TOWARDS MITIGATING AUDIO ADVERSARIAL EXAMPLES

Zhuolin Yang
Shanghai Jiao Tong University

Bo Li
UC, Berkeley

Pin-Yu Chen
IBM Research

Dawn Song
UC, Berkeley

ABSTRACT

Audio adversarial examples targeting at automatic speech recognition systems have recently been made possible in different tasks, such as speech-to-text translation and speech classification. Here we aim to explore the robustness of these audio adversarial examples generated via various attack strategies by applying different defensive methods to recover the original audio sequence. In addition, we also show that by inspecting the temporal consistency in speech signals, we can detect audio adversarial examples with above 91% success rate.

1 INTRODUCTION

Deep Neural Networks (DNNs) have been widely adopted in a variety of applications (Krizhevsky et al., 2012; Hinton et al., 2012; Levine et al., 2016). However, recent work has demonstrated that DNNs are vulnerable to adversarial perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). An adversary can add negligible perturbations to inputs and generate adversarial examples to mislead DNNs, especially in image-based machine learning tasks (Goodfellow et al., 2015; Carlini & Wagner, 2017; Liu et al., 2017; Chen et al., 2017b;a)

Beyond images, given the wide application of DNN-based audio recognition systems, such as google home and Amazon Alexa, adversarial examples have also been studied recently (Carlini & Wagner, 2018; Alzantot et al., 2018; Cisse et al., 2017). In this paper, we aim to explore the robustness of such audio adversarial examples against different defensive recovery methods. As a first attemp towards mitigating audio adversarial examples, we conduct extensive experiments to show that with different baseline defenses, such as signal quantization, local smoothing, and down-sampling, the adversarial examples can be recovered with high-quality recognition results without affecting the benign examples too much. In addition, we utilize the temporal consistency of random perturbations to audio adversarial examples, and propose a new detection method based on perturbation integrity. We perform the detection on both Libris (Nixdorf et al., 2010) and CommonVoice (Carlini & Wagner, 2018) datasets against the state-of-the-art attack method (Carlini & Wagner, 2017), and show that our proposed detection methods can achieve above 91% detection success rate.

2 BACKGROUND ON AUDIO ADVERSARIAL EXAMPLES

Generally speaking, an adversarial example is an input x^* that is similar to a natural input x but will yield different output after passing through the neural network. Currently, there are two different types of audio adversarial attacks: Speech-to-label attack and Speech-to-text attack. Speech-to-label attack aims to find an adversarial example x^* close to the original audio x but yields a different (wrong) label. Moustafa Alzantot et al. proposed a genetic algorithm to generate such adversarial examples (Alzantot et al., 2018). In the targeted case, Speech-to-text attack requires the transcribed output of the adversarial audio to be the same as the desired output, which has been made possible by Carlini et el. using optimization-based techniques in (Carlini & Wagner, 2018).

3 DEFENSIVE METHODS TOWARDS MITIGATING ADVERSARIAL NOISES

As a first attempt, we applied some primitive defensive methods on audio data to recover them from adversarial noises without decreasing their quality too much. These defensive methods are useful methods which are easy to implement, fast to operate and achieved some promising results.

- **Quantization:** By rounding the amplitude of audio sampled data into the nearest q-divisible integer number, the adversarial perturbation can be disrupted since its amplitude is usually small in the input space. We choose $q = 128, 256, 512, 1024$ as our parameters.
- **Local smoothing:** We use a sliding window of a fixed length for local smoothing to reduce the adversarial noises. For a sampled audio data x_i, we consider its k-nearest sampled data $x_{i-k+1}, x_{i-k}, \ldots, x_{i+k-1}$ as a local reference sequence and replace x_i by the smoothed value (average, median, etc) based on its reference sequence.
- **Down sampling:** Based on sampling theory, it is possible to down-sample an audio file without sacrificing the quality of the recovered signal while mitigating the adversarial noises in the reconstruction process. In our experiments, we down-sample the original 16kHZ audio data to using 8kHZ frequency and then perform signal recovery.

4 TEMPORAL CONSISTENCY BASED DETECTION

Consider how an adversarial audio is generated: produce some unnoticeable noises on the origin audio elaborately to ensure every targeted phrase can be transcribed correctly. In another aspect, we argue that adversarial audio can be fragile since it needs complete audio information to resolve temporal dependency. If we cut the audio data into two parts, the natural audio data should be transcribed normally in each part despite of some phrases near the cut position. However, adversarial audio may not have this feature since the success of adversarial noises relies on the sequential nature of speech, rendering it inherently vulnerable to changes in temporal consistency of the input.

Consequently, we proposed a temporal-consistency-based approach to detect adversarial audios. Let x be the input audio and x^* be the first half part of x. Let $f(\cdot)$ denote the speech recognizing and transcribing process. To measure the similarity between $f(x)$ and $f(x^*)$ (they usually have different lengths), we only compare the prefix of $f(x)$ that has the same length as $f^*(x)$ and ignore the rest. We use word error rate (WER), character error rate (CER) and longest common prefix ratio (LCP ratio) to evaluate the differences between $f^*(x)$ and $f(x^*)$. We will show that all of these metrics can effectively differentiate the adversarial examples and achieve high AUC scores.

5 EXPERIMENTAL RESULTS AND DISCUSSION

5.1 ANALYSIS FOR DEFENSES

In our experiments, we measure the robustness of audio adversarial attacks by Carlini et al. (Carlini & Wagner, 2018) and Moustafa Alzantot et al. (Alzantot et al., 2018). For Carlini's attack, we separately choose 50 audio files from two audio dataset (CommonVoice, Librispeech) and generate Carlini's attacks based on CTC-loss since C&W-loss based attack is easier to defend against. We evaluate our defensive methods based on WER and CER metrics and also report the ratio between

Table 1: Defense behavior on Commonvoice

Defense Methods	OriginWER(%)	OriginCER(%)	AdvWER(%)	AdvCER(%)
Without defense	37.7 (0.18)	18.5 (0.10)	95.8 (10.1)	83.0 (31.6)
Median-4	43.4 (0.21)	20.4 (0.11)	83.0 (0.46)	46.5 (0.34)
Downsample	47.2 (0.23)	23.3 (0.13)	77.6 (0.43)	43.9 (0.31)
Quantization-128	47.3 (0.24)	25.7 (0.15)	80.7 (0.52)	49.0 (0.40)
Quantization-256	52.5 (0.28)	29.2 (0.18)	73.4 (0.43)	43.6 (0.31)
Quantization-512	64.1 (0.35)	37.5 (0.24)	73.7 (0.43)	44.2 (0.31)
Quantization-1024	72.1 (0.42)	50.4 (0.37)	76.9 (0.45)	53.0 (0.41)

Table 2: Examples of defense/detection approaches

Type	Decode results
Adv.	enough said the boy
Origin	they snuffed about the fur tree and rustled among the branches
Defensed	they snuffed about the vertre and ruslled amon the branches
Origin	i would fe no if i am destined for so glorious a career cried the tree rejoicing
the first half of Origin	i would fe no if i am destined for so
Adv.	this is an adversarial example
the first half of Adv.	th is is an adverso

Table 3: AUC Results of Detection

Dataset	LSTM	TC(WER)	TC(CER)	TC(LCP ratio)
Commonvoice	0.712	**0.936**	0.916	0.859
Librispeech	0.645	**0.911**	0.902	0.729

the transcribing error rate referenced by ground truth and the adversarial target (the ratio is shown in brackets). Results are shown in Table 1 and 4. From the results we can see that most of our defensive methods (e.g., Median-4, Downsample and Quantization-256) can effectively reduce the adversarial noises without affecting the origin audio's transcribed results too much. Table 2 are some attack examples defended by Quantization-256. It shown that some mistakes caused by some similar phrase replacement or some spell errors which is a kind of reasonable. In order to obtain the better transcribed results which can fixed some reasonable mistakes, we used DeepSpeech's language model. The results are mentioned in Table 6, 7.

In Moustafa Alzantot's attack, we implemented their attack with 500 iterations and # of adversarial change limited to 5, then generated 50 adversarial examples on each attack tasks as shown in Figures 1, 2. The attack success rate is 84% on the labeled audio datasets, and the average unchanged label rate is 10.6%. We use the best defensive method as mentioned before as our defense. The defend results are very promising as shown in Fig 3, 4. The attack successful rates decreased to only 2.1%, and 63.8% adversarial audio is converted back to their original label. We also measure the possible affects on origin audio files caused by our defensive methods: the origin audio without defense can be classified correctly with rates 89.2% and the rates decreased to 89.0% after applying our defensive method, which means the effects of defense on benign instances is negligible.

5.2 Analysis for Detection

We start our detect experiment based on Carlini's CTC-loss attack and used 100 randomly picked audiofiles equally from CommonVoice and Librispeech audio dataset. Some detect samples are shown in Table 2. As you can see, while transcribing the sentence, the first half part of an natural audio changed a little comparing to the ground truth, but for the first part of adversarial audios, it changed a lot. Also this method can be effective even the adversarial audio hide speech from being transcribed: it revealed the underlying phrases.

We used three metric including WER, CER, LCP Rates mentioned before. Results from Table 9, 10 showed the differences between natural audio and adversarial audio under those three metrics. So We can use these three metrics as our detection model and evaluated its performance by using AUC score together with a trained LSTM which has 64 hidden layer features as our baseline model. The comparison was shown on Table 3. As you can see, the LSTM model' performance is not good enough and the WER metric achieved 0.936 on CommonVoice and 0.911 on Librispeech, which proved to be an simple but effective detect method.

REFERENCES

Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy, 2017*, 2017.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. *arXiv preprint arXiv:1801.01944*, 2018.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Show-and-fool: Crafting adversarial examples for neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017a.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. *arXiv preprint arXiv:1709.04114*, 2017b.

Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1097–1105, 2012.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 17(39):1–40, 2016.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

Don Nixdorf, Eric Schiffman, Richard Ohrbach, Gary Anderson, and John Look. Proposed revision of the ichd-ii criteria for headache or facial pain attributed to temporomandibular joint (tmj) disorders. In *Abstracts of the 13th World Congress of Pain*. IASP (International Association for the Study of Pain and Omnipress), 2010.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

6 APPENDIX

Table 4: Defense behavior on Librispeech

Defense Methods	OriginWER(%)	OriginCER(%)	AdvWER(%)	AdvCER(%)
Without defense	12.4 (0.07)	7.05 (0.06)	105.3 (65.8)	91.7 (611.3)
Median-4	16.4 (0.09)	8.0 (0.07)	57.9 (0.34)	27.5 (0.26)
Downsample	24.2 (0.13)	13.0 (0.11)	60.9 (0.35)	31.2 (0.28)
Quantization-128	13.4 (0.07)	7.6 (0.06)	66.1 (0.40)	37.1 (0.38)
Quantization-256	**16.3 (0.09)**	**8.9 (0.08)**	**48.6 (0.28)**	**24.0 (0.22)**
Quantization-512	27.5 (0.15)	13.8 (0.12)	47.0 (0.27)	23.0 (0.19)
Quantization-1024	46.8 (0.27)	25.4 (0.23)	52.3 (0.30)	30.0 (0.28)

Table 5: Defend samples

Type	Decode results
Adv.	enough said the boy
Origin	they snuffed about the fur tree and rustled among the branches
Defensed	they snuffed about the vertre and ruslled amon the branches
Origin	and he leaned against the wa lost in reveriey
Defensed	and he leaned against the wall losting reavelete
Origin	and then what happens then
Defensed	as the what happenis them

Table 6: Defense behavior on Commonvoice with language model

Defense Methods	OriginWER(%)	OriginCER(%)	AdvWER(%)	AdvCER(%)
Without defense	27.5 (0.14)	14.3 (0.08)	95.9 (9.13)	80.1 (9.65)
Median-4	27.0 (0.14)	14.6 (0.08)	73.6 (0.59)	42.4 (0.29)
Downsample	31.2 (0.17)	17.6 (0.10)	69.6 (0.53)	41.2 (0.29)
Quant-128	34.4 (0.19)	21.3 (0.12)	75.9 (0.70)	45.3 (0.36)
Quant-256	42.9 (0.25)	26.7 (0.16)	70.7 (0.57)	41.8 (0.29)
Quant-512	52.4 (0.29)	37.1 (0.24)	68.5 (0.40)	45.0 (0.32)
Quant-1024	62.4 (0.36)	47.2 (0.35)	70 (0.41)	51.2 (0.39)

Table 7: Defense behavior on Librispeech with language model

Defense Methods	OriginWER(%)	OriginCER(%)	AdvWER(%)	AdvCER(%)
Without defense	3.05 (0.02)	1.46 (0.01)	102.8 (15.1)	86.5 (16.3)
Median-4	3.6 (0.02)	1.7 (0.01)	35.1 (0.23)	19.0 (0.17)
Downsample	11.8 (0.06)	5.7 (0.05)	41.2 (0.28)	21.8 (0.19)
Quant-128	3.2 (0.02)	1.5 (0.01)	49.7 (0.35)	28.2 (0.22)
Quant-256	**3.5 (0.02)**	**1.7 (0.01)**	**29.1 (0.18)**	**15.4 (0.13)**
Quant-512	12.0 (0.07)	6.6 (0.06)	25.1 (0.15)	13.3 (0.12)
Quant-1024	30.7 (0.19)	20.3 (0.18)	36.6 (0.23)	24.1 (0.22)

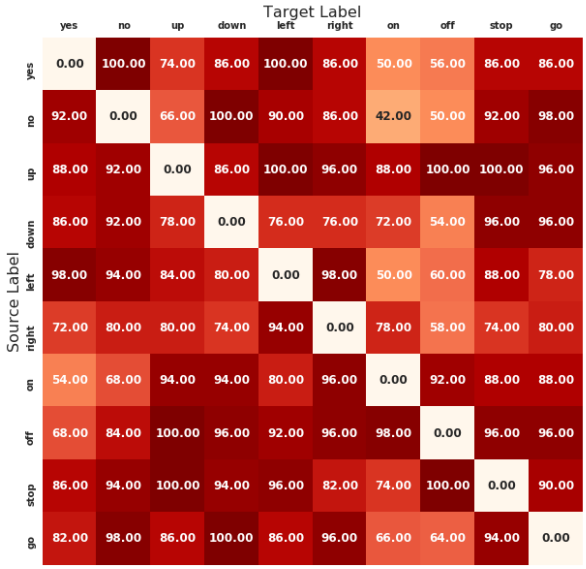

Figure 1: Successful attack rates

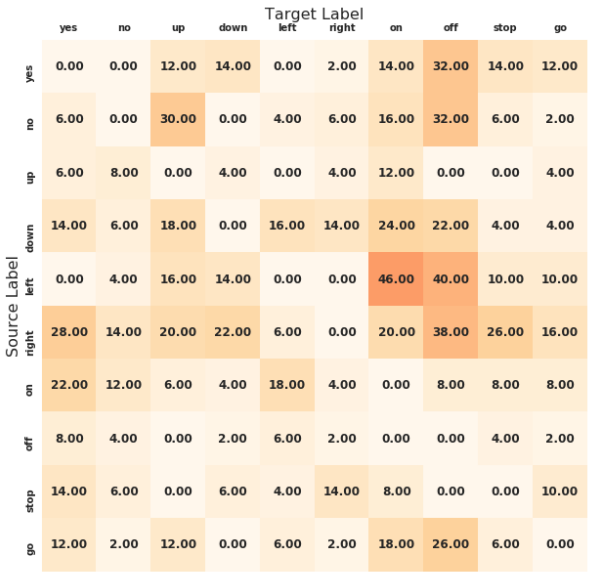

Figure 2: Unchanged label rates

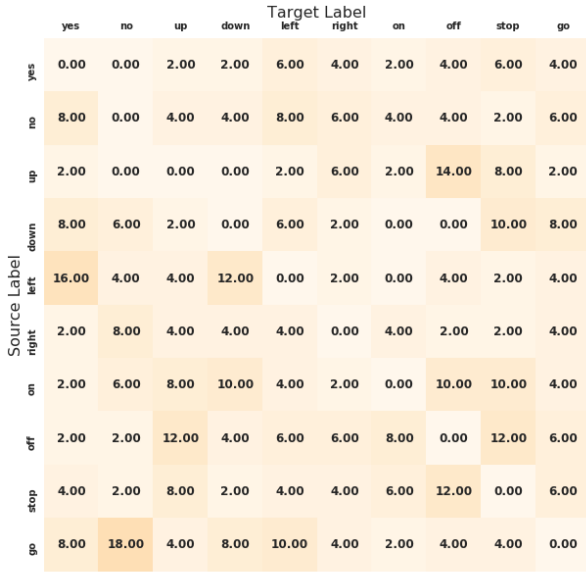

Figure 3: Successful attak rates after defense

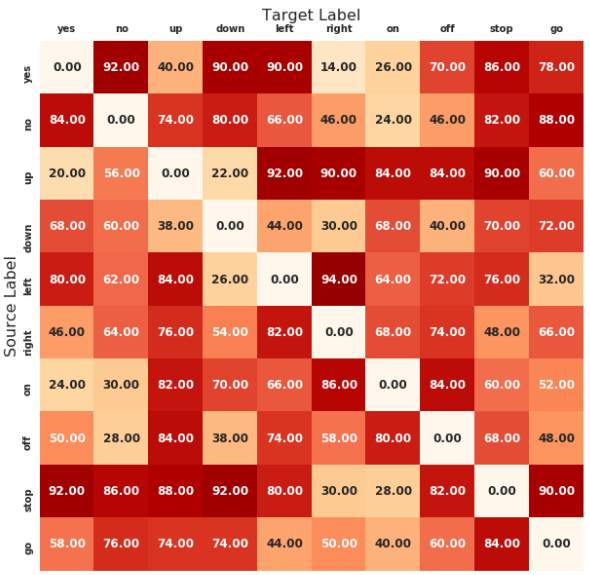

Figure 4: Unchanged label rates after defense

Table 8: Detect samples

Type	Decode results
Origin	i would fe no if i am destined for so glorious a career cried the tree rejoicing
the first half of Origin	i would fe no if i am destined for so
Adv.	this is an adversarial example
the first half of Adv.	th is is an adverso
Origin	have you not met them anywae
the first half of Origin	have you not
Adv.	*sil*
the first half of Adv.	have you not
Origin	this because you were not born
the first half of Origin	this because you were
Adv.	enough said the boy
the first half of Adv.	yens egasia or

Table 9: Detection results on Commonvoice

Metrics	Origin	Adversarial
WER(%)	37.4	96.0
CER(%)	16.2	61.1
LCP ratio(%)	27.3	2.08

Table 10: Detection results on Librispeech

Metrics	Origin	Adversarial
WER(%)	21.6	65.8
CER(%)	5.02	33.4
LCP ratio(%)	38.5	20.3