Poster: Room-Scale Over-the-Air Audio Adversarial Examples

Tao Chen
City University of Hong Kong
tachen6-c@my.cityu.edu.hk

Longfei Shangguan
Microsoft
longfei.shangguan@microsoft.com

Zhenjiang Li
City University of Hong Kong
zhenjiang.li@cityu.edu.hk

Kyle Jamieson
Princeton University
kylej@cs.princeton.edu

***Abstract*—This poster presents our recent work Metamorph, a system that can generate over-the-air audio adversarial examples working in a room scale. We find that the device and channel frequency selectivity with different characteristics could fail the previous audio adversarial attacks, and we propose a generate-and-clean two-phase design to tackle this issue. Evaluation shows the effectiveness of the Metamorph design in both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) environments.**

I. Introduction

Driven by deep neural networks (DNN), speech recognition (SR) techniques are advancing rapidly and are widely used as a convenient human-computer interface in many daily life scenarios. However, recent studies [1], [3] have investigated a crucial problem — given any audio clip I (with transcript T), by adding a carefully chosen small *perturbation sound* δ (imperceptible to people), the resulting audio $I+\delta$ (which is called audio adversarial example [1]) will be recognized as some other targeted transcript T' ($\neq T$) by a receiver's SR without transmissions. A natural question is to ask: will the audio adversarial example $I+\delta$ still be recognized as the targeted transcript T' after transmission over the air? In other words, can $I+\delta$ played by a sender fool the SR at the receiver? If so, consequences can be serious, since this introduces crucial cyber security risks that an attacker could hack or deploy a speaker to play malicious adversarial examples, hiding voice commands that are imperceptible to people, for launching a targeted audio adversarial attack remotely. Such malicious voice commands might cause unsafe driving (*e.g.*, fooling the voice control interface in a car), denial of services (*e.g.*, switching off sensors in cyber-physical systems), and launching spam or phishing attacks (*e.g.*, updating the phone's blacklist).

Through our study, we find that previous attacks [1], [3] fail after the over-the-air transmission is mainly because the *effective* audio signal received by SR after the transmission is $H(I+\delta)$, in stead of $I+\delta$, where $H(\cdot)$ represents the signal distortion from the acoustic channel, *e.g.*, attenuation, multi-path, *etc.*, and also the distortion from the device hardware (speaker and microphone). Due to $H(\cdot)$, the effective adversarial example may not lead to T' any more. Of course, if we can measure $H(\cdot)$ from the sender to the victim receiver, δ can be trivially pre-coded, by satisfying $SR(H(I+\delta))=T'$. However, such a measurement is not practical becomes it requires the attacker to hack the victim device in advance and then programs it to send a feedback signal conveying $H(\cdot)$. To unveil a real-world threat, the open question is whether we can find a generic and robust δ that survives at any location in space, even when the attacker may not have a chance to measure $H(\cdot)$ in advance. In the rest of this poster, we briefly introduce our design in Metamorph.

II. Design

A. Understanding Over-the-Air Audio Transmission

When an attacker initializes an over-the-air attack, the audio first goes through the transmitter's loudspeaker, then enters the air channel, and finally arrives at the victim's microphone. Overall, the adversarial audio is affected by three factors: *device distortion*, *channel effect*, and *ambient noise*.

We first setup a loudspeaker-microphone pair in an anechoic chamber (avoiding noise and multi-path) and observe that the frequency-selectivity caused by hardware is not strong and is similar to each other as shown in Figure 1(a), because the mobile devices are typically optimized for humans' hearing. The device frequency-selectivity is not extremely strong (compared with the channel's), while it can fail the previous audio adversary attack [1] already as shown in Figure 1(d).

Next, we investigate the frequency-selectivity from channel. We also conduct similar experiments in three typical indoor scenarios (an office, a corridor, and a home apartment) with varying distances (0.5 m to 8 m). Figure 1(b)–(c) shows channel frequency-selectivity is highly unpredictable over long distances (*e.g.*, 8 m) because the multi-path effect becomes more significant and environment-dependent, while more common features can be observed over short-distance transmissions (*e.g.*, 0.5 m) because LoS paths dominate the channel's effect and mainly causes attenuation in this case. However, the tightly glued device frequency-selectivity still affects.

We finally investigate the impact of the ambient noise by tuning the volume of added background noise to adversarial examples, and feeding the synthesized adversarial examples to the SR model directly. Figure 1(e) shows when SNR is reasonably large, *e.g.*, >22 dB, character success rates (CSRs) are all close to one. Because the attacker can decide when to launch the attack, the loud noise can be avoided. Therefore, we mainly focus on the frequency-selectivity introduced by the hardware and the acoustic channel in the Metamorph design.

B. "Generate-and-clean" Two-Phase Design

Above understanding inspires that (at least) within a reasonable distance before the channel frequency-selectivity dominates and causes $H(\cdot)$ to become highly unpredictable, we can focus on extracting the aggregate distortion effect from both device and channel. Once the core impact is captured,

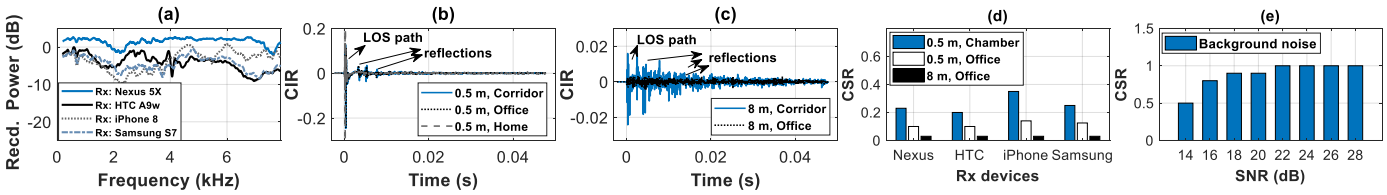


Figure 1: (a) Device frequency-selectivity curves from four receivers measured in an anechoic chamber. (b–c) Channel impulse responses measured over both short and long links in three indoor environments. (d) Character success rate (CSR) for the adversarial examples in [1] transmitted in the anechoic chamber and office. (e) CSRs achieved with different noise levels.

we can factor it into the audio adversary example generation. Therefore, we propose a "generate-and-clean" two-phase design. We first consider the major impact of these frequency selectivities by using multiple channel impulse response (CIR) measurements from different devices with different transmission distances in different environments to pre-code the impact of $H(\cdot)$ to the generation of the initial audio adversarial example. The upper part (dashed box) of Figure 2 illustrates this generation procedure. The obtained adversary example can fool SR after a short-distance transmission, *e.g.*, 1 m.

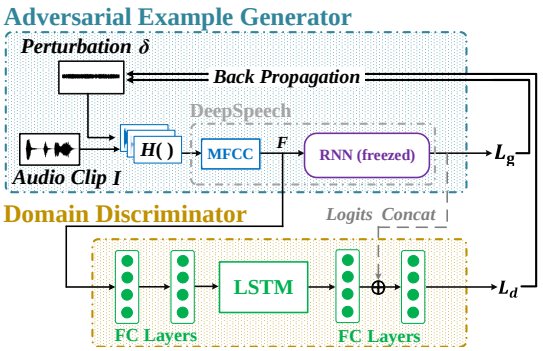


Figure 2: Illustration of the design architecture.

However, the attack still fails when the distance increases. This is because the frequency-selectivity becomes much more unpredictable and environment-dependent over long links, and the CIRs measured in advance thus become less effective. To tackle this challenge, we introduce a *domain discriminator* [4] as depicted in Figure 2 to clean the initial δ by removing the environment-related effect. The goal of the discriminator alone is to distinguish different domains (environment- and device-specific features) in the prior CIR measurements. However, the device- and environment-specific features can be removed with a proper loss function as follows:

$$L_{loss} = L_g - \beta \cdot L_d, \quad (1)$$

where L_g and L_d denote the losses of the adversary example generator and domain discriminator, respectively. In the training, the discriminator itself aims to minimize its own loss L_d. By minimizing the overall loss L_{loss} in Eqn. (1), the generator's loss L_g still gets minimized but the L_d tends to be maximized. This means that the discriminator tends to distinguish the domains incorrectly, so that the environment-dependent features can be gradually removed from the generated adversary example in Figure 2.

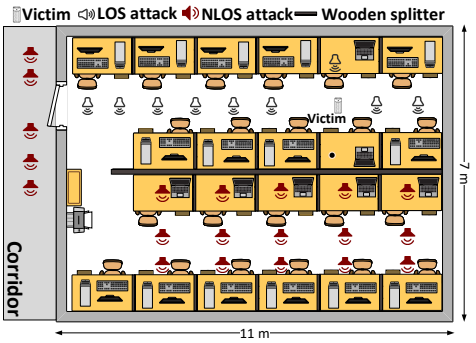


Figure 3: Floorplan used in our experiments.

C. System Performance

We evaluate the attack success rate achieved by Metamorph in a multi-path prevalent office as shown in Figure 3. We focus on a white-box attack setting and adopt DeepSpeech [2] as a concrete attack target. Through the experiment, we find that Metamorph achieves over 90% attacking success rate at the distance up to 6 m. In Metamorph, we also propose an audio quality improvement design. When this design is enabled, over 90% successful rate can be achieved up to 3 m but the audio quality can be improve significantly. On the other hand, the attacking success rate slightly drops to 85.5% on average over 11/20 none-line-of-sight locations.

III. Conclusion

In this poster, we present Metamorph that generates room-scale over-the-air acoustic adversary examples. We present our measurement studies and introduce the system design. The evaluation result shows the efficacy of the Metamorph design.

References

[1] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Deep Learning and Security workshop*, 2018.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.

[4] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *Proceedings of ICML*, 2017.

Room-Scale Over-the-Air Audio Adversarial Examples

Tao Chen[1], Longfei Shangguan[2], Zhenjiang Li[1], Kyle Jamieson[3]

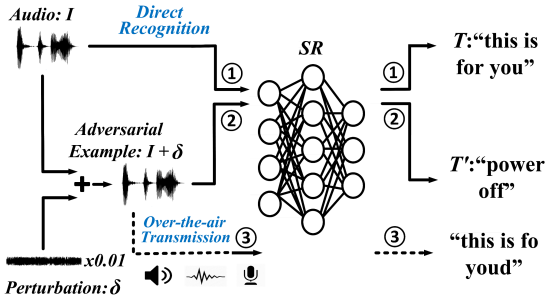
[1]City University of Hong Kong, [2]Microsoft, [3]Princeton University

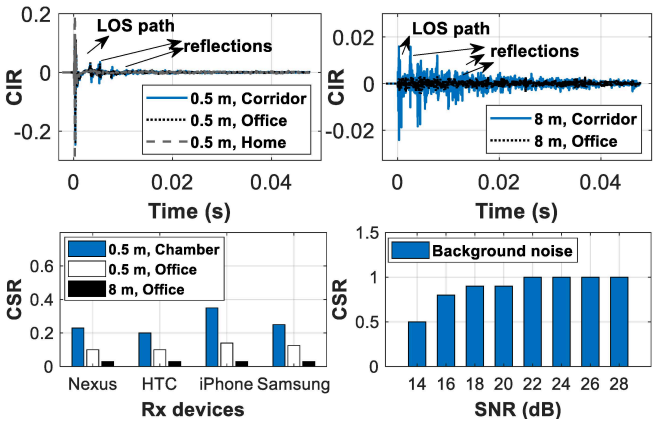
Introduction

- Audio adversarial examples could potentially attack the neural network of speech recognition (SR) systems, e.g., DeepSpeech.
- To unveil a real-world threat, one open question is whether we can find a generic and robust adversarial example that survives over-the-air at any location in the space.
- We present ***Metamorph***, a system which can generate over-the-air audio adversarial examples in the room-scale environment.

Design Considerations

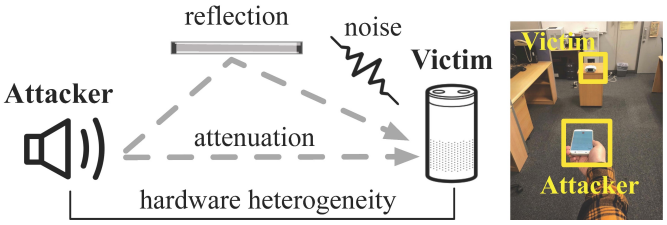
Audio adversarial examples



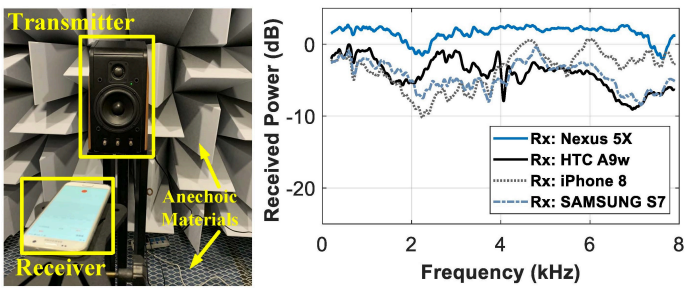


❖ **Noise:** strong noise can be avoided by attacker.

Factors challenging over-the-air attacks

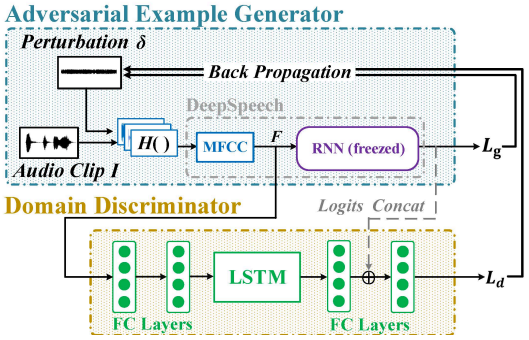


❖ **Hardware:** its frequency-selectivity is not strong.



❖ **Channel:** frequency-selectivity from channel is the main factor. However, it shows quite different features over short and long links.

Our approach: "Generate-and-Clean"



Evaluation