Metamorph: Injecting Inaudible Commands into Over-the-air Voice Controlled Systems

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 paper presents Metamorph, a system that generates imperceptible audio that can survive over-the-air transmission to attack the neural network of a speech recognition system. The key challenge stems from how to ensure the added perturbation of the original audio in advance at the sender side is immune to unknown signal distortions during the transmission process. Our empirical study reveals that signal distortion is mainly due to device and channel frequency selectivity but with different characteristics. This brings a chance to capture and further pre-code this impact to generate adversarial examples that are robust to the over-the-air transmission. We leverage this opportunity in Metamorph and obtain an initial perturbation that captures the core distortion's impact from only a small set of prior measurements, and then take advantage of a domain adaptation algorithm to refine the perturbation to further improve the attack distance and reliability. Moreover, we consider also reducing human perceptibility of the added perturbation. Evaluation achieves a high attack success rate (90%) over the attack distance of up to 6 m. Within a moderate distance, *e.g.*, 3 m, Metamorph maintains this high success rate, yet can be further adapted to largely improve the audio quality, confirmed by a human perceptibility study.**

I. Introduction

Driven by deep neural networks (DNN), speech recognition (SR) techniques are advancing rapidly [46] and are widely used as a convenient human-computer interface in many settings, such as in cars [4], on mobile platforms [3], [48], in smart homes or cyber-physical systems (*e.g.*, Amazon Echo/Alexa [1], Mycroft [7], etc.), and in online speech-to-text services (*e.g.*, SwiftScribe [10]). In general, SR converts an audio clip input I to the corresponding textual transcript T being spoken, denoted $SR(I) = T$.

In the context of the extensive research effort devoted to SR, this paper studies a crucial problem related to SR from a security perspective — given any audio clip I (with transcript T), by adding a carefully chosen small *perturbation sound* δ (imperceptible to people), will the resulting audio $I+\delta$ be recognized as some other targeted transcript T' ($\neq T$) by a receiver's SR after transmission of $I+\delta$ over the air? In other words, can $I+\delta$ (an adversarial waveform that still sounds like T to a human listener) played by a sender fool the SR neural network at the receiver?

Network and Distributed Systems Security (NDSS) Symposium 2020
23-26 February 2020, San Diego, CA, USA
ISBN 1-891562-61-4
https://dx.doi.org/10.14722/ndss.2020.23055
www.ndss-symposium.org

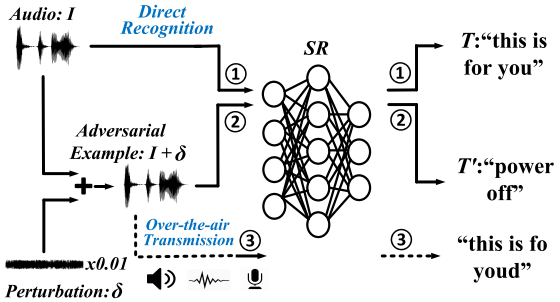

Figure 1: ① Transcript T of audio clip I is "**this is for you**". ② By adding a small δ, the adversarial example $I+\delta$ can be correctly recognized as "**power off**" without transmission [17]. This target transcript T' is selected by the attacker. ③ After over-the-air transmission, however, $I+\delta$ is no longer adversarial. Recognized transcript is similar to the original T, instead of T'.

If so, consequences are serious, since this introduces a crucial security risk 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 (*i.e.*, a T' chosen by the selection of δ). Such malicious voice commands might cause:

1) Unsafe driving. Malicious commands could be embedded into the music played by a hacked in-car speaker to fool the voice control interface and cause an unsafe driving potentially, *e.g.*, tamper the navigation path to disturb the driver's driving, suddenly change personalization settings (like volume up), *etc.*

2) Denial of service. The attacker could inject hidden commands to turn on the airplane mode of a mobile device and disables its wireless data, switch off the sensors in cyber-physical systems, *etc.*

3) Spam and phishing attacks. The attacker may delete or add appointments in the victim's calendar, update the phone blacklist or visit a phishing website on the victim device.

Recent studies [17], [46] have investigated the first step of this attack, *i.e.*, generating an adversarial example $I+\delta$ to directly fool a SR without actual over-the-air audio transmission. As Figure 1 depicts, the transcript T ("this is for you") of the input audio I can be recognized as T' ("power off") after adding a small perturbation δ. However, these works also find that the proposed technique fail after over-the-air transmission (*e.g.*, the recognized transcript becomes "this is

fo youd" instead of "power off" in Figure 1). This is because after the transmission, the *effective* audio signal received by SR is $H(I+\delta)$, where $H(\cdot)$ represents signal distortion from the acoustic channel, *e.g.*, attenuation, multi-path, *etc.*, and also distortion from the device hardware (speaker and microphone). Due to $H(\cdot)$, the effective adversarial example may not lead to T' any more. There are also follow up works [56], [57] try to compensate the channel effect by directly feeding the channel state information collected at other places into the training model. However, these proposals are far from becoming a real-world threaten primarily due to the short attacking range (*e.g.*, < 1 m) and physical presence of the attack device (*e.g.*, fail in none-line-of-sight conditions).

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, the channel measurement is not practical because it requires the attacker to hack the victim device in advance and then programs it to send a feedback signal conveying $H(\cdot)$. To create 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*.

To answer this question, we first conduct micro-benchmarks to understand how the over-the-air transmission affects acoustic adversarial attack. Our micro-benchmark results reveal that the signal distortion is mainly due to the frequency selectivity caused by both multi-path propagation and device hardware. Specifically, we first experiment in an acoustic anechoic chamber (avoiding multi-path) and find that as devices are optimized for humans' hearing, the hardware distortion on the audio signal shares many common features in the frequency domain cross devices and undermines the over-the-air adversarial attack already. In practice, the problem is naturally more challenging since the channel frequency selectivity will be further superimposed, which could become stronger and highly unpredictable as the distance increases. Although it is difficult to separate these two frequency selectivity sources and conduct precise compensation, as the multi-path effect varies over distance and the hardware distortion shares similar features cross devices, this 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. Once the core impact is captured, we can factor it into the sound signal generation.

With these considerations, we develop Metamorph with a "generate-and-clean" two-phase design. In phase one, we collect a small set of $H(\cdot)$ measurements as a prior dataset, to generate an initial δ that captures the major impact of the frequency-selectivity from these measurements (including both device and channel frequency selectivity) collected in different environments with different devices. The first phase achieves an initial success for the over-the-air attack, but this primary δ inevitably preserves some measurement-specific features, still limiting the attack performance. Therefore, in the second phase, we further leverage domain adaptation algorithms to clean δ by compensating the common device-specific feature and also minimizing the unpredictable environment dependent feature from these $H(\cdot)$ measurements to further improve the attack distance and reliability.

We finally consider the impact on audio quality of the generated adversarial example and minimize perceptibility by people with two mechanisms. First, we customize the added δ, so that the resulting noise heard is like a real-world background sound, *e.g.*, music. We call this as a "acoustic graffiti", so that the audience may believe this is part of the original audio clip. Second, we find we only need to add δ to a part of audio I that contributes most to the SR recognition, reducing the volume of perturbation bits added to I.

We include all above design elements in a prototype system named *Metamorph*. Similar to other recent attacks [17], [46], this paper also focuses on the *white-box* setting (detailed in §II-A), and we utilize the state-of-the-art speech recognition system, DeepSpeech [27] developed by Baidu, as a concrete attack target. Even with Metamorph, we believe that plenty of research opportunities remain possible in the future, while this paper already serves as a wake-up call to alarm people to the potential real-world threat from the useful and apparently non-detrimental speech recognition techniques. The key experimental results are as follows.

- Metamorph achieves over 90% attacking success rate at the distance up to 6 m (when prioritized to reliability) and 3 m (when prioritized to audio quality) in a multi-path prevalent office scenario. The attacking success rate slightly drops to 85.5% in most none-line-of-sight settings on average.
- Metamorph performs consistently for different victim receivers and is robust to the victim movement with a moderate moving speed, *e.g.*, 1 m/s.
- The user perceptibly study on 50 volunteers shows up to 99.5% imperception rate to identify any word (content) change over 2000 adversarial example instances. Adversarial samples generated by Metamorph are released in [9].

Contribution. This paper makes following contributions. We empirically understand the factors that limits prior audio adversarial attacks with the over-the-air setting. We propose a series of effective solutions to address the identified design challenges and enable the over-the-air attack in both LOS and NLOS environment. We develop a prototype system and conduct extensive real-world experiments to evaluate performance.

II. Preliminaries

A. Attack Model

The attacker's goal is to launch a targeted adversarial attack on a victim receiver, by fooling the neural network of its speech recognition system without the owner's awareness. The attacker adds a *perturbation waveform* δ to the owner's audio clip I (transcript T) to generate a voice command recognized as T' by the receiver. We consider the attack model regarding to the following aspects in the paper.

Speaker device. Attacker can either directly play or hack a deployed speaker device (*e.g.*, in-car speaker or Amazon Echo in a room) in the vicinity of the victim receiver to play the adversarial audio $I+\delta$. Because the speaker is controlled by the attacker, the frequency selectivity introduced by the transmitter device can be compensated by the training if the attacker adds some channel impulse response measures from this device, or the attacker can simply select a high-quality

device to minimize the impact from the transmitter's frequency selectivity and skip such an explicit compensation.

Perturbation δ. For each audio clip I, the generated δ only works for this audio I, not for other audio clips.

Measurement-free for audio distortion. Attacker can play any targeted sneaky commands to the victim receiver, while we do not assume that she can measure the audio signal distortion $H(\cdot)$ at the victim side, *e.g.*, no prior measurement or information is needed in advance to launch this attack, because the attacker may not be able to enter into the room or the receiver's location may change.

Victim device. Attacker can launch the attack when the receiver device is not in use by the owner, or the owner is temporarily away from the device. In addition, the attacker does not need to know the specific victim device to be used in this attack, because our design considers and compensates this diversity in the adversarial example generation.

Ambient noise. Attacker can tune the speaker volume according to the noise level around the victim device, and our current design mainly works with moderate noise levels, *e.g.*, SNR (Signal-to-Noise Ratio) is greater than 25, which is available in many indoor scenarios (*e.g.*, office or home).

Audio quality. The perturbation δ should be imperceptible to human beings. Although encoding the perturbation δ on the high-frequency band (>20 kHz) by a common speaker could be inaudible to human beings, it fails to initiate adversarial attack since the speech recognition system analyzes the voice input mainly on the audible frequency, *e.g.*, <8 kHz [27].

White-box setting. Similar as recent attacks [17], [46], we also focus on the white-box setting, assuming the awareness of the speech recognition system's particulars. Similar to recent works [17], [27], [56], we adopt DeepSpeech [8], [27] as a concrete attack target. DeepSpeech is an end-to-end speech recognition system that has been widely adopted by a bunch of voice assistant products (*e.g.*, Mycroft [7]) and online speech-to-text services (*e.g.*, SwiftScribe [10]), as a concrete target.

B. Primer on Audio Adversarial Attack

Before we elaborate the Metamorph design in §III, we first provide a brief primer on audio adversarial attack. First, to convert one audio clip I to its transcript T, there are two major steps in the speech recognition (SR) system:

- **Step one**: The audio input I is divided into short frames (*e.g.*, 20 ms) [17]. The neural network of SR then takes these frames as input and extracts the Mel-Frequency Cepstral Coefficients (MFCC) feature for each frame, based on which each frame will be recognized as one of the following **tokens** [26]: 1) English letters: 'a' to 'z'; and 2) two special characters: 'space' and a predefined token 'ε', which means "empty" corresponding to the frames without meaningful contents, *e.g.*, voiceless consonants.
- **Step two**: The recognized raw token sequence can be then reduced to the final recognized transcript, according to two Connectionist Temporal Classification (CTC) rules [17], [23]: a) merge all the consecutively duplicated tokens as one token; and b) then exclude all the ε tokens. For instance, the raw token sequence "$n\ n\ d\ \varepsilon\ \varepsilon\ s\ s\ \varepsilon\ s$" will be reduced to "$n\ d\ s\ s$".

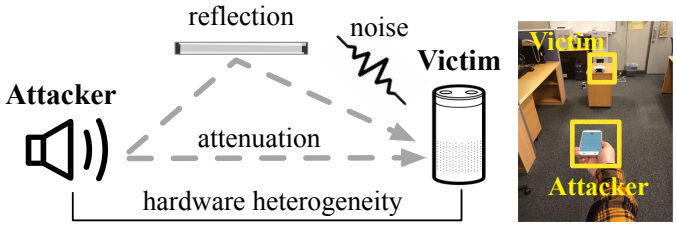

Figure 2: **An illustration of in-field audio adversarial attack**. The voice command sent from the attacker experiences distortion, attenuation, and multi-path propagation before arriving at the victim's microphone.

Formulation. With the SR principle aforementioned, the state-of-the-art adversarial attack [17] can be formulated as:

$$\begin{aligned} \text{minimize} \quad & dB_I(\delta), && (1)\\ \text{such that} \quad & SR(I) = T, && (2)\\ & SR(I+\delta) = T', && (3) \end{aligned}$$

where $T \neq T'$, T' is chosen by the attacker and $dB_I(\delta)$ is the audio sound distortion measured in Decibels (dB), *i.e.*, $dB_I(\delta) = dB(I+\delta) - dB(I)$.

Solving δ. The formulation above can be further rephrased as follows to solve the perturbation δ [17]:

$$\arg\min_\delta \ dB_I(\delta) + \alpha \cdot L(SR(I+\delta), T'), \qquad (4)$$

where $L(\cdot)$ and α are the loss function and the weighting factor, respectively. Two points are worth noting:

- As each divided short audio frame (*e.g.*, 20 ms) further contains multiple sampling points (*e.g.*, 320), the obtained δ is a set of values indicating the perturbations to be added to the amplitude of each frame's sampling points in I.
- To solve Eqn. (4), we need to know the working particulars of the target SR for computing the exact loss (*i.e.*, a white-box attack). After δ is resolved, the adversarial example $I+\delta$ can be inherently achieved [17].

With the preliminary information above, the next section reports our empirical understanding of the acoustic channel, followed by the Metamorph design.

III. Design

A. Understanding Over-the-Air Audio Transmission

When an attacker initializes an audio adversarial attack, the audio clip first goes through the transmitter's loudspeaker, then enters the air channel, and finally arrives at the victim's microphone, as shown in Figure 2. Overall, the adversarial audio clip is affected by three factors: *device distortion*, *channel effect*, and *ambient noise*. To survive the adversarial examples from the over-the-air transmission, we need to first carefully understand the effects of these three factors.

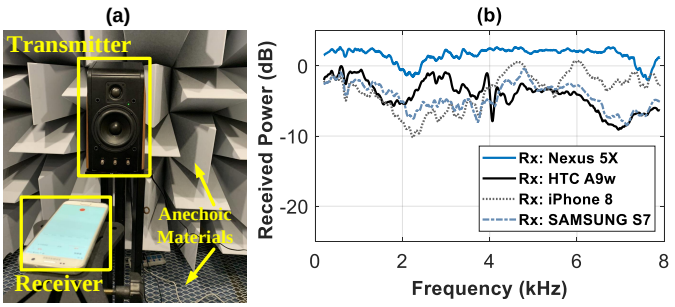

Figure 3: (a) Experiment setup in the anechoic chamber. (b) Device frequency-selectivity curves from four receivers.

1) Device Distortion: Both the attacker loudspeaker and the victim microphone introduce *frequency-selectivity*[1] to the transmitted audio signal, which can distort the audio adversarial example and undermine this attack after the over-the-air transmission. To separate the device frequency-selectivity and focus on its effect, we setup a loudspeaker-microphone pair in an anechoic chamber (avoiding noise and multi-path), as Figure 3(a) shows. In practice, the attack can be initiated on attacker's own device (loudspeaker), hence the loudspeaker can be selected with small device frequency-selectivity to avoid an explicit compensation of transmitter's hardware distortion and facilitate the attack. Thus in Figure 3(a), we use a high-end speaker HiVi M200MKIII [5] that has a relatively flat frequency response over the audible frequency band, to minimize the effect of the transmitter and focus on the receiver's (victim device) frequency-selectivity. The speaker transmits a swept sine wave [21] to multiple receivers at 0.5 m, ranging from 20Hz to 20kHz, and we cut it up to 8 kHz to analyze the frequency selectivity (SR, *e.g.*, DeepSpeech, uses this range).

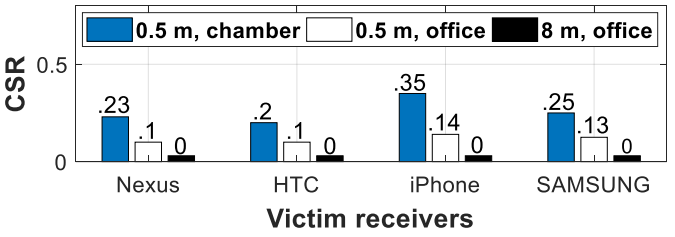

Figure 4: Character success rate (CSR) for the adversarial examples transmitted in the anechoic chamber and office.

Result. We plot the frequency response curve of each receiver in Figure 3(b). We observe that these frequency response curves exhibit a similar profile in 0–8 KHz frequency band. This is understandable since the microphone on smart devices is typically optimized for human speech, hence their frequency response should be similar to each other. However, due to the hardware heterogeneity, each curve exhibits different frequency-selectivity details. For example, we observe 6 dB frequency selectivity on 2–4 kHz frequency band for iPhone 8, while only 3 dB for SAMSUNG S7 is on the same frequency band. We further transmit the adversarial examples generated by Carlini *et al.* [17] in the chamber and observe that the device frequency-selectivity alone could fail this attack[2], *e.g.*, character success rate (CSR) is low in Figure 4 ("0.5 m, chamber"), and incorrect characters always exist in each recognized transcript from all the receivers.

[1]Frequency-selectivity refers to the non-uniform frequency response across the frequency band [38], *e.g.*, 0–8 kHz in the audible band.

[2]The attack proposed in [17] is outlined in Section II-B.

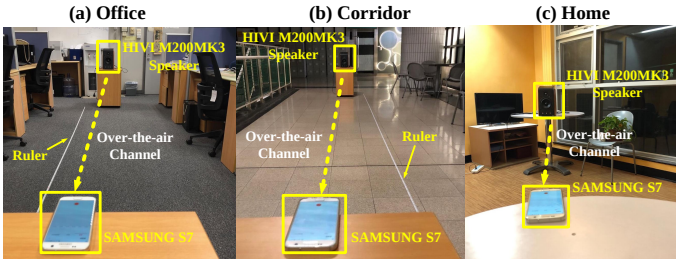

Figure 5: Tx-Rx pairs in office, corridor and home.

However, as depicted in Figure 3(b), the device frequency-selectivity overall is not extremely strong (some characters are still correct in Figure 4) and these frequency-selectivity curves share many similarities. Moreover, device frequency-selectivity is hardware's inherent feature, not related to the transmission distance. So the device frequency selectivity in principle can be measured and compensated. In fact, with a proper design (§III-B), this device effect can be implicitly considered when we deal with the acoustic channel, which also causes frequency selectivity. Since the channel's effect varies over distance, we next examine the acoustic channel.

2) Channel Effect: The impact of acoustic channel on the transmitted signal is mainly from the *attenuation* and the *multi-path* two aspects.

Attenuation. Attenuation leads to a signal strength reduction. It would not undermine the adversarial attack, because the SR system usually normalizes the amplitude of the input audio in the MFCC feature extraction [51]. In our experiment, we have also validated that when we scale the amplitude of an audio input $I+\delta$, the same transcript can be always obtained from the speech recognition system.

Multi-path. Multi-path is environment-dependent. It also introduces frequency-selectivity to the received signal due to the constructive and destructive interference [55], and may potentially impact the adversarial attack.

To understand the impact of mult-path in acoustic channels, we setup a transmitter-receiver pair (*e.g.*, M200MKIII loudspeaker sends the swept sine wave to the smart phone receiver) in three typical indoor attacking scenarios: an office, a corridor and a home apartment, as shown in Figure 5. We first look at channel state information (CSI) in these three environments and plot the result in Figure 6(a)–(b). CSI is the frequency domain response, which can unveil the frequency-selectivity directly. Ideally, CSI can be accurately obtained by $\frac{FFT(y(t))}{FFT(x(t))}$, where $x(t)$ and $y(t)$ are the transmitted and received signal, respectively. However, as the acoustic signal will go through the hardware (loudspeaker and the microphone) during transmission, the frequency selectivity from the CSI measurement is the combined one from both channel and device.

From Figure 6(a), we observe a moderate frequency selectivity in office, corridor and home environments when the receiver is in close proximity to the transmitter, *e.g.*, 0.5 m. These three CSI curves exhibit a similar frequency selectivity.

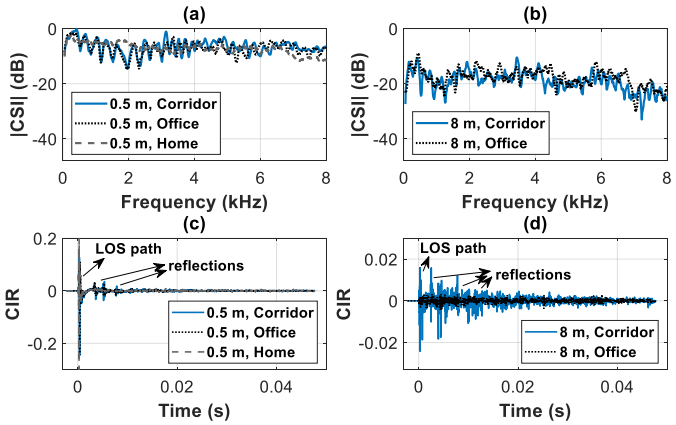

Figure 6: Frequency spectrum (a–b) and their channel impulse responses (c–d) measured over both short and long acoustic links in three typical indoor environment. We do not measure long link channel at home due to the space limit.

To better understand this result, we plot the channel impulse response (CIR[3]) of these three channels in Figure 6(c). All these three CIR curves exhibit a huge power gap between the line of sight (LOS) path and reflection paths, indicating that the LOS path dominates the signal transmission over such short acoustic links. This unequal power distribution over different paths renders the superposition of multi-path signals resemble to the LOS signal, as shown in Figure 7(a). Accordingly, the channel along would not cause significant frequency selectivity over such short links. The slight CSR declination in Figure 4 ("0.5m, office") also confirms this.

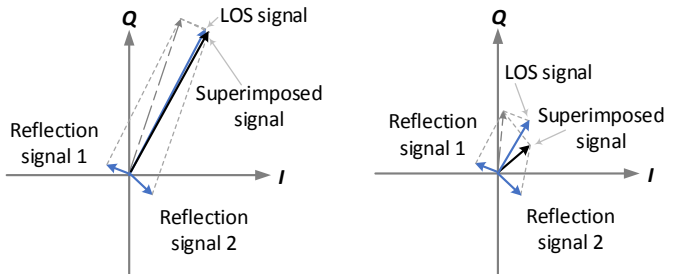

Figure 7: Superposition of multi-path signals in (a) short and (b) long acoustic link settings.

As we expand the link distance, *e.g.*, 8 m, the CSI profiles (we skip the long link setting at home due to the space limitation) exhibit a stronger and dissimilar frequency selectivity in Figure 6(b). We further plot their CIRs and observe a decreased power gap between the LOS path and reflection paths (Figure 6(d)). This result indicates that signals propagate among these paths, when adding together, would cause significant frequency selectivity due to the constructive and destructive interference, as shown in Figure 7(b). We further play the adversarial examples generated by [17] in the long acoustic link settings (8 m) and observe that these adversarial attacks never succeed in Figure 4 ("8m, office").

Observation. Above results reveal that the frequency selectivity due to channel fundamentally challenges the over-the-air audio adversarial attack. For long links, the multi-path effect becomes more significant and unpredictable (environment dependent). For short links, the multi-path effect itself may not be very strong, but the tightly glued device frequency-selectivity still affects. Fortunately, the hardware's distortion on audio signal will not change over distance and shares similar frequency selectivity features (§III-A). The key inspiration to us is hence that within a reasonable distance (before the channel frequency selectivity dominates and causes the overall signal distortion to become highly unpredictable), if we have a chance to capture the core impact of the overall distortion from both channel and device, we can pre-code it in the adversarial example generation.

Although deriving a theoretical model to describe the feasible attack distance is still open, in this paper, we demonstrate that the attacker can leverage learning algorithms to launch the over-the-air adversarial attack within a reasonably long distance, *e.g.*, 6 m, that can achieve both a high successful rate (§III-B) and a good audio quality (§III-C).

3) Ambient Noise: We finally investigate the impact of the ambient noise on the adversarial attack. We collect three types of typical background noises: ambient human voice, background music, and engine noise. We then tune the volume of these three background noises to different levels and synthesize them with the adversarial example. To avoid the frequency selectivity introduced by the device hardware and the acoustic channel, we feed these synthesized adversarial examples to the speech recognition system directly.

Result. We vary the signal-to-noise ratio (SNR) from 14 to 28 dB in Figure 8(a) and calculate the character success rate (CSR) for these three types of synthesized adversarial attacks. We observe that when the SNR is reasonably large (noise is small), *e.g.*, >26 dB (such as playing an adversarial example (76 dBSPL) in a normal human conversation (40-50 dBSPL) environment), the CSRs are all close to one for these three synthesized adversarial examples. This is reasonable since the weak noises are easily overwhelmed by the voice commands. In §IV, we also have a similar observation from the real-world attack. CSR decreases slightly as we tune up the volume of the noise (a lower SNR). In particular we find CSR with the human voice noise drops rapidly as we slightly decrease the SNR from 26 dB to 22 dB.

To understand the reason behind, we further plot the frequency spectrum of these three kinds of noises in Figure 8(b). Compared with the engine and background noises, the human voice shows more significant frequency selectivity, and thus should have a higher impact on the adversarial attack. However, as 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. Practical Audio Adversarial Examples

From the empirical study, our key insight is to cope with the frequency-selectivity introduced by both the device and channel. The device frequency-selectivity is more predictable, while the channel's impact varies over distance. However, even within a reasonable attacking distance (when the channel frequency-selectivity is moderate), it is still unfeasible

[3]CIR is similar to the concept of room impulse response (RIR) in the audio signal processing domain [13]. Both describe signal's time domain response.

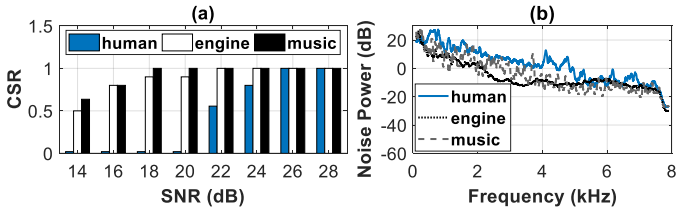

Figure 8: (a) Character success rate (CSR) in different noise levels. (b) Frequency responses of three typical noises.

to enumerate all possible frequency-selectivity curves in the adversarial example generation. Therefore, in Metamorph, we will conduct a small set of prior frequency-selectivity measurements and further leverage learning algorithms to extract the core impact from these measurements, so that we can factor it into the adversarial example generation, achieved by a "generate-and-clean" two-phase design.

- In phase one (§III-B1), we generate an initial δ that mines and considers the major impact of frequency selectivity from these measurements conducted in different environments with different devices. Of course, it may also preserve some measurement-dependent features (to minimize the optimization loss), still limiting the attack performance.
- In phase two (§III-B2), we further leverage learning algorithms to clean δ by compensating the common device-specific feature and also minimizing the unpredictable environment dependent feature from these frequency selectivity measures to further improve the attack performance.

1) Generating Initial Examples: Motivated by Expectation Over Transformation (EOT) method invented by vision-based adversarial attack [15], we introduce the following three steps to generate the initial audio adversarial examples.

Step 1. When we transmit the swept sine wave and receive it over the air, the derived channel impulse response (CIR) includes the frequency-selectivity from both device and channel. Therefore, we can collect multiple (M) such measurements from M sender-receiver transmission pairs with different distances in arbitrary environments. To simplify this measurement process and include more device heterogeneity, we can directly leverage some public acoustic CIR datasets. We utilize four such datasets, including AIR [28], MARDY [53], REVERB [32] and RWCP [37], and adopt M as 370 in our current design (the description of these datasets and our configuration is in §IV).

Step 2. Next we train δ by minimizing the following optimization, subjected to M constraints $SR(H_i(I+\delta)) = T'$, where $i = 0,\ldots,M$. Mathematically, δ can be obtained by extending the formulation in Eqn. (4) to:

$$\begin{aligned} & \arg\min_{\delta} \ \alpha \cdot dB_I(\delta) + L_{ctc}, \qquad (5) \\ = \ & \arg\min_{\delta} \ \alpha \cdot dB_I(\delta) + \frac{1}{M}\sum\nolimits_i L(SR(H_i(I+\delta)), T'), \end{aligned}$$

where $dB_I(\delta)$ is the sound quality distortion in dB and $L(\cdot)$ in the second line of Eqn. (5) is the CTC loss [23] to quantify the difference between the target transcript T' and SR's recognition result by taking $H_i(I+\delta)$ as input. In Eqn. (5), the hyper-parameter α trades off the audio quality and attack success.

The upper part (dashed box) of Figure 9 illustrates this audio adversarial example generation procedure. The original

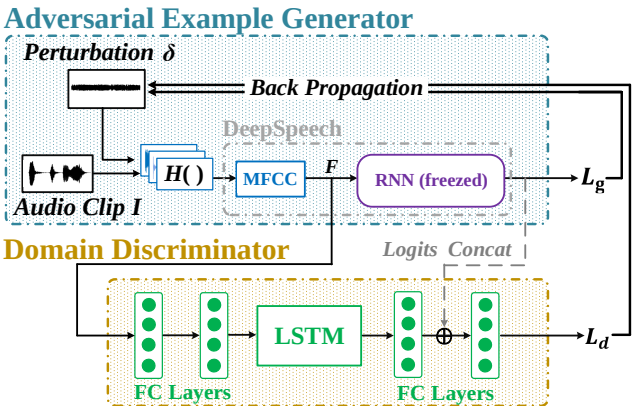

Figure 9: Illustration of initial adversarial example generator and the domain discriminator, where L_g represents all the loss factors except the loss L_d from domain discriminator.

audio clip I and the perturbation δ (which is the variable to be optimized) are processed by the M measurements of $H_i(\cdot)$. The resulting audios $H_i(I+\delta)$ are then passed to the neural network of our attacking target DeepSpeech. DeepSpeech will first extract the MFCC feature of each audio input $H_i(I+\delta)$, denoted as F_i, based on which its recurrent neural network (RNN) can recognize the transcript T_i for the current input $H_i(I+\delta)$. As stated in Eqn. (5), the loss function here is the CTC loss L_{ctc}, which quantifies the distance between the target transcript T' and T_i, and the optimization of δ aims to minimize the overall CTC loss cross all M audio $H_i(I+\delta)$ inputs. Note that in this process, only δ will be trained and DeepSpeech already has a fixed neural network. We use it for the calculation of the CTC loss merely.

Step 3. After step 2, the composed audio $I+\delta$ is not only an adversarial example. The obtained δ already considers the future impact from the frequency-selectivity due to the transmission. We can then play $I+\delta$ over the air to fool the receiver's SR at the new locations.

Result. With the primary design above, the generated adversarial example has pre-coded the impact from frequency-selectivity, it can thus potentially fool SR after the transmission. Figure 10(a) shows an encouraging result. TSR measures the success rate of the *entire* transcript and we can see that the adversarial examples generated by this initial design now can survive after short-range over-the-air transmissions, *e.g.*, $<$ 1m.

However, TSR rapidly drops when the distance increases. This is because the received signal suffers from frequency selectivity that varies over different channels, while the limited CIR datasets used for training fail to cover all channel conditions. To better understand the performance achieved by the initial design, in Figure 10(b), we also plot the success rate of the recognized characters in the target transcript. Result shows that when TSR dramatically decreases as the distance varies from 1 m to 2 m, the character success rate (CSR) remains relatively high, *e.g.*, 0.9. Even the distance is 4 m, CSR is still above 0.5, which indicates that most characters can survive from the over-the-air transmission. However, due to the more severe frequency selectivity over longer distances, more characters in T' fail to be recognized.

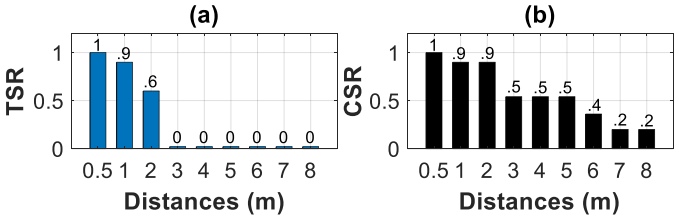

Figure 10: Performance of initial adversarial example generation. (a) Transcript success rate (TSR) and (b) Character success rate (CSR) in different attack distances. Our final design further extends the effective attacking distance to 6 m.

2) Enhancing Adversarial Examples: As the perturbation δ obtained from the initial adversarial example generation inevitably contain some device- and environment-specific features from the M channel measurements (to minimize the optimization loss), its performance will be limited at new locations, especially when the attacking distance is long and the multi-path's impact becomes stronger. To alleviate this issue, we plan to clean the initial δ by excluding its embraced measurement-specific features. After this operation, a more generic and robust perturbation δ can be obtained, which can improve the attacking distance and accuracy at new locations.

Inspired by the huge success in domain adaptation techniques [22] for object detection [39], semantic segmentation [61] and person re-identification [24], we introduce a *domain discriminator* as depicted in Figure 9 to clean the initial δ. The term "domain" here refers to the acoustic signal transmissions using different devices and settings (distances and environments). The goal of the discriminator alone is to distinguish different domains in the M prior measures. However, with a proper loss function design (below), the device- and environment-specific features can be further removed.

Domain discriminator. To design the domain discriminator, we classify the M measurements into 21 different environments, according to different transmission distances (with the one-meter step size), different rooms used in these measurements and different devices (different datasets use different devices). The discriminator then takes the MFCC feature vector F as input in Figure 9 to recognize these domains.

In particular, the MFCC feature vector F is first processed by two fully-connected layers of the discriminator to extract the measurement-dependent features. Since the audio file is a temporal sequence, the extracted features will be then processed by a RNN module, *e.g.*, Long Short-Term Memory (LSTM). To further ensure the recognition of both the initial adversarial example generator and the domain discriminator, as suggested by [29], [60], the feature vector (before the loss calculation in the generator) can be integrated into the discriminator. Therefore, we apply this integration after the LSTM in Figure 9. After the integration, we insert one more fully-connected layer to extract their overall feature before a soft-max for the domain recognition.

Loss function. We denote the loss function of the discriminator as L_d. With the discriminator, our goal can be achieved by minimizing the following integrated loss:

$$L_{loss} = L_{ctc} - \beta \cdot L_d, \quad (6)$$

where β is the weighting factor for L_d, which is configured in §IV-A. The goal of the discriminator itself aims to minimize L_d. But as L_{ctc} and L_d are connected by minus, by minimizing L_{loss}, we essentially

- 1) minimize the loss of the adversarial example generator, *i.e.*, the adversarial example is still functionable.
- 2) try the best to "cheat" the discriminator to maximize its loss L_d and make it tend to distinguish the domains incorrectly, so that the measurement specific features can be gradually removed from the MFCC feature vector, by adjusting the perturbation δ.

Improving loss to alleviate over-fitting. With the integrated loss function defined in Eqn. (6), we find that the loss function can be further improved with the following observation.

For those primary adversarial examples that are failed to be recognized as the targeted transcript T' in Figure 10, we compare all the intermediate results inside DeepSpeech when we convert $I+\delta$ and $H(I+\delta)$ to their corresponding transcripts before and after the transmission, respectively. We observe that for many characters c_j that did not survive after the transmission, the likelihood (calculated by SR) to recognize their corresponding CTC tokens (*i.e.*, English letters, space or the special token ε stated in §II-B) is high before the transmission, *e.g.*, 0.9, but this likelihood becomes very small at the receiver side after the transmission, *e.g.*, reduced to 0.1, so that another (incorrect) character token with a higher likelihood is selected in the recognized transcript.

This phenomenon suggests that the primary adversarial examples are not reliable enough, and the significant confidence reduction is likely an occurrence of over-fitting in δ for these inaccurately recognized characters. To address this issue, we can further improve the loss function in Eqn. (6), by adding a term L_{of} to alleviate the over-fitting [29]. The key idea is to introduce certain (N) "noises", so that before and after adding these noises, the recognized CTC token sequences, denoted as s and s^n respectively, should be similar (otherwise it is likely an over-fitting). Its similarity can be measured by

$$L_{of} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{n=1}^{N} \mathrm{JSD}(s_i||s_i^n), \quad (7)$$

where $\mathrm{JSD}(\cdot)$ is the Jensen-Shannon divergence [29]. Putting them all together, the improved integrated loss function is

$$L_{loss} = L_{ctc} + \gamma \cdot L_{of} - \beta \cdot L_d, \quad (8)$$

based on which robust adversarial examples can be generated. As shown in §IV, transcript success rate after enhancement can be .95 when the attack distance is even up to 6 m.

C. Improving Audio Quality

With the practical audio adversarial example generated in §III-B that can survive from the over-the-air transmission, in this subsection, we further consider its audio quality. In particular, we propose two mechanisms to minimize the perception of the added perturbation δ by human's ear. First, we propose to customize the perturbation shape, so that it sounds more similar as some real-world sound, *e.g.*, bird's chirp. We name it as a "acoustic graffiti". With this design, the audience may believe that the added perturbation is a part of the original

audio clip (§III-C1). Second, we find that we only need to train δ for covering a part of the original audio clip I (in the time domain), which could further reduce the percentage of contents in I to be modified by δ (§III-C2).

1) Acoustic Graffiti: To alleviate the perception of the target command information (which might be leaked by the added perturbation δ), we propose to customize (or reshape) the added perturbation, so that it sounds similar as some real-world background noise. In particular, the attacker can visit the nearby environment of the victim receiver, identify the noises that could appear in this environment, and then record them. If the on-site visit is not possible, the attacker can instead select any other audio template that would not raise the victim's concern, such as the soft music, the source audio itself, general ambient sounds (traffic sound for example), *etc.*

For one selected acoustic graffiti template, the attacker first normalizes the amplitude of both the perturbation δ and the template audio (scaling them to the same unit) and then computes the loss introduced by the shape difference between the perturbation and the template audio $\hat{N}$. The optimization loss will be updated as follows:

$$L_{loss} = (L_{ctc} + \gamma \cdot L_{of} - \beta \cdot L_d) + \eta \cdot dist(\delta, \hat{N}), \tag{9}$$

where $dist(\cdot)$ measures the MFCC difference between δ and $\hat{N}$. With this updated loss, δ is customized to be similar as the acoustic graffiti template.

2) Reducing Perturbation's Coverage: As stated in §II-B, the audio clip I is divided into frames (*e.g.*, 20 ms) by SR for processing and each frame contains multiple sampling points (*e.g.*, 320), the perturbation δ essentially alters (increases or decreases) the amplitude of each sampling point. In the formulation to train δ in Eqns (1)-(3), the objective is to minimize the sampling point's amplitude changing to ensure a good audio quality. Next by referring to the selected graffiti template, the perturbation then sounds more like an acoustic graffiti. In this section, we find we can reduce the amount of frame sampling points to be altered by δ, *i.e.*, coverage of δ, to further improve the audio quality.

To recognize one audio clip I as the corresponding transcript T by SR, different frames usually have a different importance in this recognition [20], [25]. However, during the training of δ, it is unclear which frame sampling points from $I+\delta$ could contribute more to the recognition of the target transcript T' in advance, since δ keeps being updated in the training. To overcome this issue, we add an L_2 regularization in the loss function to punish perturbation amplitude [20], With this L_2 regularization term, the perturbation value can maintain to be small. We can thus treat such very small perturbation values as 0 and their corresponding frame sampling points in I will not be altered. With L_2 regularization and graffiti template, the attacker can finally train δ again by:

$$\begin{aligned} \arg\min_{\delta} \ & \alpha \cdot dB_I(\delta) + L_{ctc} \\ & + \gamma \cdot L_{of} - \beta \cdot L_d + \eta \cdot dist(\delta, \hat{N}) + \mu \cdot L_2, \end{aligned} \tag{10}$$

where μ is the weighting factor for L_2, which is configured in §IV-A. For the δ obtained from Eqn. (10), we can define a perturbation coverage mask $C = \{C_f\}$, where f is the sampling point index, as follows:

$$C_f = \begin{cases} 1, & \text{if s} < \delta_f, \\ 0, & \text{otherwise,} \end{cases}$$

where s is the threshold to determine whether δ for each sampling point f is small enough, *e.g.*, $s = 20$ in the amplitude's representation range from -2^{15} to 2^{15} (int-16). Thus, $C \cdot \delta$ will ignore those very small perturbation values and thus reduce the δ's coverage. Figure 11(a) depicts one δ obtained from §III-B. When Eqn. (10) is adopted, the resulting δ is shown in Figure 11(b), and we can see that many perturbation values in δ are very small. By applying the mask C, we can obtain the masked $C \cdot \delta$ as the final perturbation.

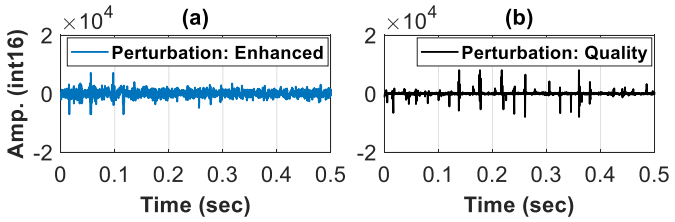

Figure 11: Perturbations trained by (a) the enhanced adversarial example generation in §III-B2 and (b) further with the improved audio quality in §III-C.

IV. Evaluation

In this section, we first introduce the evaluation setup, including data collection and training, hardware and software, evaluation metrics, parameter settings and comparison methods. We then present field studies, which comprehensively evaluate both the attack success rate and audio quality in both line-of-sight (LOS) and none-line-of-sight (NLOS) settings. We finally describe micro-benchmark results in terms of hardware diversity, ambient noise, victim movement, *etc.*

A. Experiment Setup

1) Data Collection and Training: To demonstrate Metamorph could generate over-the-air adversarial examples with a small set of prior $H(\cdot)$ measurements, we only use 370 channel impulse response (CIR) measures from four public acoustic CIR dataset (AIR [28], MARDY [53], REVERB [32] and RWCP [37]) for the perturbation generation. No CIRs are collected from our experimental environment directly. These four CIR datasets are recorded in different rooms (*e.g.*, anechoic chamber, lecture and meeting room, stairway, corridor, church.) with various link distance (0–3 m). Our selected 370 CIRs cover 21 different environments[4]. With this setting, we observe that using these CIR traces can achieve a good attack performance already and also lead to a reasonable computation overhead as stated below.

Metamorph is implemented using tensorflow 1.8.0 [11] and trained by Adam optimizer [31], together with a our proposed domain discriminator, on a high-end server equipped with two

[4]When future research studies employ our approach, they do not need to design the domain discriminator specifically for their anticipated environments neither. If the domain discriminator needs to be more generic, they can further include additional CIR traces covering more environments, *e.g.*, these datasets contain over 50 different environments in total.

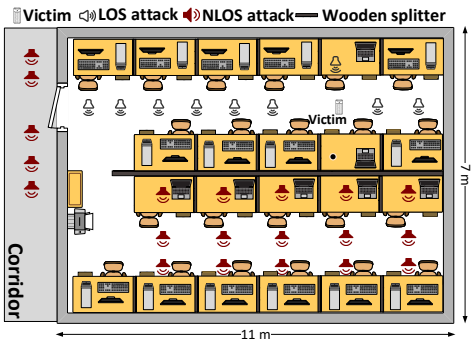

Figure 12: Floorplan of the field study. We initiate both LOS and NLOS adversarial attacks in an office building.

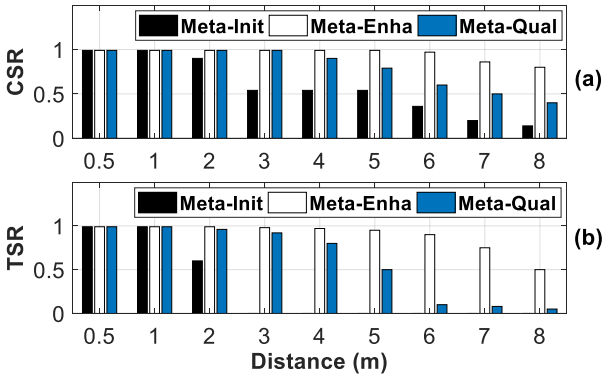

Figure 13: Performance of LOS attack by three comparison schemes. (a) Character successful rate (CSR) and (b) Transcript successful rate (TSR) in different attack distances.

NVIDIA GTX 1080Ti GPU and 32GB RAM. The training time of an adversarial example depends on the length of this adversarial audio clip. For example, generating a 6-second adversarial example takes around five to seven hours on a single NVIDIA GTX 1080Ti GPU, respectively. The training process in the future can be accelerated when more GPUs can be used in parallel. We then conduct trace-driven evaluations to quantify the system performance. In particular, we initiate the adversarial attack using different receivers (including a Google Nexus 5X, Samsung Galaxy S7, HTC A9W and iPhone 8) and one default transmitter (HiVi M200MKIII [5]) across 29 different locations, as shown in Figure 12. At each location, we play each adversarial example 100 times. The receiver records the received adversarial examples and feeds them into the targeting neural network for speech recognition, *i.e.*, DeepSpeech. We then evaluate using following metrics.

2) Metrics: Our experiments primarily rely on the following three metrics to evaluate Metamorph's performance:

- **Character success rate (CSR)** is defined as the ratio of **characters** being successfully interpreted to the total number of characters conveyed by the adversarial example.
- **Transcript success rate (TSR)** is defined as the ratio of **transcripts** being successfully interpreted to the total number of transcript conveyed by all the adversarial examples.
- **Mel Cepstral Distortion (MCD)** [19] measures the sound quality by comparing the distance between the target sound (the encoded audio adversarial) and the reference sound (the original sound). MCD is calculated by: $MCD = (10/ln(10)) \cdot \sqrt{2 \cdot \sum_{i=1}^{24} (mc_i^t - mc_i^e)^2}$, where mc_i^t and mc_i^e denote target and the estimated MCD, respectively. Lower MCD indicates better sound quality.

3) Comparison Schemes: We evaluate following schemes:

- **Meta-Init** is the initial version of Metamorph (§III-B).
- **Meta-Enha** is the domain discriminator-based version of Metamorph (§III-B2). It minimizes the effects of the device- and environment-specific features from perturbation to improve the attack distance and reliability.
- **Meta-Qual** represents the audio quality improved version of Metamorph (§III-C).

4) System Configurations: Metamorph contains several parameters. According to our detailed investigation in Appendix, we adopt the default β, γ, η and μ from the final loss function in Eqn. (10) as 0.05, 500, 1e-4, 1e-12 respectively in the experiments. On the other hand, the ratio of the characters being en-coded into the source audio to the total number of source audio frames, defined as frame utilize rate (FUR), is set to be less than or equal to 0.2 by default (Section IV-C).

We generate two types of adversarial examples (music and speech) with different source and target transcripts, detailed in Table 5 (Appendix). The source musics are labelled in Table 5 directly, and the speech adversarial examples are generated based on 11 different speech samples from the public Mozilla Common Voice Dataset [6]. For each adversarial example, we generate three versions using three comparison schemes.

B. Field Study

1) LOS Attack: We first initiate adversarial attacks at different locations that all have a clear LOS path to the victim microphone. Figure 13 shows the averaged TSR and CSR achieved by three versions of Metamorph in different link distance settings. We divide the link distance into three categories: short-range (0.5–1 m), mid-range (2–6 m), and long-range (6–8 m).

CSR performance. We observe that the initial version Meta-Init achieves nearly 100% CSR in short range settings. As we expand the link distance to the mid-range settings, the multi-path effect grows. Since the initial version has limited robustness to the multi-path effect, we thus see that CSR drops significantly to around 50%. As we further increase the attack distance to 7 m and 8 m (long-range), Meta-Init rarely succeeds, with a CSR of only around 20%.

In contrast, since the enhanced version Meta-Enha leverages the domain discriminator to minimize the channel effect, we can see its CSR remains in a constantly high level (around 100%) over both short and middle range link settings. CSR performance then drops to around 80% in long link distance settings. This result demonstrates the effectiveness of our domain discriminator-based "cleaning" design.

The CSR performance of the audio quality improved version Meta-Qual is higher than the initial method and lower than the enhanced one. Its CSR value is constantly high when the link is shorter than 3 m.

As we expand the link distance further, the CSR performance of Meta-Qual drops, yet it is still higher than that of

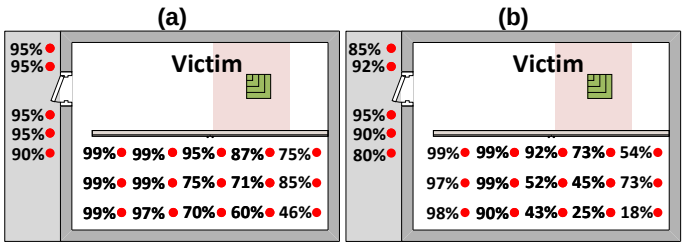

Figure 14: (a) CSR and (b) TSR of the NLOS attack for the enhanced method **Meta-Ehan** at different locations. The noise floor at the victim microphone is around 35 dBSPL.

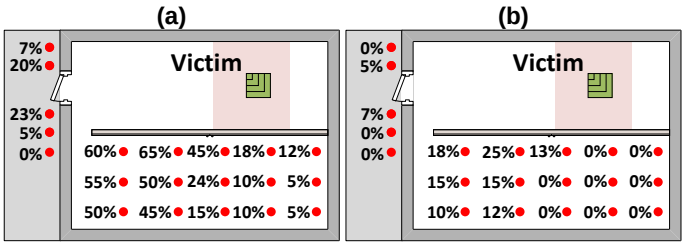

Figure 15: (a) CSR and (b) TSR of the NLOS attack for the quality improved method **Meta-Qual** at different locations. The noise floor at the victim microphone is around 35 dBSPL.

the initial method. The CSR drop is mainly due to the reduced perturbation coverage (§III-C). The resulting perturbation in Meta-Qual becomes sparse and concentrates on modifying only a part of the original audio I, which actually trades off the performance of success rate and reliability for audio quality.

TSR performance. we further calculate the transcript success rate (TSR) for each attack method and plot the result in Figure 13(b). TSR shares a similar trend with CSR yet with different success rate details. Specifically, we find that the adversarial transcripts are all successfully interpreted when the attacker loudspeaker is within the short range of the victim microphone. As we expand the link distance to 2 m, TSR of Meta-Init drops to around 60%. The initial method never succeeds as we expand the link distance further.

In contrast, the enhanced method never fails within the 5 m attacking range. As we expand the attacking range to 6 m and further to 8 m, TSR of the enhanced method drops to 90% and then 50%. The Meta-Qual method, on the other hand, succeeds over 90% within the 3 m attacking range, which is better than the initial method. TSR of Meta-Qual then drops to 50% and then 5% as we expand the attacking range to 5 m and further to 8 m.

2) NLOS Attack: We next evaluate the performance of Metamorph in the NLOS conditions. Launching adversarial attack in the NLOS environment is more challenging as the blocking materials not only attenuates the acoustic signals but also introduce frequency selectivity due to the non-uniform distribution of blocking materials. In this experiment we launch attacks with the adversarial examples trained by Meta-Enha and Meta-Qual. Figure 14(a-b) shows CSR and TSR of Meta-Enha across different locations. To imitate the real attack where the attacker is unaware of the exact location of the victim device, we place the attacker speaker facing towards the blockage (*e.g.*, the wall or the wooden splitter) across all testing locations. The victim microphone, on the other hand, is facing towards the wall on the left throughout the experiment. When we move speaker in the room, the facing direction (angle) between two devices varies from about 45° to 135°.

Figure 14 shows that Metamorph achieves consistently high CSR across the majority of attacking locations. CSR drops to 46% at the corner of this office building, primarily due to the severe multi-path introduced by walls, tables, and monitors nearby. TSR also shows a similar trend with different success rate details. We observe Meta-Enha achieves over 85% TSR across 11/20 NLOS attacking locations. In particular, we find that attacker could initiate the attack with a consistently high TSR in the corridor. This result demonstrates the efficiency of our domain adaptation algorithm and practicality of our adversarial attack. As the victim microphone is facing towards the wall on the left throughout the experiment, the performance is generally better when the speaker is placed to the left-hand side of the room. Moreover, we observe both CSR and TSR are relatively low of Meta-Qual in Figure 15(a-b) and suggest to launch Meta-Enha in NLOS attacks.

3) Audio Quality: In this experiment we quantify the audio quality of adversarial examples generated by different methods using the MCD metric (introduced in §IV-A). A lower MCD value indicates a higher similarity between the adversarial example and the original audio. We find Metamorph has different audio quality behaviors with the music (M) and human speech (S) as audio source. Hence, we plot the MCDs of Meta-Enha and Meta-Qual separately to achieve a more comprehensive view.

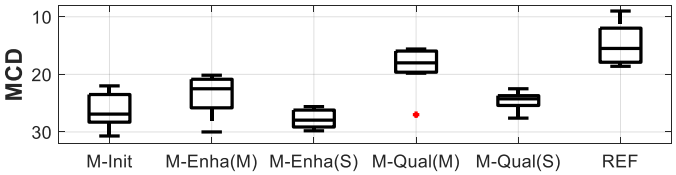

Figure 16: MCD values achieved by different methods (with a reversed y-axis representation).

MCD comparison. Figure 16 shows MCD comparison result, wherein we also plot the MCD of the adversarial example generated by Carlini *et al.* [2] as a reference (REF). From the figure, we can see that REF achieves the lowest MCD value (15.5) on average, followed by Meta-Qual(M) (18), Meta-Enha(M) (22.5), Meta-Qual(S) (24.2), Meta-Enha(S) (27.9), and Meta-Init (27). Meta-Enha(S) achieves the highest MCD (27.9). Meta-Qual achieves lower MCDs (better audio quality) than Meta-Enha for each type of the audio files, indicating that our proposed mechanisms in §III-C improve the audio quality successfully. On the other hand, the adversarial examples generated from musics outperform those generated from speeches in general, because music files usually have a higher power than the speech files, while their added perturbations have similar amplitude levels. Therefore, music-based adversarial examples could have a higher SNR to achieve a lower MCD.

User perceptibility study. While the above objective MCD measurement justifies the similarity between the adversarial example and the original source audio, these MCD values fail to reflect the subjective opinions from human beings, *e.g.*,

	No Diff.	Word change		Audio quality level				Reason description		
		Yes	*No*	*1*	*2*	*3*	*4*	*A*	*B*	*C*
M-Enha(M) (%)	9.7	0.8	99.2	49.5	43.6	6.9	0	89.9	7.5	2.6
M-Enha(S) (%)	2.0	1.9	98.1	14.5	64.7	20.0	0.8	88.1	8.9	3.0
M-Qual(M) (%)	12.3	0.5	99.5	55.0	40.8	4.2	0	90.5	7.5	2.0
M-Qual(S) (%)	9.7	0.7	99.3	47.2	47.3	5.5	0	91.3	5.6	3.1

TABLE 1: Results of the first trial of the user perceptibility study. The meaning of each option is explained in Table 4 (Appendix).

whether the adversarial examples can be easily perceived by humans. We thus invite 50 volunteers (20 females and 30 males) with diverse ages ranging from 18 to 50 and conduct a perceptibility evaluation of the audio quality. These volunteers are non-paid for this study and have no hearing disease. We utilize the adversarial examples listed in Table 5 (Appendix) to conduct the following two trials of user perceptibility studies.[5]

a) In the first trial, volunteers will sequentially listen to each set of audios organized as follows: "[(one original audio, the adversarial example generated from this audio by **Meta-Enha**), *60s pause*, (the same original audio, the adversarial example generated from this audio by **Meta-Qual**), *60s pause*]". In each 60s pause, volunteers assess the audio quality of the adversarial example (they just heard) compared with the original audio by answering the following four questions.

Volunteers first select whether this adversarial example has a same audio quality as the original audio, including both the noise level and the audio content. If the answer is *Yes* (*i.e.*, no difference), the assessment of this adversarial example is complete for the first trial; Otherwise, volunteers will further answer the following questions for this adversarial example.

- **Word change (*Yes* or *No*)**: any word (content) change is perceived compared to the original audio's transcript?
- **Audio quality level**: we have provided four options (1–4) reflecting different audio quality levels for volunteers to rate.
- **Description**: we have also provided two options (A and B) to describe how or where, the volunteers think, such noises come from. If none of them fits, they can also select "C (Others)" and describe using their own words.

Due to the page limit, the explanations of above three questions and the options are detailed in Table 4 (Appendix). Table 1 summarizes the results. Although nearly 90% adversarial examples are thought not exactly the same as their original audios, among those examples, 98.1% to 99.5% of them do not cause the hearing of any content (word) change to the volunteers, *i.e.*, the heard content is still the original transcript. In terms of the audio quality level, 64.7% of the adversarial examples from Meta-Enha(S) are rated to be slightly loud (level 2), and even 20% of them cause the missed hearing of certain audio content occasionally due to noise. However, around 47.2% to 55.0% adversarial examples are rated to be clear (level 1) for Meta-Enha(M), as well as for both Meta-Qual(M) and (S), implying the effectiveness of our Meta-Qual design on improving the audio quality. In the description field, for 88.1% to 91.3% of the adversarial examples, volunteers feel that the noises are coming from the hardware (*e.g.*, recording microphones, cheap speakers) (Option A). For 5.6% to 8.9% of them, they feel that it is due to the low-quality of audio clip itself (Option B). For the rest 3%, volunteers describe like "mixture of options A and B", "sound dithering from the old tape recorder", "buzzing effect", *etc.*

b) After a 10-minute rest, volunteers start the second trial. At the beginning of this trial, we play original audios one more time to refresh the volunteers' impressions on the audio quality of these original audios. Then we play audio clips (either an original audio or an adversarial example) in a random order, and volunteers are not aware they are about to hear an original audio or an adversarial one each time. After hearing each played audio, volunteers need to decide whether this played audio is an original audio clip immediately.

	M-Enha(M)	M-Enha(S)	M-Qual(M)	M-Qual(S)	Original audio
Ratio (%)	36.7	19.5	42.4	39.4	88.9

TABLE 2: Results from the second trial of the user study.

Table 2 summarizes the result. We can see with this experimental setting, even about 10% original audios are recognized incorrectly. For M-Enha(S), a small portion of adversarial examples are recognized as the original audios, while the music audios can increase this ratio to 36.7%. With our audio quality improvement design, M-Qual(M) and M-Qual(S) can further improve the ratio to 42.4% and 39.4%, respectively.

Conclusion. According to this field study, we conclude that within a moderate attack distance (*e.g.*, 3 m), Meta-Qual can be firstly considered. For the long links, Meta-Enha(M) is prioritized than Meta-Enha(S), if the music source can be selected in the attack.

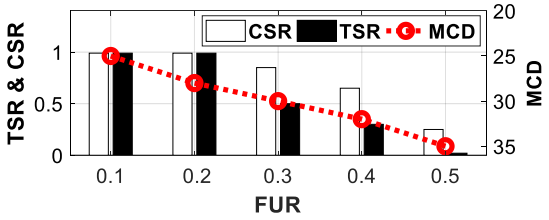

Figure 17: TSR, CSR, and MCD in different FUR settings (with a reversed y-axis representation for MCD).

C. Micro-Benchmarks

We next conduct micro-benchmarks to understand the effect of each designing factors on Metamorph's performance. Suggested by the field study, Meta-Enha and Meta-Qual can achieve effective attacking results (*e.g.*, abover 90% TSRs) at distances of five and three meters, respectively. We thus adopt these link distances in the micro-benchmarks.

[5]The questions in our user study do not involve any confidential information about volunteers, which will not cause them any potential risks (psychologically, physically, socially, etc.). The study obtains university's ethical approval.

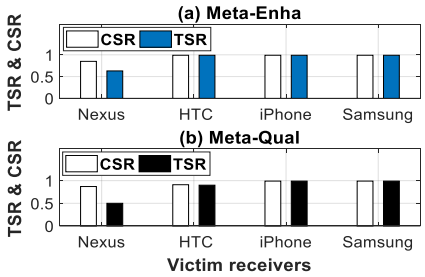

Figure 18: CSR and TSR achieved by (a) Meta-Enha and (b) Meta-Qual across different victim receivers.

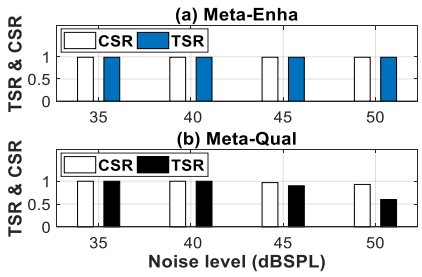

Figure 19: CSR and TSR achieved by (a) Meta-Enha and (b) Meta-Qual in different noise floor settings.

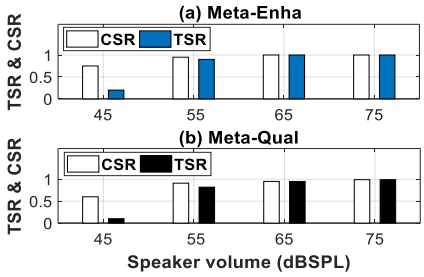

Figure 20: CSR and TSR achieved by (a) Meta-Enha and (b) Meta-Qual in various speaker volume.

1) Effect of Transcript Length: Given a source audio, the audio quality degrades with the growth of adversarial transcripts that being inserted in this audio source. In this experiment, we define a new metric *frame utilize rate* (FUR) as the ratio of characters being en-coded into the source audio to the total number of source audio frames. The experiment setup is same as the previous one. Figure 17 shows the result. As expected, audio quality decreases (MCD increases) with the growth of FUR. On the other hand, we also witness a decreasing trend of both TSR and CSR as we increase the FUR from 0.1 to 0.5. This is understandable as a larger FUR value indicates more adversarial characters are en-coded into the source audio, hence more characters are prone to errors. Suggested by this result, we set the maximum FUR to 0.2 by default in the current Metamorph.

2) Effect of Device Frequency Selectivity: We first examine whether the attack performance of Metamorph is insensitive to different types of victim devices. We setup a five-meter (for Meta-Enha) and three-meter (for Meta-Qual) acoustic link to launch the attack. We fix the transmitter and then exchange the receivers to examine the corresponding CSR and TSR. Figure 18 shows the TSR and CSR achieved by Meta-Qual and Meta-Enha across four type of receivers. We observe that the high-end iPhone and Samsung smartphone achieve consistently high TSR and CSR, which are both around 100%. CSR and TSR of HTC smartphone (less expensive) drops gradually to around 90%. While the CSR of Nexus (cheapest one among four testing phones) maintains in a reasonable level (80%), we witness a significant TSR drop (50%) on it, probably due to the inferior hardware components used in this smartphone.

TSR of Nexus then grows from 50% to around 65% when we use Meta-Enha method to train the adversarial phrases. We also observe that TSR of HTC smartphone even jumps to around 100% in the same setting. On the other hand, both iPhone and Samsung smartphone maintains a consistently high TSR and CSR. The result demonstrates that Metamorph achieves overall satisfying robustness to the middle-end and high-end smartphone. Its performance degrades when using low-end smartphone, and we leave the way to compensate for that as our future work.

3) Effect of Ambient Noise: We next examine the effect of ambient noise. The experiment setup is same as the previous one. The attacker speaker plays the adversarial examples at 75 dBSPL. We further play another music clip as a background noise and examine system performance under different noise levels from 35 dBSPL to 50 dBSPL, *e.g.*, the corresponding SNR varies from 40 to 25. From Figure 19 we observe that Meta-Qual achieves consistently high TSR and CSR in 35 dBSPL and 40 dBSPL noise floor settings (*e.g.*, a quite room). TSR decrease slightly to 85% when the noise floor grows to 45 dBSPL (*e.g.*, in a common human conversion), and then drops to around 60% as we further increase noise floor to 50 dBSPL. On the other hand, we observe TSR for Meta-Enha method maintains in a high level in all these four noise floor settings. This result shows Metamorph is robust to moderate ambient noise levels, *e.g.*, SNR is greater than 25.

4) Effect of Speaker Volume: Moreover, we further vary the transmission power from 45, 55, 65 to 75 dBSPL and examine the system performance with the ambient noise around 35 dBSPL. Figure 20 shows the performance of Meta-Enha and Meta-Qual. When the speaker volume is 65 and 75 dBSPL (SNR is 30 and 45 respectively), both TSR and CSR are nearly 100%. When the speaker volume is tuned to 55 dBSPL (SNR is 20), the attack performance slightly degrades, *e.g.* TSR of Meta-Enha degrades to 0.9 and TSR of Meta-Qual degrades to around 0.82. When speaker volume is further reduced to 45 (SNR is 10), the attack successful rates become low.

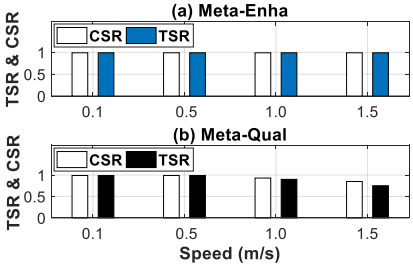

Figure 21: CSR and TSR achieved by (a) Meta-Enha and (b) Meta-Qual in different attacker moving speed settings.

5) Effect of Victim Device Movement: We finally investigate the possibility of attacking when the victim device is moving, which is a nature and practical scenario for the adversarial attack. In this experiment we place the attacker speaker on a table and play adversarial examples generated by Meta-Qual and Meta-Enha, respectively. We hold the victim device in hand and move towards and backwards the table at a different yet relatively constant speed (0.1 m/s, 0.5 m/s, 1.0 m/s, and 1.5 m/s). The result is shown in Figure 21. We observe CSR for Meta-Qual is consistently high (>90%) when the attacker moves at both low (0.5m/s) and normal speed (1.5m/s). TSR for Meta-Qual, on the other hand, decreases slightly when the attacker moves at 1.5m/s. Both CSR and

Schemes	Target model	Attack model	Over-the-air	Attack scenes	Successful rate	Audio quality (MCD)
Black-box Attacks [43], [47]	DeepSpeech	Black-box	No	-	-	-
Qin *et al.* [42]	Lingvo [30]	White-box	No	Simulated	-	-
Carlini *et al.* [17]	DeepSpeech	White-box	No	-	-	-
Abdullah *et al.* [12]	DeepSpeech	White-box	Yes	0.3m (1 foot)	15/15 (trials)	-
CommanderSong [57]	Kaldi [41]	White-box	Yes	1.5m	78 %	22.3
Yakura *et al.* [56]	DeepSpeech	White-box	Yes	0.5m	80 %	25.1
Meta-Enha	DeepSpeech	White-box	Yes	6m / NLOS	90 % / 85.5 %	25.2
Meta-Qual	DeepSpeech	White-box	Yes	3m	90 %	21.1

TABLE 3: The state-of-the-art audio adversarial attacks. "-" indicates the information is not available. We compute MCD value for [56] and [57] based on their released attack samples.

TSR for Meta-Enha are around 100% in all four moving speed settings. The result shows that both Metamorph versions are robust to the victim's normal movement.

V. Related Work

Audio adversarial examples. Early study [50] reveals the possibility to conduct an adversarial attack on speech recognition (SR) systems, while the generated adversarial examples can be easily perceived by human [46]. Alzantot *et al.* [14] later attack a command word recognition model without the listener's perception. Motivated by [14], Taori *et al.* [49] further attack DeepSpeech [27]. However, their major limitation is that the recognized command contain no more than two words [56].

Recently, Carlini *et al.* [16] realize an attack on general HMM-based RS systems without the constraint of the command's word number, and later they introduce a targeted audio adversarial attack on the state-of-the-art SR system DeepSpeech in [17]. Study [46] further introduces an attack with the dedicated temporal alignment and back-propagation designs, and Liu *et al.* [36] propose a weighted-sampling method to reduce the search space. Qin *et al.* [42] propose a set of frequency masking algorithms to improve the imperceptibly of adversarial attacks. Felix *et al.* [33] design adversarial examples to attack voice authentication system. Moustapha *et al.* [18] proposes a general adversarial example generation method, which can works on any gradient-based machine learning models. Moreover, there are also few works leveraging evolutionary algorithms to initiate black-box attack [43], [47]. However, the adversarial examples generated from these works cannot survive after the over-the-air transmission. Later on researchers try to make these adversarial attacks work in real-world scenarios (as listed in Table 3). Yuan *et al.* [57] integrate the commands into a song and Abdullah *et al.* [12] leverage the similar frequency domain feature vectors extracted from multiple source audios to generate audio adversarial examples that can initially succeed after over-the-air transmission. Yakura *et al.* [56] further propose to inject the CIR collected at other places into the training model and achieve descent success rate. However, they mainly work in short range, *e.g.*, 0.3 m to 1 m, and/or require the physical presence of the attack devices.

Embedding bits into audio. In the literature, there are also some existing works that propose to embed bits into audios for different application designs. For example, Dhwani [38] utilizes the acoustic signals to develop a secure near-field communication protocol. GeneWave [54] proposes an efficient authentication design for mobile devices. The study [58] further introduces a secure communication design without using keys. These works mainly focus on the security-related application designs. There are also some prior works that propose more general methods to embed bits into sounds to achieve a side-channel information delivery [35], [40], [52]. These designs mainly embed bits into the high-frequency band, such as 18 kHz – 20 kHz, to minimize the perception of human. The generation of audio adversarial examples, in both Metamorph and prior attacks, also add bits into audios. However, these bits are usually added in the audible range, *e.g.*, 0 kHz – 8 kHz, because SR mainly uses this range for the recognition. Therefore, the audio quality is one crucial consideration in the adversarial example generation.

Microphone non-linearity. In the literature, some recent studies, like [44], [45], [59], successfully realize a series of inaudible attacks on the speech recognition by harnessing the non-linearities of the diaphragm of microphone and the power amplifier of receiver [44]. The attacker can inject the sneaky voice commands to the speech recognition system of the victim receiver, and the device's owner cannot hear such commands. However, these recent inaudible attacks all require the special speaker hardware to play ultrasonic acoustic signals, incurring the extra hardware requirement. Moreover, it is successfully defended in [45]. These works do not belong to the adversarial attack, which are parallel to Metamorph and do not address our unique challenges in this paper.

Assorted topics related to Metamorph. There are also some other types of adversarial examples and the most representative example is the image-based ones [15], [34]. For the image adversarial example generation, there exists a similar problem — whether the image adversarial examples can survive when they are taken by a camera? RP_2 [20] recently reports a successful attack by taking the varying of distances and angles between the camera and the adversarial image into consideration in the perturbation training. However, the technical challenges in acoustic channels are different compared with the existing image-based adversarial attacks.

On the other hand, to improve the attacking distance, we also utilize the domain discriminator training methods [29], [60]. Inspired by these existing works, we further propose a dedicated domain discriminator to exclude the device- and environmental-dependent features from the prior measurements in the training of the adversarial example perturbation.

VI. Conclusion

This paper presents Metamorph to generate over-the-air audio adversarial examples. We first conduct extensive empirical studies to understand this attack in the over-the-air setting

and observe that the reason undermining prior designs is the frequency-selectivity caused by both device and channel. To cope with this issue, we propose a "generate-and-clean" two-phase design and also consider the audio quality of generated adversarial examples. The evaluation shows the efficacy and good performance of Metamorph.

ACKNOWLEDGMENT

We sincerely thank the anonymous reviewers for their helpful comments and feedback. This work is supported by the GRF grant from Research Grants Council of Hong Kong (Project No. CityU 11217817). This work is also supported by NSF Award CNS-1617161.

REFERENCES

[1] "Amazon Echo and Alexa," https://alexa.amazon.com.

[2] "Audio Adversarial Examples," https://github.com/carlini/audio_adversarial_examples.

[3] "Google Now," https://en.wikipedia.org/wiki/Google_Now.

[4] "In-Car Voice Commands NLP for Self-Driving Cars," https://aitrends.com/ai-insider/car-voice-commands-nlp-self-driving-cars/.

[5] "M200MKIII+ Bluetooth Bookshelf Speakers," https://swanspeakers.com/product/m200mkiii-bluetooth-bookshelf-speakers/.

[6] "Mozilla Common Voice Dataset," https://voice.mozilla.org/en/datasets.

[7] "Mycroft," https://mycroft.ai/.

[8] "Project DeepSpeech," https://github.com/mozilla/DeepSpeech.

[9] "Project Website of Metamorph," https://acoustic-metamorph-system.github.io/.

[10] "SwiftScribe," https://swiftscribe.ai/.

[11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceedings of USENIX OSDI*, 2016.

[12] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," in *Proceedings of NDSS*, 2019.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 1979.

[14] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," in *Proceedings of NIPS*, 2017.

[15] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of ICML*, 2018.

[16] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands." in *Proceedings of USENIX Security Symposium*, 2016.

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

[18] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Proceedings of NIPS*, 2017.

[19] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proceedings of IEEE ICASSP*, 2009.

[20] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of IEEE CVPR*, 2018.

[21] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES Convention*, 2000.

[22] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of ICML*, 2015.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*, 2006.

[24] S. L. R. H. Guillaume Delorme, Xavier Alameda-Pineda, "Camera adversarial transfer for unsupervised person re-identification," *https://arxiv.org/abs/1904.01308*, 2019.

[25] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *Proceedings of ACM CCS*, 2018.

[26] A. Hannun, "Sequence modeling with ctc," *Distill*, 2017.

[27] 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.

[28] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of IEEE DSP*, 2009.

[29] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of ACM MobiCom*, 2018.

[30] e. Jonathan Shen, "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," *https://arxiv.org/abs/1902.08295*, 2019.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of IEEE WASPAA*, 2013.

[33] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," *https://arxiv.org/pdf/1801.03339*, 2018.

[34] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *ICLR Workshop*, 2017.

[35] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *Proceedings of IEEE INFOCOM*, 2015.

[36] X. Liu, K. Wan, and Y. Ding, "Adversarial attack on speech-to-text recognition models," *arXiv preprint arXiv:1901.10300*, 2019.

[37] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of LREC*, 2000.

[38] R. Nandakumar, K. K. Chintalapudi, V. Padmanabhan, and R. Venkatesan, "Dhwani: secure peer-to-peer acoustic nfc," in *Proceedings of ACM SIGCOMM*, 2013.

[39] T. Y. K. A. Naoto Inoue, Ryosuke Furuta, "Cross-domain weakly-supervised object detection through progressive domain adaptation," *https://arxiv.org/abs/1803.11365*, 2018.

[40] A. S. Nittala, X.-D. Yang, S. Bateman, E. Sharlin, and S. Greenberg, "Phoneear: interactions for mobile devices that hear high-frequency sound-encoded data," in *Proceedings of ACM SIGCHI*, 2015.

[41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proceedings of IEEE ASRU*, 2011.

[42] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," *arXiv preprint arXiv:1903.10346*, 2019.

[43] B. C. N. V. Rohan Taori, Amog Kamsetty, "Targeted adversarial examples for black box audio systems," *https://arxiv.org/pdf/1805.07820*, 2018.

[44] N. Roy, H. Hassanieh, and R. Roy Choudhury, "Backdoor: Making microphones hear inaudible sounds," in *Proceedings of ACM MobiSys*, 2017.

[45] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, "Inaudible voice commands: The long-range attack and defense," in *Proceedings of USENIX NSDI*, 2018.

[46] 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.

[47] S. M. Shreya Khare, Rahul Aralikatte, "Adversarial black-box attacks for automatic speech recognition systems using multi-objective genetic optimization," *https://arxiv.org/abs/1811.01312*, 2018.

[48] "Apple Siri," https://www.apple.com/siri/.

[49] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," *arXiv preprint arXiv:1805.07820*, 2018.

[50] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *Proceedings of USENIX WOOT*, 2015.

[51] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Proceedings of Interspeech*, 2013.

[52] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages behind the sound: real-time hidden acoustic signal capture with smartphones," in *Proceedings of ACM MobiCom*, 2016.

[53] J. Y. Wen, N. D. Gaubitch, E. A. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the mardy database," in *Proceedings of IWAENC*, 2006.

[54] P. Xie, J. Feng, Z. Cao, and J. Wang, "Genewave: Fast authentication and key agreement on commodity mobile devices," in *Proceedings of IEEE ICNP*, 2017.

[55] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wifi," in *Proceedings of ACM MobiCom*, 2015.

[56] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," *arXiv preprint arXiv:1810.11793*, 2018.

[57] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," *arXiv preprint arXiv:1801.08535*, 2018.

[58] B. Zhang, Q. Zhan, S. Chen, M. Li, K. Ren, C. Wang, and D. Ma, "Priwhisper: Enabling keyless secure acoustic communication for smartphones," *IEEE internet of things journal*, 2014.

[59] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of ACM CCS*, 2017.

[60] 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.

[61] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of ECCV*, 2018.

APPENDIX

A. Configuration of System Parameters

The final loss function of Metamorph in Eqn. (10) includes five parameters, including α, β, γ, η and μ. In this subsection, we introduce how they are configured in this paper.

Parameter α. This parameter is based on the audio adversarial example generation method proposed in [17], which aims to balance the audio sound distortion, described by Decibels (dB), and the attack successful rate, described by the Connectionist Temporal Classification (CTC) loss. Although it is possible to train it directly, a more efficient mechanism is implemented in [2] to avoid a direct parameter tuning. In our implementation, we also adopt this mechanism without turning α directly.

Parameters β and γ. These two parameters are introduced in Metamorph to ensure the good attack performance after the over-the-air transmission of the adversarial example. Parameter β balances the adversarial example generation and the ability to distinguish domains by the domain discriminator. We vary[6] β from 0.005 to 0.5 in Figure 22(a). From the results, we observe that both the transcript successful rate (TSR) and character successful rate (CSR) at a moderate attack distance of 3 m can achieve a better performance (*e.g.*, > 0.95) when β is 0.05. The audio quality, measured by MCD (Mel Cepstral Distortion), keeps relative stable in this experiment. We thus experimentally adopt 0.05 as the default β setting in current Metamorph. With this setting, more experiments from the evaluation section show the good system performance at other attack distances as well. On the other hand, parameter γ is introduced to reduce the over-fitting. Through the experiment in Figure 22(b), we observe that when we increase γ, *e.g.*, 500 or 1000, TSR approaches to nearly 100%. The audio quality degrades only slightly. However, when we further increase γ, both TSR/CSR and audio quality drop rapidly. Therefore, we adopt 500 as the default γ setting in the current Metamorph.

Parameters η and μ. These two parameters are introduced in Metamorph to mainly improve the audio quality of the generated adversarial example. Parameter η controls the utility of the audio graffiti. In Figure 22(c), we vary η from 1e-5 to 1e-3. The result shows that when η increases, the audio quality, measured by MCD, keeps improving, while CSR and TSR drop significantly when η is greater than 1e-4. Hence, we adopt 1e-4 as the default η setting in current Metamorph. On the other hand, parameter μ is introduced to reduce the perturbation coverage. Through the experiment in Figure 22(d), we observe that the increase of μ also leads to the improvement of the audio quality MCD, while the CSR and TSR will drop concurrently. As a result, we adopt 1e-12 as the default μ setting in the current Metamorph.

These default parameters introduced above are utilized in the experimental evaluations in Section IV.

B. User Perceptibility Study Questions

In the first trial of the experiments conducted in the user perceptibility study of Section IV-B, volunteers will sequentially listen to each set of audios following the organization

[6]Principle of each parameter's varying range is to ensure its product with its loss function will be comparable to other terms in Eqn. (10).

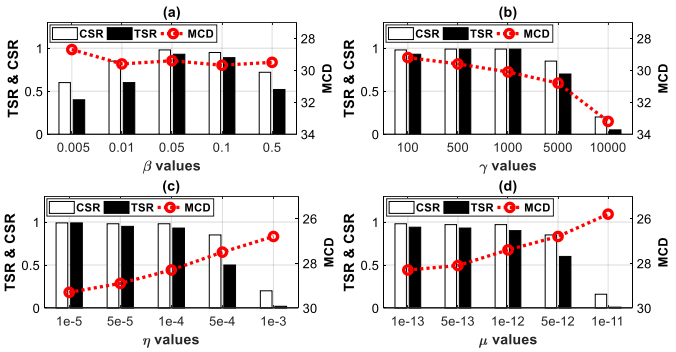

Figure 22: Experimental configurations for system parameters β, γ, η and μ (with a reversed y-axis representation for MCD).

below: "[(one original audio, the adversarial example generated from this audio by **Meta-Enha**), *60s pause*, (the same original audio, the adversarial example generated from this audio by **Meta-Qual**), *60s pause*]". During each pause, the volunteers immediately assess the audio quality of each adversarial example compared with the original audio. Volunteers first select whether each adversarial example has the same audio quality as the original audio (Y or N), including both the noise level and the audio content. If the answer is Y, the assessment of this adversarial example in the first trial is complete; Otherwise, volunteers will further select for three questions related to 1) **word (content) change**, 2) **audio quality level** and 3) **noise description**. The explanation of each question is in Table 4.

		Explanations
Word change	**Y**	Word (content) change perceived.
	N	No word (content) change perceived.
Quality level	**1**	Noise is small and audio content is clear.
	2	Noise is a slightly loud, but it does not impact my hearing of the audio content.
	3	Noise is loud and I cannot hear the audio content occasionally.
	4	Noise is annoying and I cannot hear the audio content consistently.
Description	**A**	Noise is brought by the hardware, *e.g.*, microphone recording, cheap speaker, etc.
	B	Noise is due to the low-quality of the audio clip itself.
	C	Others (using your own words).

TABLE 4: Explanations of word change, audio quality level and description three fields in the user perceptibility study.

C. Adversarial Examples Used in Evaluation

We generate two types of adversarial examples (music and speech) with different source and target transcripts, which are detailed in Table 5. The source musics are labelled in the table directly, and the speech adversarial examples are generated based on 11 different speech samples from the public Mozilla Common Voice Dataset [6].

No.	Source audio transcripts (musics)	Target commands
1	"[no transcript]"–*Bach, Violin*	"hello world"
2	"chase your dreams and remember me sweet bravery"–*Owl City, To The Sky*	"power off"
3	"I feel earth move under my feet I feel the sky"–*Carole King, I Feel The Earth Move*	"pay the money"
4	"lyrical acrobat stunts while I'm practicing that I'll still be able to break a motherfuckin'table over the back of a couple"–*Eminem, Rap God*	"turn off the light"
5	"well the kid is into losin' sleep and he don't come home for half the week"–*Van Halen, And the Cradle Will Rock*	"airplane mode on"
6	"[no transcription]"–*Van Halen, Guitar*	"browse to evil dot com"
7	"somebody mix my medicine"–*The pretty Reckless, My Medicine*	"turn off the cellular network"
8	"[no transcription]"–*Chopin, Piano*	"update the phone blacklist"
9	"I am a mountaineer in the"–*Owl City, Hello Seattle*	"silence the phone"
No.	**Source audio transcripts (speeches)**	**Target commands**
1	"hold your nose to keep the smell from disabling your motor functions"	"clear all appointments on calendar"
2	"your son went to server at a distant place and became a centurion"	"open the door"
3	"the shower's in there"	"restart"
4	"and you know it"	"open the camera"
5	"this is no place for you"	"flashlight on"
6	"if I had told you you wouldn't have seem the pyramids"	"play the scary music"
7	"I told you to have the ice box fixed"	"call nine one one"
8	"their faces were hidden behind blue veils with only their eyes showing"	"send me your messages"
9	"we are refugees from the tribal wars and we need money the other figure said"	"log in paypal"
10	"isn't the party also to announce his engagement to joanna"	"show fake traffic information"
11	"he stood irresolute for a moment and then scrambled out of the pit"	"shut down the power source"

TABLE 5: Source audios and target transcripts used in Metamorph, where "[no transcription]" means that there is no transcript when the classical music is played. The source musics are labelled in the table and the source audio for speeches are from the Mozilla Common Voice Dataset.