Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Robust Audio Adversarial Example for a Physical Attack

Hiromu Yakura[1,2*] and **Jun Sakuma**[1,2]
[1]University of Tsukuba
[2]RIKEN Center for Advanced Intelligence Project
hiromu@mdl.cs.tsukuba.ac.jp, jun@cs.tsukuba.ac.jp

Abstract

We propose a method to generate audio adversarial examples that can attack a state-of-the-art speech recognition model in the physical world. Previous work assumes that generated adversarial examples are directly fed to the recognition model, and is not able to perform such a physical attack because of reverberation and noise from playback environments. In contrast, our method obtains robust adversarial examples by simulating transformations caused by playback or recording in the physical world and incorporating the transformations into the generation process. Evaluation and a listening experiment demonstrated that our adversarial examples are able to attack without being noticed by humans. This result suggests that audio adversarial examples generated by the proposed method may become a real threat.

1 Introduction

In recent years, deep learning has achieved vastly improved accuracy, especially in fields such as image classification and speech recognition, and has come to be used practically [LeCun *et al.*, 2015]. On the other hand, deep learning methods are known to be vulnerable to adversarial examples [Szegedy *et al.*, 2014, Goodfellow *et al.*, 2015]. More specifically, an attacker can make deep learning models misclassify examples by intentionally adding a small perturbation to the examples. Such examples are referred to as adversarial examples.

While many papers discussed image adversarial examples against image classification models, little research has been done on audio adversarial examples against speech recognition models, even though speech recognition models are widely used at present in commercial applications like Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana and devices like Amazon Echo and Google Home. For example, [Carlini and Wagner, 2018] proposed a method to generate audio adversarial examples against DeepSpeech [Hannun *et al.*, 2014], which is a state-of-the-art speech recognition model. However, this method targets the case in which the waveform of the adversarial example is input directly to the model, as shown in Figure 1(A). In other words, it is not feasible to attack in the case that the adversarial example is played by a speaker and recorded by a microphone in the physical world (hereinafter called the *over-the-air* condition), as shown in Figure 1(B).

The difficulty of such an over-the-air attack can be attributed to the reverberation of the environment and noise from both the speaker and the microphone. More specifically, in the case of the direct input, adversarial examples can be generated by determining a single data point that fools the targeted model using an optimization algorithm for a clearly described objective. In contrast, under the over-the-air condition, adversarial examples are required to be robust against unknown environments and equipment.

Considering that audio signals spread through the air, the impact of a physical attack using audio adversarial examples would be larger than that using image adversarial examples. For an attack scenario using an image adversarial example, the adversarial example must be presented explicitly in front of an image sensor of the attack target, e.g., the camera of an auto-driving car. In contrast, audio adversarial examples can simultaneously attack numerous targets by spreading via outdoor speakers or radios. If an attacker hijacks the broadcast equipment of a business complex, it will be possible to attack all the smartphones owned by people inside via a single playback of the audio adversarial example.

In the present paper, we propose a method by which to generate a robust audio adversarial example that can attack speech recognition models in the physical world. To the best of our knowledge, this is the first approach to succeed in generating such adversarial examples that can attack complex speech recognition models based on recurrent networks, such as DeepSpeech, over the air. Moreover, we believe that our research will contribute to improving the robustness of speech recognition models by training models to discriminate adversarial examples through a process similar to adversarial training in the image domain [Goodfellow *et al.*, 2015].

1.1 Related Research

Some studies have proposed methods to generate audio adversarial examples against speech recognition models [Alzantot *et al.*, 2018, Taori *et al.*, 2018, Cissé *et al.*, 2017, Schönherr *et al.*, 2018, Carlini and Wagner, 2018]. These methods are

*Contact Author

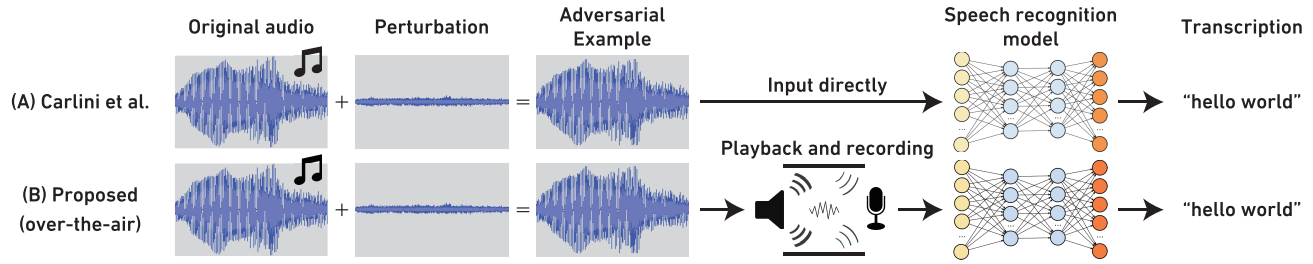

Figure 1: Illustration of the proposed attack. [Carlini and Wagner, 2018] assumed that adversarial examples are provided directly to the recognition model. We propose a method that targets an over-the-air condition, which leads to a real threat.

divided into two groups: black-box and white-box settings.

In the black-box setting, in which the attacker can only use the score that represents how close the input audio is to the desired phrase, [Alzantot *et al.*, 2018] proposed a method to attack a speech command classification model [Sainath and Parada, 2015]. This method exploits a genetic algorithm to find an adversarial example, which is recognized as a specified command word. Inspired by this method, [Taori *et al.*, 2018] proposed a method to attack DeepSpeech [Hannun *et al.*, 2014] under the black-box setting by combining genetic algorithms and gradient estimation. One limitation of their method is that the length of the phrase that the attacker can make the models recognize is restricted to two words at most, even when the obtained adversarial example is directly inputted. [Cissé *et al.*, 2017] performed an attack on Google Voice application using adversarial examples generated against DeepSpeech-2 [Amodei *et al.*, 2016]. The aim of their attack was changing recognition results to different words without being noticed by humans. In other words, they could not make the targeted model output desired words and concluded that attacking speech recognition models so as to transcribe specified words "seem(s) to be much more challenging." From these points, current methods in the black-box settings are not realistic for considering the attack scenario in the physical world.

In the white-box setting, in which the attacker can access the parameters of the targeted models, [Yuan *et al.*, 2018] proposed a method to attack Kaldi [Povey *et al.*, 2011], a conventional speech recognition model based on the combination of deep neural network and hidden Markov model. [Schönherr *et al.*, 2018] extended the method such that generated adversarial examples are not noticed by humans using a hiding technique based on psychoacoustics. Although [Yuan *et al.*, 2018] succeeded in attacking over the air, their method is not applicable to speech recognition models based on recurrent networks, which are becoming more popular and highly functional. For example, Google replaced its conventional model with a recurrent network based model in 2012[1].

In that respect, [Carlini and Wagner, 2018] proposed a white-box method to attack against DeepSpeech, a recurrent network based model. However, as mentioned previously, this method succeeds in the case of the direct input, but not in the over-the-air condition, because of the reverberation of the environment and noise from both the speaker and the microphone. Thus, the threat of the obtained adversarial example is limited regarding the attack scenario in the physical world.

1.2 Contribution

The contribution of the present paper is two-fold:

- We propose a method by which to generate audio adversarial examples that can attack speech recognition models based on recurrent networks under the over-the-air condition. Note that such a practical attack is not achievable using the conventional methods described in Section 1.1. We addressed the problem of the reverberation and the noise in the physical world by simulating them and incorporating the simulated influence into the generation process.
- We show the feasibility of the practical attack using the adversarial examples generated by the proposed method in evaluation and a listening experiment. Specifically, the generated adversarial examples demonstrated a success rate of 100% for the attack through both speakers and radio broadcasting, although no participants heard the target phrase in the listening experiment.

2 Background

In this section, we briefly introduce an adversarial example and review current speech recognition models.

2.1 Adversarial Example

An adversarial example is defined as follows. Given a trained classification model $f : \mathbb{R}^n \rightarrow \{1, 2, \cdots, k\}$ and an input sample $\boldsymbol{x} \in \mathbb{R}^n$, an attacker wishes to modify $\boldsymbol{x}$ so that the model recognizes the sample as having a specified label $l \in \{1, 2, \cdots, k\}$ and the modification does not change the sample significantly:

$$\tilde{\boldsymbol{x}} \in \mathbb{R}^n \; s.t. \; f(\tilde{\boldsymbol{x}}) = l \wedge \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| \leq \delta \tag{1}$$

Here, δ is a parameter that limits the magnitude of perturbation added to the input sample and is introduced so that humans cannot notice the difference between a legitimate input sample and an input sample modified by an attacker.

Let $\boldsymbol{v} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$ be the perturbation. Then, adversarial examples that satisfy Equation 1 can be found by optimizing this

[1] https://ai.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html

in which $Loss_f$ is a loss function that represents how distant the input data are from the given label under the model f:

$$\underset{\boldsymbol{v}}{\operatorname{argmin}} \underset{f}{Loss}\left(\boldsymbol{x}+\boldsymbol{v}, l\right)+\epsilon\|\boldsymbol{v}\| \tag{2}$$

By solving the problem using optimization algorithms, the attacker can obtain an adversarial example. In particular, when f is a differentiable model, such as a regular neural network, and a gradient on $\boldsymbol{v}$ can be calculated, a gradient method such as Adam [Kingma and Ba, 2015] is often used.

2.2 Image Adversarial Example for a Physical Attack

Considering attacks on physical recognition devices (e.g., object recognition of auto-driving cars), adversarial examples are given to the model through sensors. In the example of the auto-driving car, image adversarial examples are given to the model after being printed on physical materials and photographed by a car-mounted camera. Through such a process, the adversarial examples are transformed and exposed to noise. However, adversarial examples generated by Equation 2 are assumed to be given directly to the model and do not work for such scenarios.

In order to address this problem, [Athalye *et al.*, 2018] proposed a method to simulate transformations caused by printing or taking a picture and incorporate the transformations into the generation process of image adversarial examples. This method can be represented as follows using a set of transformations $\mathcal{T}$ consisting of, e.g., enlargement, reduction, rotation, change in brightness, and addition of noise:

$$\begin{aligned}\underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathbb{E}_{t\sim\mathcal{T}}\Big[&\underset{f}{Loss}\left(t\left(\boldsymbol{x}+\boldsymbol{v}\right), l\right)\\ &+\epsilon\left\|t\left(\boldsymbol{x}\right)-t\left(\boldsymbol{x}+\boldsymbol{v}\right)\right\|\Big]\end{aligned} \tag{3}$$

As a result, adversarial examples are generated so that images work even after being printed and photographed.

2.3 Audio Adversarial Example

As explained in Section 1.1, [Carlini and Wagner, 2018] succeeded to attack against DeepSpeech, a recurrent network based model. Here, the targeted model has time-dependency and the same approach as image adversarial examples is not applicable. Thus, based on the fact that the targeted model uses Mel-Frequency Cepstrum Coefficient (MFCC) for the feature extraction, they implemented MFCC calculation in a differentiable manner and optimized an entire waveform using Adam [Kingma and Ba, 2015].

In detail, the perturbation $\boldsymbol{v}$ is obtained against the input sample $\boldsymbol{x}$ and the target phrase $\boldsymbol{l}$ using the loss function of DeepSpeech as follows:

$$\underset{\boldsymbol{v}}{\operatorname{argmin}} \underset{f}{Loss}\left(MFCC\left(\boldsymbol{x}+\boldsymbol{v}\right), \boldsymbol{l}\right)+\epsilon\|\boldsymbol{v}\| \tag{4}$$

Here, $MFCC\left(\boldsymbol{x}+\boldsymbol{v}\right)$ represents the MFCC extraction from the waveform of $\boldsymbol{x}+\boldsymbol{v}$. They reported the success rate of the obtained adversarial examples as 100% when inputting waveforms directly into the recognition model, but did not succeed at all under the over-the-air condition.

To the best of our knowledge, there has been no proposal to generate audio adversarial examples, which work under the over-the-air condition, targeting speech recognition models using a recurrent network.

3 Proposed Method

In this research, we propose a method by which to generate a robust adversarial example that can attack DeepSpeech [Hannun *et al.*, 2014] under the over-the-air condition. The basic idea is to incorporate transformations caused by playback and recording into the generation process, similar to [Athalye *et al.*, 2018]. We introduce three techniques: a band-pass filter, impulse response, and white Gaussian noise.

3.1 Band-pass Filter

Since the audible range of humans is 20 to 20,000 Hz, normal speakers are not made to play sounds outside this range. Moreover, microphones are often made to automatically cut out all but the audible range in order to reduce noise. Therefore, if the obtained perturbation is outside the audible range, the perturbation will be cut during playback and recording and will not function as an adversarial example.

Therefore, we introduced a band-pass filter in order to explicitly limit the frequency range of the perturbation. Based on empirical observations, we set the band to 1,000 to 4,000 Hz, which exhibited less distortion. Here, the generation process is represented as follows based on Equation 4:

$$\begin{aligned}&\underset{\boldsymbol{v}}{\operatorname{argmin}} \underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right), \boldsymbol{l}\right)+\epsilon\|\boldsymbol{v}\|\\ &where\ \tilde{\boldsymbol{x}}=\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)\end{aligned} \tag{5}$$

In this way, it is expected that the generated adversarial examples will acquire robustness such that they function even when frequency bands outside the audible range are cut by a speaker or a microphone.

3.2 Impulse Response

Impulse response is the reaction obtained when presented with a brief input signal, called an impulse. Based on the fact that impulse responses can reproduce the reverberation in the captured environment by convolution, a method of using impulse responses from various environments in the training of a speech recognition model to enhance the robustness to the reverberation has been proposed [Peddinti *et al.*, 2015]. Similarly, we introduced impulse responses to the generation process in order to make the obtained adversarial example robust to reverberations.

In addition, considering the scenario of attacking numerous devices at once via outdoor speakers or radios, we want the obtained adversarial example to work in various environments. Therefore, in the same manner as [Athalye *et al.*, 2018], we take an expectation value over impulse responses recorded in diverse environments. Here, Equation 5 is extended like Equation 3, where the set of collected impulse responses is $\mathcal{H}$ and the convolution using impulse response h is

$Conv_h$:

$$\underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathbb{E}_{h\sim\mathcal{H}}\left[\underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right),\boldsymbol{l}\right)+\epsilon\left\|\boldsymbol{v}\right\|\right]$$
$$where\ \tilde{\boldsymbol{x}}=\underset{h}{Conv}\left(\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)\right) \quad (6)$$

In this way, it is expected that the generated adversarial examples will acquire robustness such that they are not affected by reverberations produced in the environment in which they are played and recorded.

3.3 White Gaussian Noise

White Gaussian noise is given by $\mathcal{N}\left(0,\sigma^2\right)$ and used for emulating the effect of many random processes that occur in nature. For example, it is used in the evaluation of speech recognition models to measure their robustness against the background noise [Hansen and Pellom, 1998]. Consequently, we introduce white Gaussian noise in the generation process in order to make the obtained adversarial example robust to background noise. Here, Equation 6 is extended as follows:

$$\underset{\boldsymbol{v}}{\operatorname{argmin}}\, \mathbb{E}_{h\sim\mathcal{H},\boldsymbol{w}\sim\mathcal{N}(0,\sigma^2)}\left[\underset{f}{Loss}\left(MFCC\left(\tilde{\boldsymbol{x}}\right),\boldsymbol{l}\right)+\epsilon\left\|\boldsymbol{v}\right\|\right]$$
$$where\ \tilde{\boldsymbol{x}}=\underset{h}{Conv}\left(\boldsymbol{x}+\underset{1000\sim4000\mathrm{Hz}}{BPF}\left(\boldsymbol{v}\right)\right)+\boldsymbol{w} \quad (7)$$

In this way, it is expected that the generated adversarial examples will acquire robustness such that they are not affected by noise caused by recording equipment and the environment. Note that the white Gaussian noise should also be added before the convolution for the purpose of emulating thermal noise caused in both the playback and recording devices. However, we added the noise only after the convolution because doing so makes the optimization easier and Equation 7 was sufficiently robust in the empirical observations.

4 Evaluation

In order to confirm the effectiveness of the proposed method, we conducted evaluation experiments. We played and recorded audio adversarial examples generated by the proposed method and verified whether these adversarial examples are recognized as target phrases.

4.1 Implementation

We implemented Equation 7 using TensorFlow[2]. Since calculating the expected value of the loss is difficult, we instead evaluated the sample approximation of Equation 7 with respect to a fixed number of impulse responses sampled randomly from $\mathcal{H}$. For optimization, we used Adam [Kingma and Ba, 2015] in the same manner as [Carlini and Wagner, 2018].

[2]Our full implementation is available at https://github.com/hiromu/robust_audio_ae

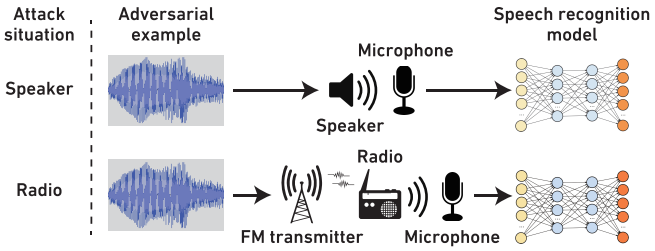

Figure 2: Two attack situations of the evaluation: speaker and ratio. In the first situation, the adversarial examples were played and recorded by a speaker and a microphone. In the second situation, the adversarial examples were broadcasted using an FM radio.

4.2 Settings

For the input sample $\boldsymbol{x}$, we prepared two different audio clips of four seconds cut from *Cello Suite No. 1* by Bach and *To The Sky* by Owl City. The first clip is the same as the publicly released samples[3] of [Carlini and Wagner, 2018]. The second clip is the same as the publicly released samples[4] of [Yuan *et al.*, 2018]. The difference between the clips is that the first clip is an instrumental piece and does not include singing voices, whereas singing voices are included in the second song by Owl City.

For the target phrase $\boldsymbol{l}$, we prepared three different cases: "hello world," "open the door[5]," and "ok google[6]." Considering that [Carlini and Wagner, 2018] tested their method with 1,000 phrases that were randomly chosen from a speech dataset, three phrases appear to be insufficient to evaluate the efficiency of our attack. However, unlike the direct attack as performed by [Carlini and Wagner, 2018], our evaluation involves a number of playback cycles in the physical world. This means that our experimental evaluation in the over-the-air setting requires actual time for playing back the generated audio adversarial examples. For example, our evaluation of a single combination of the input sample and the target phrase requires more than 18 hours in a quiet room without interruption because it involves playing 500 intermediate examples 10 times each with an interval of several seconds. For this reason, we focused on these three phrases considering the attack scenarios.

For the set of impulse responses $\mathcal{H}$, we collected 615 impulse responses from various databases [Kinoshita *et al.*, 2013, Nakamura *et al.*, 2000, Jeub *et al.*, 2009, Wen *et al.*, 2006, Härmä, 2001], which are constructed primarily for research on dereverberation.

For the playback and the recording, we prepared two different attack situations, as shown in Figure 2, in order to confirm that the attack using the generated adversarial examples is applicable via a wide range of offensive means. First, we played and recorded the adversarial examples using a speaker and a microphone (JBL CLIP2 / Sony ECM-PCV80U) in a meet-

[3]https://nicholas.carlini.com/code/audio_adversarial_examples/

[4]https://sites.google.com/view/commandersong/

[5]This phrase is used in [Yuan *et al.*, 2018] to discuss an attack scenario using voice commands.

[6]This phrase is used as a trigger word of Google Home.

	Input sample	Target phrase	SNR
(A)	Bach	hello world	9.3dB
(B)	Bach	open the door	5.3dB
(C)	Bach	ok google	0.2dB
(D)	Owl City	hello world	11.8dB
(E)	Owl City	open the door	13.4dB
(F)	Owl City	ok google	2.6dB

Table 1: Details of the generated audio adversarial examples, which showed 100% success by both the speaker and the radio and having the maximum value of SNR[8].

ing room with a distance of approximately 0.5 meters. We also examined whether the generated adversarial examples could attack through the radio using HackRF One[7], a Software Defined Radio (SDR) capable of transmission or reception of radio signals. We broadcasted at 180.0MHz FM and received the signal with a portable radio (Sony ICF-P36) in the same room, while the playback was recorded by a microphone (Sony ECM-PCV80U). In both cases, we played each adversarial example 10 times to suppress random fluctuation in the physical world and evaluated the recognition results obtained by DeepSpeech.

4.3 Metrics

For the evaluation metrics of the obtained adversarial example, we used the signal-to-noise ratio (SNR) of the perturbation, the success rate of the attack, and the edit distance of the recognition results. The SNR is given by $10\log_{10}\frac{P_x}{P_v}$ for the power of the input sample $P_x = \frac{1}{T}\sum_{t=1}^{T} x_t^2$ and the power of perturbation $P_v = \frac{1}{T}\sum_{t=1}^{T} v_t^2$. In other words, a larger SNR is associated with a smaller perturbation and a smaller likelihood for a human to notice.

The success rate of the attack is the ratio of the number of times that DeepSpeech transcribed the recorded adversarial example as the target phrase among all trials. The success rate becomes non-zero only when DeepSpeech transcribes adversarial examples as the target phrase perfectly.

Thus, we also introduced the edit distance between the recognition results and the target phrase to confirm the progress of the generation process. The edit distance reveals the progress more precisely, even when the success rate is 0%. Here, the edit distance is defined as the minimum number of procedures required to transform one string into the other by inserting, deleting, and replacing one character.

4.4 Results

The progress of the generation process is presented in Figure 3. The figure shows that, as the generation progresses, the SNR decreases and the edit distance of the recognition results to the target phrase also decreases. The detailed results of the generated adversarial examples showed certain levels of the success rate are presented in Table 1.

[7]https://greatscottgadgets.com/hackrf/

[8]These audio files are available at https://yumetaro.info/projects/audio-ae/.

	Input sample	Target phrase	SNR	Attack situation	Success rate	Edit dist.
(G)	Bach	hello world	11.9dB	Speaker	60%	1.1
				Radio	50%	1.3
(H)	Bach	open the door	6.6dB	Speaker	60%	1.8
				Radio	60%	1.8
(I)	Bach	ok google	4.2dB	Speaker	80%	0.6
				Radio	70%	0.9
(J)	Owl City	hello world	12.2dB	Speaker	70%	0.9
				Radio	50%	1.5
(K)	Owl City	open the door	14.6dB	Speaker	90%	0.2
				Radio	100%	0.0
(L)	Owl City	ok google	8.7dB	Speaker	90%	0.6
				Radio	70%	0.9

Table 2: Details of the generated audio adversarial examples, which showed at least 50% success by both the speaker and the radio and having the maximum value of SNR[8].

As shown in Table 1, in all combinations of the input sample and the target phrase, the proposed method generated adversarial examples that showed 100% success by both the speaker and the radio. On the other hand, the magnitude of the perturbation required to achieve 100% success differed depending on the input sample and the target phrase. In the previous method [Yuan *et al.*, 2018] targeted at Kaldi [Povey *et al.*, 2011] under the over-the-air condition, an SNR of less than 2.0 dB was reported in all cases. In other words, considering that (D) through (F) of Table 1 use the same input sample, the proposed method is able to generate an adversarial example with less perturbation while targeted at a more complex speech recognition model.

Table 2 showed the adversarial examples having a success rate of at least 50% by both the speaker and the radio with the maximum value of SNR. We found that much less perturbation was required to achieve a success rate of 50%, as compared to Table 1, in all cases. In other words, the attack using these adversarial examples will succeed once in two attempts and can be a major threat when the attacker allows uncertainty of the attack.

In all cases in Table 1 and Table 2, the adversarial examples generated with Bach's Cello Suite No. 1 have larger SNR compared to the case of Owl City. This result supports the discussion of [Yuan *et al.*, 2018], whereby some phonemes from singing voices in the input sample work together with the injected small perturbations to form the target phrases. Considering such an effect, we can determine that the result is due to the song by Owl City having more phonemes to help form the target phrases, as compared to Bach's instrumental piece, and requires less perturbation.

We also found that the recognition results of the adversarial examples in Table 2 changed only slightly between the cases of the speaker and the radio. This result suggests that the proposed method makes the generated adversarial example robust for FM transmission also. For example, the addition of white Gaussian noise in the proposed method would also work for the noise caused by FM transmission. Moreover, as mentioned in Section 1, one of the major concerns of an

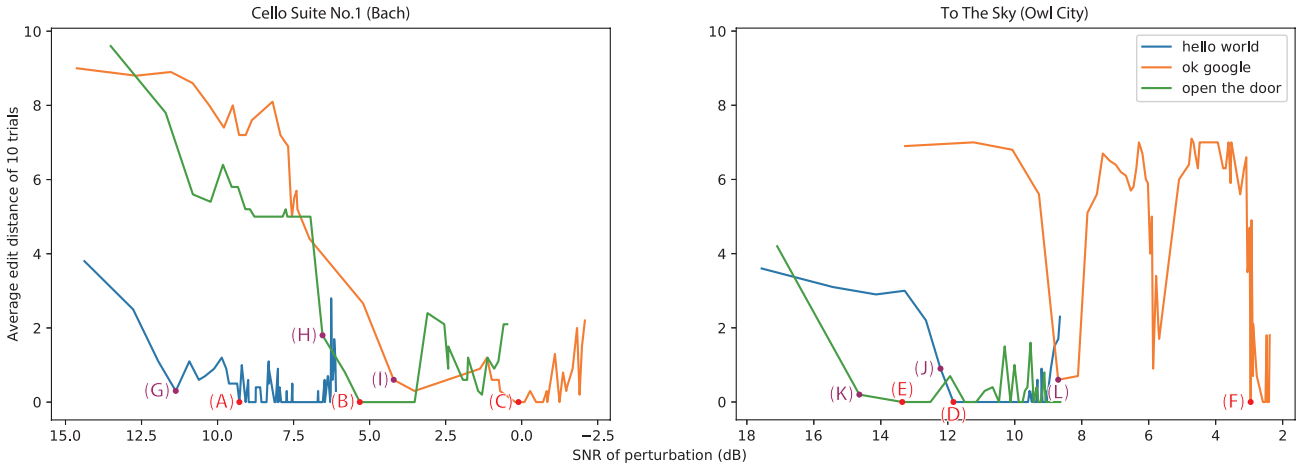

Figure 3: Progress of the generation process of the adversarial examples. As the generation progresses, the SNR and the edit distance in the speaker situation decreases. The detailed results of the highlighted adversarial examples are shown in Table 1 and Table 2.

Used techniques			Input sample	
Band-pass filter	Impulse response	Gaussian noise	Bach	Owl City
			–	–
✓			–	–
	✓		–	–
		✓	–	–
✓	✓		–	–
✓		✓	−4.2dB	−3.8dB
	✓	✓	–	–
✓	✓	✓	9.3dB[9]	11.8dB[9]

Table 3: Results of switching in the presence of each technique. Only the case of combining the band-pass filter and white Gaussian noise succeeded to generate, though it requires much more perturbation than the case of Table 1.

audio adversarial example is that it can attack numerous targets simultaneously. In this respect, the success of the attack through the radio might have a significant impact because such an attack can be made without making victims play an adversarial example actively on their own.

4.5 Effect of Each Technique

We then investigated the individual effect of the three techniques on the success of the proposed method. In detail, we evaluated the effect of the three techniques with changing the combinations in the generation. Once we obtained an adversarial example that is recognized as the target phrase using the speaker in a similar environment as Section 4.2, we compared its SNR in Table 1. Here, we used "hello world" as the targeted phrase because it is suggested to be relatively easy to generate according to Table 1.

The results are shown in Table 3. We note that the generation without any of the proposed techniques is equivalent to [Carlini and Wagner, 2018]. Here, all the cases except the combination of the band-pass filter and white Gaussian noise could not generate adversarial examples that can attack under the over-the-air condition. In addition, the succeeded combination requires much more perturbation than the case of using all the three techniques.

From these results, it is suggested that we can generate adversarial examples that work over the air without the help of the impulse responses by using white Gaussian noise, whereas the band-pass filter seems to be essential considering the limitation of physical devices. At the same time, the impulse responses are considered to make adversarial examples robust specifically for reverberations, which results in the reduction of the perturbation, as discussed in Section 3.2.

[9]These values are from Table 1.

ok google	turn off	open the door	happy birthday
good night	call john	hello world	airplane mode on

Table 4: List of choices presented to participants in the listening experiments. We chose simple phrases of lengths similar to those of "hello world" or "open the door," concentrating on phrases that are used as voice commands.

5 Listening Experiment

In order to consider an attack scenario using the generated adversarial examples, whether humans can notice is important. If an attacker can make intended phrases to be recognized without it being noticed by humans, then an attack exploiting speech recognition devices will be possible.

For example, [Yuan *et al.*, 2018] conducted listening experiments using Amazon Mechanical Turk (AMT) in the proposal of the attack for Kaldi [Povey *et al.*, 2011]. As a result, they reported that only 2.2% of the participants realized that the lyrics had changed from the original songs used as input samples, whereas approximately 65% noticed abnormal noises in the generated adversarial examples.

We similarly conducted listening experiments using AMT in order to confirm whether humans notice an attack.

	Hear anything abnormal	Hear a target phrase	With presentation of the choices listed in Table 4		
			Correct	Incorrect	Not sure
(A)	36.0%	0.0%	4.0%	28.0%	68.0%
(B)	56.0%	0.0%	4.0%	32.0%	64.0%
(C)	48.0%	0.0%	4.0%	24.0%	72.0%
(D)	32.0%	0.0%	4.0%	28.0%	68.0%
(E)	44.0%	0.0%	8.0%	16.0%	76.0%
(F)	48.0%	0.0%	0.0%	32.0%	68.0%

Table 5: Results of the listening experiments of Table 1. Although a certain number of participants felt abnormal, most of the participants could not hear the target phrases, even when presented with choices.

5.1 Settings

We used the six generated adversarial examples (A) through (F) of Table 1, which were recognized as target phrases with a success rate of 100%. We conducted an online survey separately for each adversarial example with 25 participants. They listened to the adversarial example three times, and, after each listening, we asked each of the following questions:

1. Did you hear anything abnormal? (For affirmative responses, we asked them to write what they felt.)
2. *(With the disclosure that some voice is included)* did you hear any words? (For affirmative responses, we asked them to write down the words.)
3. *(With the presentation of eight phrases in Table 4)* which phrase do you believe was included?

5.2 Results

The results are shown in Table 5. Although a certain number of participants felt abnormal, no one could hear the target phrases in all cases.

In detail, for example, only 32% of the participants felt abnormal about Table 5(D) and provided comments like, "It was not very clear," "The music seemed a bit fuzzy," and "It sounded like birds in the background." Although Table 5(B) showed the highest rate of the participants feeling abnormal, only comments similar to (D), such as, "It was like hearing over a bad Skype connection or phone call," were provided. For (D) through (F) of Table 5, which are generated on the song by Owl City, we found more comments related to the sound quality, such as, "It sounds highly compressed," compared to the case of Bach's Cello Suite. However, an indication of any messages or utterances was not available in all cases.

Furthermore, even when presented with choices for the target phrases, more than half of the participants responded, "I could not catch anything." In particular, no one could choose the correct choice, even though seven participants chose the incorrect choices for Table 5(F). Moreover, in the other five adversarial examples, only one or two participants chose the target phrase correctly. Note that these results were obtained under the condition in which we explicitly instructed the participants to listen for the adversarial examples and presented them with choices for the target phrases. Thus, we believe this result does not deter the attack scenario, which usually seeks a situation that is less likely to be noticed.

Based on the above considerations, we conclude that the generated adversarial examples sound like mere noise and are almost unnoticeable to humans, which can be a real threat. In addition, the obtained comments suggest directions for future investigation of attack scenarios. For example, we might able to use birdsong as the input samples or play the samples through a telephone to make adversarial examples more difficult to notice.

6 Conclusion

In this research, we proposed a method by which to generate audio adversarial example targeting the state-of-the-art speech recognition model that can attack practically in the physical world. We were able to generate such robust adversarial examples by introducing a band-pass filter, impulse response, and white Gaussian noise to the generation process in order to simulate the transformations caused by the over-the-air playback. In the evaluation, we confirmed that the adversarial examples generated by the proposed method can have smaller perturbations than the conventional method, which cannot deal with recurrent networks. Moreover, the results of listening experiments confirmed that the obtained adversarial examples are almost unnoticeable to humans. To the best of our knowledge, this is the first approach to successfully generate audio adversarial examples for speech recognition models that use a recurrent network in the physical world.

In the future, we would like to examine a detailed attack scenario and possible defense methods regarding the generated audio adversarial examples. We would also like to consider the possibility of realizing a robust speech recognition model using adversarial training, as discussed for the image classification [Goodfellow *et al.*, 2015].

Acknowledgments

This work was partly supported by KAKENHI (Grants-in-Aid for scientific research) Grant Numbers JP19H04164 and JP18H04099.

References

[Alzantot *et al.*, 2018] Moustafa Alzantot, Bharathan Balaji, and Mani B. Srivastava. Did you hear that? adversarial examples against automatic speech recognition. In *Proceedings of the 2017 NIPS Workshop on Machine Deception*, pages 1–6, 2018.

[Amodei *et al.*, 2016] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Han, Patrick LeGresley, Xiangang Li, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Sheng Qian, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Chong Wang, Yi Wang, Zhiqian Wang, Bo Xiao, Yan Xie, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International*

Conference on Machine Learning, volume 48, pages 173–182, 2016.

[Athalye *et al.*, 2018] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35rd International Conference on Machine Learning*, pages 284–293, 2018.

[Carlini and Wagner, 2018] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *Proceedings of the 1st IEEE Workshop on Deep Learning and Security*, pages 1–7, 2018.

[Cissé *et al.*, 2017] Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pages 6980–6990, 2017.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–11, 2015.

[Hannun *et al.*, 2014] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv*, 1412.5567:1–12, 2014.

[Hansen and Pellom, 1998] John H. L. Hansen and Bryan L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the 1998 International Conference on Spoken Language Processing*, pages 2819–2822, 1998.

[Härmä, 2001] Aki Härmä. Acoustic measurement data from the varechoic chamber. Technical report, Agere Systems, 2001.

[Jeub *et al.*, 2009] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of the 16th International Conference on Digital Signal Processing*, pages 1–5, July 2009.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–13, 2015.

[Kinoshita *et al.*, 2013] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas. The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech. In *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, 2013.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Nakamura *et al.*, 2000] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of the 2nd Language Resources and Evaluation Conference*, pages 965–968, 2000.

[Peddinti *et al.*, 2015] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pages 2440–2444, 2015.

[Povey *et al.*, 2011] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 1–4, 2011.

[Sainath and Parada, 2015] Tara N. Sainath and Carolina Parada. Convolutional neural networks for small-footprint keyword spotting. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pages 1478–1482, 2015.

[Schönherr *et al.*, 2018] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv*, 1808.05665:1–18, 2018.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, pages 1–10, 2014.

[Taori *et al.*, 2018] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. Targeted adversarial examples for black box audio systems. *arXiv*, 1805.07820:1–9, 2018.

[Wen *et al.*, 2006] Jimi Y. C. Wen, Nikolay D. Gaubitch, Emanuël A. P. Habets, Tony Myatt, and Patrick A. Naylor. Evaluation of speech dereverberation algorithms using the mardy database. In *Proceedings of the 10th International Workshop on Acoustic Signal Enhancement*, pages 1–4, 2006.

[Yuan *et al.*, 2018] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In *Proceedings of the 27th USENIX Security Symposium*, pages 49–64, 2018.