Multi-Spectrum Based Audio Adversarial Detection

Yunchen Li, Jian Ma, Da Luo

Dongguan University of Technology, Dongguan, China

Abstract

Audio adversarial examples are emerging as a threat to automatic speech recognition (ASR) systems. Existing studies on adversarial examples and defence are mostly developed for image classification. Unlike attack methods, adversarial detection techniques cannot be directly transferred to ASR due to sequence dependency of sound waveforms and relatively less audio adversarial examples for training an adversarial detector. In this paper, we study the spectrum characteristics of audio adversarial examples and accordingly, propose a multi-spectrum based learning scheme to address these problems for audio adversarial detection. We evaluate the ASR dataset under two white-box and one black-box attacks, respectively. Compared with existing methods, our method significantly improves detection accuracy on short audio frames, especially under keyword modification attacks. In addition, through ablation experiments, it is proved that our proposed multi-spectral method achieves good results in audio adversarial detection.

Keywords

audio adversarial detection, automatic speech recognition, multi-spectrum

1. Introduction

Core to the automatic speech recognition (ASR) system is the speech-to-text procedure, it can be deeply influenced by adversarial examples [1].Many modern ASR systems like DeepSpeech [2] and Lingvo [3] use deep neural networks (DNNs), which are vulnerable to input perturbations [4]. It is possible to inject adversarial perturbations into audio segment to change recognition result [5]. There are two main cases. The first one is called the key-word modification, which would changes some key words in the adversarial transcript. The second one is called the sentence modification, in which the entire transcript could be replaced.

Here comes an example.

- Original transcript: I have gave them to Alice.
- Keyword modification: I have gave them to Bob.
- Sentence modification: I did not give to anyone.

The adversarial perturbation is almost imperceptible and is a serious threat to the growing applications of ASR such as Google Home and Amazon Alexa. Therefore, it is important to address the detection problem of audio adversarial examples.

Several ASR attacks are demonstrated in the literature [6–8]. Some of them are adapted to audio systems with key techniques such as gradient descent transferred from the domain of image classification. Unlike attack methods, audio adversarial detection poses different challenges comparing with its counterpart in the image domain. Firstly, it is much slower and more complex to generate audio adversarial examples based on Recurrent Neural Networks (RNN) [9] due to sequence dependency. This causes much less adversarial samples for training a binary classifier as an audio adversarial detector. Moreover, audio input transformation is not effective against adversarial attacks [10]. This is mainly because the data of speech possess temporal dependency that does not have hierarchical object associations.

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control
EMAIL: *15361930042@163.com (Yunchen Li)

To identify audio adversarial examples, Zeng et al. [11] compared audio transcripts by different ASR systems. Yang et al. [10] exploited the inherent temporal dependency in audio data. Jayashankar et al. [12] introduced random dropout to an ASR network at test time to reduce the attack success of adversarial perturbations but at a cost of significant system degradation on clean examples.

In this paper, we propose to study artefact characteristics of audio perturbations in the Mel-scale Frequency Cepstral Coefficients (MFCC) domain and we make the following contributions:

– We found artefact characteristics of adversarial perturbations is hiding in all the frames of audio. So we propose multi-spectrum framework for adversarial detection.

– The proposed approach can significantly improve the accuracy of audio adversarial detection rate especially in the case of black-box attack.

2. Audio Adversarial Attacks

In this paper, we consider two representative attack methods as well as newly proposed attack methods for creating audio adversarial examples: 1) C&W attack, 2) Taori attack and 3) Time limit attack. C&W attack [6] is a white-box attack proposed by Carlini and Wange to generate adversarial perturbation. This method access the parameters of the target ASR and minimize its loss function to tamper the result of ASR. Taori attack [7] is a black-box attack against ASR systems without access to the internal information of the victim ASR. It employs genetic algorithm and gradient estimation method to tamper the result of ASR. Time limit attack is an audio adversarial sample white-box attack method based on time domain limitation, which hides the noise of the audio adversarial sample in the speech information part of the audio, so as to make the speech information part of the audio, so as to make the disturbance noise imperceptible.

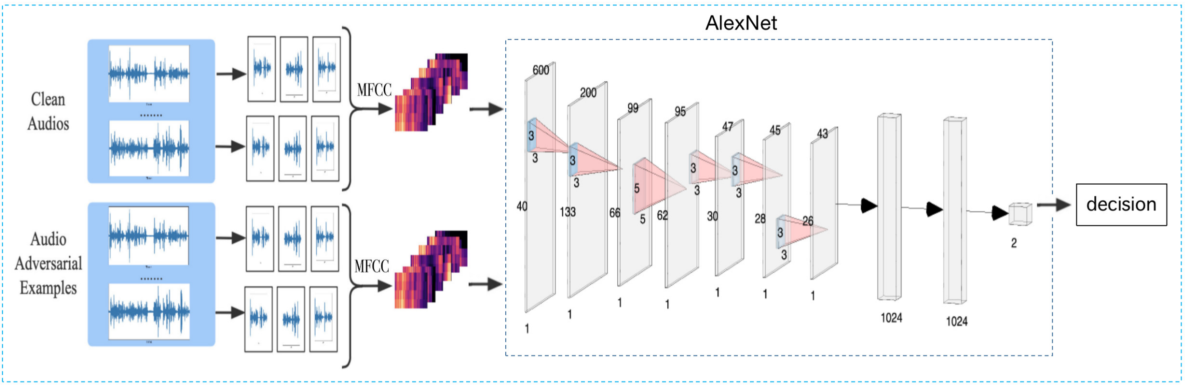


Figure 1 The proposed MS-AlexNet framework.

One of the interesting question is that where the artefact characteristics of audio adversarial perturbation will exist. In modern ASR such as DeepSpeech, bi-directional RNN (B-RNN) [13] is often used to obtain the semantic information of audio data by aligning speech to text information. The input sample $\boldsymbol{x} = \{x_t\}$ is a sequence of utterance and the output $\boldsymbol{y} = \{y_t\}$ is the corresponding transcript at time step t for $t = 0, 1, 2, ..., \mathrm{T}$. In the B-RNN, let the hidden layer activation function be $h(\cdot)$ and the output layer activation function be $g(\cdot)$. The forward and backward hidden states at time t are denoted by A_t and A'_t, respectively. They are updated by

$$A_t = h(U_t x_t + W_t A_{t-1}) \tag{1}$$

$$A'_t = h\big(U'_t x_t + W'_t A'_{t+1}\big) \tag{2}$$

while the model output is computed by

$$y_t = g\big(V_t A_t + V_t{}' A'_t\big) \tag{3}$$

where B-RNN parameters U and U', W and W', V and V' are weights between different layers. It can see that the forward hidden state A_t is determined by the utterance sequence of $x_0, x_1, ...x_t$, while the backward hidden state A'_t is determined by $x_{t+1}, x_{t+2}, ...x_T$. Accordingly, the output y_t depends on all utterance $x_0, x_1, ..., x_T$ in the time series.

In the attack process, define the adversarial perturbation signal as the difference between audio samples before and after speech modification, i.e., $\delta = \{\delta_t\}$ for $\delta_t = \tilde{x}_t$-x_t.

Proposition 1. To temper y_t, the entire sequence of utterance $x_0, x_1, ..., x_t, ..., x_T$ in the time series must be altered, i.e., $\delta_t \neq 0$ for $t = 0, 1, ..., T$.

Proof. Without loss of generality, denote the tempered output by $\tilde{y}_t$. In l_p-normed attacks, the objective function with respect to y_t is therefore

$$\min \|\tilde{y}_t - y_t\|_p \tag{4}$$

for $t = 0, 1, ..., T$. It is common to optimize (4) by gradient desent using equation (3). The updating rule for the gradient-based attack procedures can be generally expressed as

$\tilde{x}_t = x_t - \alpha \frac{\partial CTC\,\mathrm{los}(\tilde{y}_t, y_t)}{\partial x_t}$, since the gradient $\frac{\partial CTC\,\mathrm{Closs}(\tilde{y}_t, y_t)}{\partial x_t} \neq 0$

so we have $\delta_t = \tilde{x}_t - x_t \neq \mathbf{0}$ That is, the adversarial perturbation exists across the entire time series of the original audio sequence.

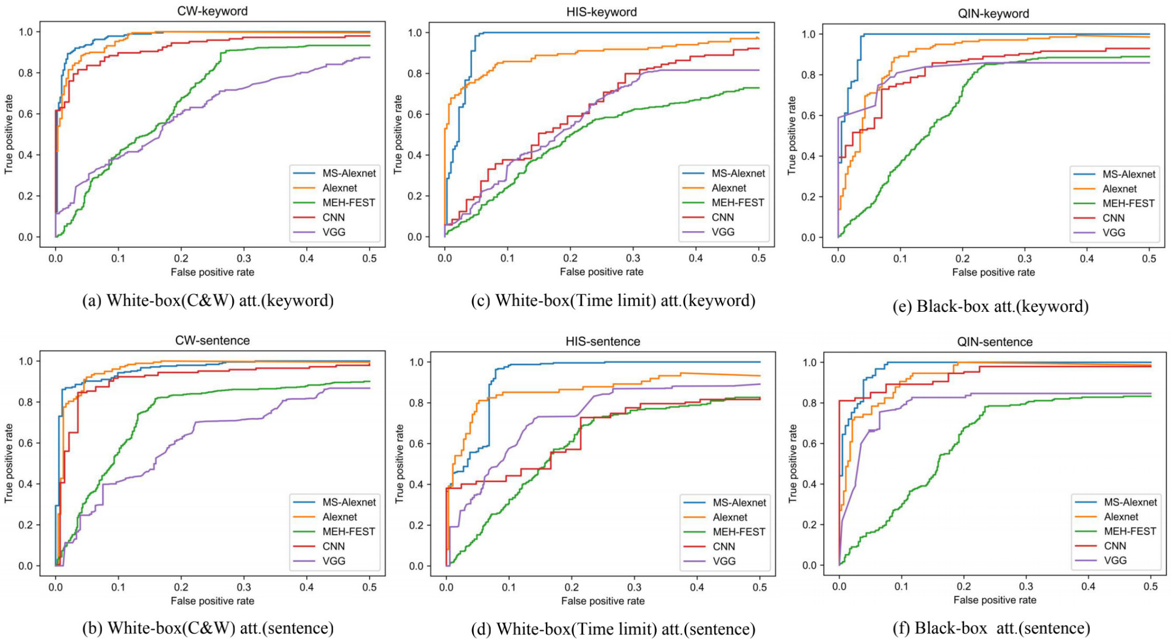
Proposition 1 indicates that the adversarial perturbations generated on BRNN are distributed through the audio time series data. Therefore, we make all ST frames $\{x_t\}$ contain adversarial signals given x is an adversarial audio sample.

3. Multi-Spectrum Based Detection

The adversarial perturbation analysis above suggests that we could discriminate adversarial characteristics based on the short time (ST) frames instead of the entire audio sequence, and for the ST frame we further utilize MFCC features.

There are at least three benefits on designing adversarial detectors in the MFCC domain: 1) The artefacts of adversarial perturbation are more significant in the power spectra of ST frames; 2) Spatio-temporal information is better exploited on the short-time basis to deal with the highly non-stationary perturbation signals, especially when the locations of such noise sources are varying across time as demonstrated in our case; 3) By slicing the speech segment into multiple ST frames there are more adversarial samples for training an effective detector thus relieving the few-shot problem in audio adversarial detection.

In this paper, we proposed MS-AlexNet, which is a detection method combined of the idea of multi-spectrum and AlexNet [14]. The framework is shown in Figure 1. The clean audio and the adversarial audio examples are two categories that we are going to classify. The audios are divided into ST frames and then apply MFCC to obtain multi-spectrum features, which are fed into AlexNet. Specifically, the AlexNet architecture consists of 8 layers in turn (5 convolutional layers and 3 max pooling layers alternate). In the end two fully connected layers are applied and softmax binary classifier will determine the classification result.



(a) White-box(C&W) att.(keyword)

(c) White-box(Time limit) att.(keyword)

(e) Black-box att.(keyword)

(b) White-box(C&W) att.(sentence)

(d) White-box(Time limit) att.(sentence)

(f) Black-box att.(sentence)

Figure 2 Detector ROC under (a-b) white-box(C&W), (c-d) white-box(Time limit) and (e-f) black-box attacks by keyword and sentence modification, respectively.

Table 1. Adversarial detection accuracy (TPR @5% FPR) under white-box and black-box attacks for keyword and sentence modification. The best result for each row is marked in **bold**.

Attack Types		Common Voice				
		AlexNet	VGG	CNN	MEH-FEST	MS-AlexNet
White (C&W)	Keyword	0.89	0.28	0.84	0.22	**0.93**
	Sentence	**0.92**	0.25	0.85	0.33	0.90
White (Time)	Keyword	0.76	0.17	0.19	0.11	**0.98**
	Sentence	**0.79**	0.35	0.41	0.15	0.56
Black	Keyword	0.70	0.65	0.54	0.15	**1.0**
	Sentence	0.74	0.66	0.85	0.16	**0.92**

4. Experiments

We evaluate the adversarial detection accuracy of the proposed method on the open-source Chinese Mandarin speech corpus AISHELL-1 [15], which contains a 150-hour training set, a 10-hour development set and a 5-hour testset. The test set contains 7,176 utterances in total.

We generate adversarial samples using the three attack methods introduced in Section 2. Note that the C&W samples and Time limit samples are white-box attacks and the Taori samples are black-box attacks. To accomplish the black-box attack in limited time, we break long audio sequence into a 10-second series and a 5-second series to generate the Taori samples for Aishell-1. We generate 630 C&W samples, 400 Taori samples and 440 Time limit samples from Aishell-1 datasets for keyword and sentence modification, respectively. DeepSpeech [2] is used as victim ASR where white attacks are deployed on v0.4.1 and black attacks are deployed on v0.1.1. All experiments are performed on a single NVIDIA v100 machine.

We compare the proposed MS-AlexNet method with four different methods. The first method is AlexNet [14] that uses MFCC features directly from the entire audio as input, and AlexNet acts as a binary classifier to detect adversarial examples. The second method is CNN [16], which also uses MFCC to extract features as input. The CNN architecture consists of 5 layers in turn (alternating 3 convolutional layers and 2 max pooling layers). The third approach is to employ a pooling layer in the VGG [17] architecture to aggregate utterance statistics for decision making. The statistical pooling layer aggregates the mean and bias of the output of the last convolutional layer and forwards it to the dense layer. Finally, two dense layers project the statistics into a two-dimensional output space for decision making. The fourth MEH-FEST [18] calculates the minimum energy in high frequencies through the short-time Fourier transform of the audio and uses it as a detection metric.

Figure 2 plots the ROC curves for keyword and sentence modification of the five compared methods under two white-box and one black-box attacks, respectively. In general, black-box attacks are easier to detect than white-box attacks due to the addition of larger perturbation strengths to the signals, especially those used for sentence modification. The proposed MS-AlexNet, significantly outperforms adversarial detection methods of CNN, VGG, and MEH-FEST in most of the case. This demonstrates the effectiveness of audio adversarial detection in the multi-spectrum domain.

In addition, we also conduct ablation experiments, using the AlexNet model for adversarial detection without multi-spectral method, and the results show that our proposed multi-spectral has a good effect on audio adversarial detection. Table 1 displays the detection accuracy of TPR at 5% FPR for keyword and sentence modification. It can be seen that MS-AlexNet performs best under the other attacks, except for the white-box for sentence modification (C&W and Time limit), in fact under the white-box for sentence modification (C&W) AlexNet is comparable to MS-AlexNet. Experiments show that our proposed MS-AlexNet has good results under different attack methods.

5. Conclusions

In this paper, we study the short-time spectrum of audio sequences in the MFCC domain, analyzes audio adversarial characteristics, and, accordingly, propose a multi-spectrogram based detection method for audio adversarial examples against ASR. The proposed method is able to achieve significant improvement of detection accuracy, especially for keyword modification attacks under both white-box and black-box scenarios. Under different attack methods, the multi-spectrum detection method also has a good detection effect.

6. Acknowledgements

This work was financially supported by Guangdong Natural Science Key Field Project (2019KZDZX1008).

7. References

[1] Abramowitz M.: Handbook of mathematical functions, 3rd ed., New York: Dover, 1980

[2] Julkapli N. M.: A brief analysis on the application of wind-proof and dust suppression wall in the open coal storage yard, Electronic production. Vol. 14, 2013, pp. 229

[1] Yu, D., Deng, L.: Automatic Speech Recognition. Springer (2016)

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

[3] Shen, J., Nguyen, P., Wu, Y., Chen, Z., Chen, M.X., Jia, Y., Kannan, A., Sainath, T., Cao, Y., Chiu, C.C., et al.: Lingvo: a modular and scalable framework for sequence-to-sequence modeling. arXiv preprintarXiv:1902.08295 (2019)

[4] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)

[5] Cisse, M.M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. Advances in neural information processing systems 30 (2017)

[6] Carlini, N., Wagner, D.: Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW), IEEE (2018) 1–7

[7] Taori, R., Kamsetty, A., Chu, B., Vemuri, N.: Targeted adversarial examples for black box audio systems. In: 2019 IEEE Security and Privacy Workshops (SPW), IEEE (2019) 15–20

[8] Yakura, H., Sakuma, J.: Robust audio adversarial example for a physical attack. arXiv preprint arXiv:1810.11793 (2018)

[9] Mikolov, T., Karafi´at, M., Burget, L., Cernock`y, J., Khudanpur, S.: Recurrent neural network based language model. In: Interspeech. Volume 2., Makuhari (2010) 1045–1048

[10] Yang, Z., Li, B., Chen, P.Y., Song, D.: Characterizing audio adversarial examples using temporal dependency. arXiv preprint arXiv:1809.10875 (2018)

[11] Zeng, Q., Su, J., Fu, C., Kayas, G., Luo, L., Du, X., Tan, C.C., Wu, J.: A multiversion programming inspired approach to detecting audio adversarial examples. In: 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), IEEE (2019) 39–51

[12] Jayashankar, T., Roux, J.L., Moulin, P.: Detecting audio attacks on asr systems with dropout uncertainty. arXiv preprint arXiv:2006.01906 (2020)

[13] Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing 45(11) (1997) 2673–2681

[14]Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2012)

[15] Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA), IEEE (2017) 1–5

[16] Samizade, S., Tan, Z.H., Shen, C., Guan, X.: Adversarial example detection by classification for deep speech recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2020) 3102–3106

[17] Li, X., Li, N., Zhong, J., Wu, X., Liu, X., Su, D., Yu, D., Meng, H.: Investigating Robustness of Adversarial Samples Detection for Automatic Speaker Verification. In: Proc. Interspeech 2020. (2020) 1540–1544

[18] Chen, Z.: On the detection of adaptive adversarial attacks in speaker verification systems. arXiv preprint arXiv:2202.05725 (2022)