NY, USA, Dec. 2020, pp. 843–855.

[29] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 377–386.

[30] H. Guo, Y. Wang, N. Ivanov, L. Xiao, and Q. Yan, "SPECPATCH: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Los Angeles, CA, USA, Nov. 2022, pp. 1353–1366.

[31] T. Eisenhofer, L. Schönherr, J. Frank, L. Speckemeier, D. Kolossa, and T. Holz, "Dompteur: Taming audio adversarial examples," in *Proc. USENIX Secur.*, 2021, pp. 2309–2326.

[32] J. Portêlo, A. Abad, B. Raj, and I. Trancoso, "Secure binary embeddings of front-end factor analysis for privacy preserving speaker verification," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2494–2498.

[33] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018, *arXiv:1804.03209*.

[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 2616–2620.

[35] 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)*, South Brisbane, QLD, Australia, Apr. 2015, pp. 5206–5210.

[36] K. Kinoshita et al., "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 7, Dec. 2016.

[37] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.

[38] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 498–502.

[39] K. Rajaratnam and J. Kalita, "Noise flooding for detecting audio adversarial examples against automatic speech recognition," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Louisville, KY, USA, Dec. 2018, pp. 197–201.

[40] H. Kwon, H. Yoon, and K.-W. Park, "POSTER: Detecting audio adversarial example through audio modification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, London, U.K., Nov. 2019, pp. 2521–2523.

[41] S. Hussain, P. Neekhara, S. Dubnov, J. J. McAuley, and F. Koushanfar, "WaveGuard: Understanding and mitigating audio adversarial examples," in *Proc. USENIX Secur.*, 2021, pp. 2273–2290.

[42] H. Zhang, P. Zhou, Q. Yan, and X.-Y. Liu, "Generating robust audio adversarial examples with temporal dependency," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, New Orleans, LA, USA, Jul. 2020, pp. 3167–3173.

[43] X. Du, C.-M. Pun, and Z. Zhang, "A unified framework for detecting audio adversarial examples," in *Proc. 28th ACM Int. Conf. Multimedia*, WA, WA, USA, Oct. 2020, pp. 3986–3994.

[44] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dallas, TX, USA, Oct. 2017, pp. 57–71.

[45] L. Lu et al., "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1466–1474.

[46] M. E. Ahmed, I. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proc. USENIX Secur.*, Boston, MA, USA, 2020, pp. 2685–2702.

[47] Z. Li et al., "Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2021, pp. 1884–1899.

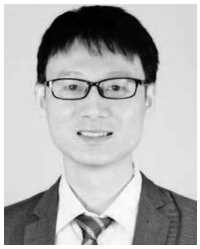
[48] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–11.

Meng Chen (Graduate Student Member, IEEE) received the B.E. degree in software engineering from Zhejiang University, where he is currently pursuing the Ph.D. degree with the School of Cyber Science and Technology. His research interests include mobile computing and AI security. He was a recipient of the Best Poster Runner-Up Award from ACM MobiCom 2022 and the Student Travel Grant of IEEE INFOCOM 2022.

Li Lu (Member, IEEE) received the B.E. degree in computer science and technology from Xi'an Jiaotong University and the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University. He is currently a tenure-track Research Professor with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University. He was also a Visiting Research Student with the Wireless Information Network Laboratory (WINLAB) and the Department of Electrical and Computer Engineering, Rutgers University. His research interests include IoT security, intelligent voice security, mobile sensing, and ubiquitous computing. He was a recipient of the ACM China SIGAPP Chapter Rising Star Award, the ACM China SIGAPP Chapter Doctoral Dissertation Award, the Best Poster Runner-Up Award from ACM MobiCom 2022, and the First Runner-Up Poster Award from ACM MobiCom 2019.

Jiadi Yu (Senior Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Prior to joining Shanghai Jiao Tong University, he was with the Stevens Institute of Technology, USA, as a Post-Doctoral Researcher. He has published more than 100 refereed papers in international leading journals and key conferences in the areas of wireless communications and networking, mobile computing, and security and privacy. His current research interests include mobile computing and sensing, cyber security and privacy, the Internet of Things (IoT), and smart healthcare. He is a Senior Member of the IEEE Communication Society.

Zhongjie Ba (Member, IEEE) received the Ph.D. degree in computer science and engineering from the State University of New York at Buffalo in 2019. He is currently a ZJU100 Young Professor with the College of Computer Science and Technology and the Institute of Cyberspace Research (ICSR), Zhejiang University, Hangzhou, China. He was a Post-Doctoral Researcher with the School of Computer Science, McGill University. His current research interests include the security and privacy aspects of the Internet of Things, artificial intelligence-powered mobile sensing, and forensic analysis of multimedia contents.



Feng Lin (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN, USA, in 2015. He is currently a Professor with the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. He was an Assistant Professor with the University of Colorado Denver, Denver, CO, USA; a Research Scientist with the State University of New York (SUNY) at Buffalo, Buffalo, NY, USA; and an Engineer with Alcatel-Lucent (currently Nokia). His current research interests include mobile sensing, Internet of Things security, biometrics, AI security, and IoT applications. He was a recipient of the Best Paper Award from ACM MobiSys20, the IEEE Globecom19, the IEEE BHI17, the Best Demo Award from ACM HotMobile8, and the First Prize Design Award from the 2016 International 3-D Printing Competition.

Kui Ren (Fellow, IEEE) received the Ph.D. degree from the Worcester Polytechnic Institute, Worcester, MA, USA. He is currently the Dean and a Professor with the School of Cyber Science and Technology, Zhejiang University. He has published extensively in peer-reviewed journals and conferences. He is a fellow of ACM; a distinguished lecturer of IEEE; and a past board member of the Internet Privacy Task Force, State of Illinois. He received several best paper awards, including the IEEE ICDCS 2017, IWQoS 2017, and ICNP 2011. He also received the NSF CAREER Award in 2011, the Sigma Xi/IIT Research Excellence Award in 2012, the UB SEAS Senior Researcher of the Year Award in 2015, the UB Exceptional Scholar Award for Sustained Achievement in 2016, and the IEEE CISTC Technical Recognition Award in 2017. He serves on the editorial boards for IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON SERVICE COMPUTING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, and *SpringerBriefs on Cyber Security Systems and Networks*.