Poster: Propose a Defense Method against Audio Adversarial Attack

Yuya Tarutani*†, Kensuke Ueda‡ and Yoshiaki Kato§
*The Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University
†Cybermedia Center, Osaka University
‡Mitsubishi Electric Corporation Advanced Technology R&D Center
§Mitsubishi Electric Corporation Information Technology R&D Center

***Abstract*—Smart speakers have security risk that attacker can operate the home devices by voice commands. Especially, an audio adversarial attack, which is an attack method that cause misrecognition in the speech recognition system, is the most dangerous. In this paper, we propose a method to detect an audio adversarial attack by using anormaly detecdion method with AutoEncoder. Our method detects attack audio based on the error between input and output of the Auto Encoder. We show the error of attack audio is larger than one of normal audio. In addition, we clarify that our method can detect most of the attack audio using sentence audio.**

I. Introduction

Smart speakers such as Google Home and Amazon Echo are spreading as devices using speech recognition. The smart speakers are used as one of the user operation interfaces in home IoT(Internet of Things) systems. On the other hand, some people concern that smart speakers have vulnerable in terms of security.

Various attack methods with vulnerabilities in deep learning are proposed. In Ref. [1], an attack method has been proposed against DeepSpeech. This approach is called *Audio Adversarial Attack*. The audio adversarial attack causes misrecognition of speech recognition system. The attack audio is generated by adding perturbation to a base audio. The human recognizes this attack audio data as the base audio. Therefore, the attack audio is played without user's knowledge by a video streaming service.

In this paper, we propose a method to detect the audio adversarial attack. Our method uses the anomaly detection method with AutoEncoder trained from normal audio data. We evaluate our method by experiment. Furthermore, we show that our method can detect the attack audio.

II. Audio adversarial attack detection system

Our system prevents unauthorized voice command by classifying attack voice commands and normal voice commands. Figure 1 shows an overview of audio adversarial attack detection system. In our system, we use the abnormality detection method with AutoEncoder.

Our system detects the attack audio by following procedure.

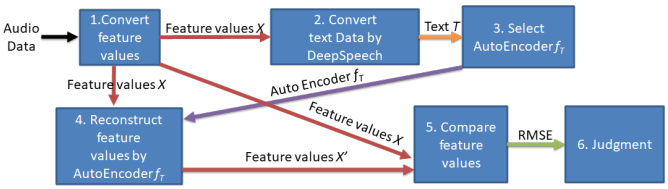


Fig. 1. Overview of audio adversarial attack detection system

1) Convert audio data to feature values X.
2) Convert the feature values X to the text T by using speech recognition system.
3) Select the Auto Encoder f_T trained using only normal data of text T.
4) Input the feature values X to the Auto Encoder f_T to get the reconstructed feature values X'.
5) Calculate the difference between the feature values X and the reconstructed feature values X'.
6) If the difference is higher than or equal to the threshold, the inputting audio is determined as attack audio. Otherwise the inputting audio is determined as normal audio.

The Audio Adversarial Attack makes the speech recognition system misrecognize by adding perturbations to normal audio data. Since the AutoEncoder learns using only normal data, an output from which an abnormal portions of the input is removed can be obtained. Therefore, the feature value reconstructed by the AutoEncoder is a feature value from which the perturbation is removed. In our method, the difference between the feature values X and the reconstructed feature values X' is used to detect attack audio data.

III. Evaluation

We implement the speech recognition system and the attack method to evaluate our method. In this evaluation, we evaluate the difference between the input and output feature values of the AutoEncoder. Then, we show that our system detects the attack audio based on the difference. The attack method [1] is published on github[1]. In this experiment, the dataset provided by Mozilla and Google

[1]https://github.com/carlini/audio_adversarial_examples/

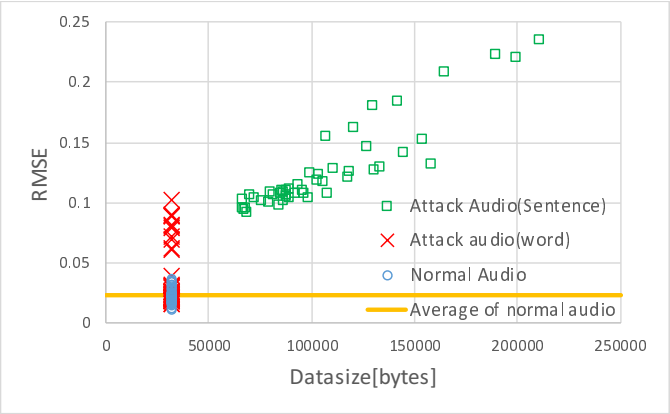


Fig. 2. Relationship between RMSE and datasize

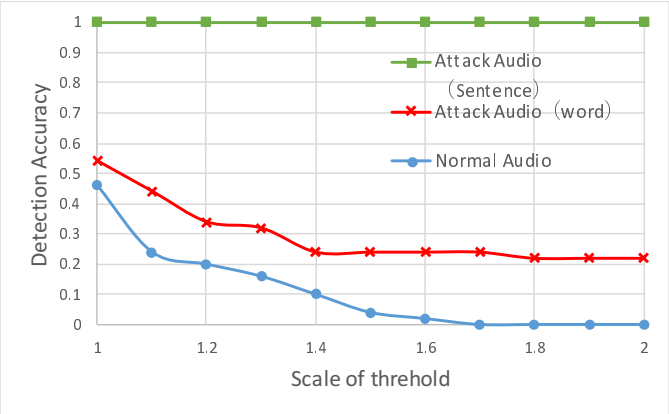


Fig. 3. Detection accuracy

uses. We use 5 word audio in Google dataset and various sentence audio in Mozilla dataset to generate attack audio. The number of word audio is 10 for each audio and the number of sentence audio is 50. The speech recognition system recognizes attack audio as "Yes". Moreover, attack audio sounds like a base audio to humans.

A. Metric

Our system uses the difference between the input feature values X and the output feature values $'X$ of AutoEncoder. The values obtained by normalizing the value of each feature vector with the maximum value and the minimum value of the feature vector are denoted by Y and Y', respectively. Then, the difference is defined as follows:

$$RMSE = \sqrt{\frac{\sum_i (y_i - y_i')^2}{n}} (y_i \in Y, y_i' \in Y') \qquad (1)$$

n is the number of dimension of feature values.

B. Result of RMSE

Figure 2 shows the relationship between RMSE and data size of audio. According to Fig. 2, the RMSE using attack audio is higher than the RMSE using normal audio. The AutoEncoder is trained with normal audio. Therefore, the output feature values are close to the feature values of normal audio. On the other hands, the feature values of attack audio are close to the feature values of the base audio. Therefore, the RMSE of attack audio is larger than one of normal audio. Moreover, the RMSE using the sentence audio is larger than the RMSE using the word audio. This is because the length and shape of the waveform are very different. As the result, the feature values of the attack audio using the sentence audio is very different from the feature values of the normal audio.

C. Detection accuracy

According to subsection III-B, the RMSE with the attack audio is larger than the RMSE with the normal audio. Therefore, the attack audio is detected by threshold based on the RMSE with normal audio. In this evaluation, we set to threshold based on the average of the RMSE with normal audio. The threshold is changed from 1 to 2 times the average of the RMSE with normal audio.

Figure 3 shows the detection accuracy of our method. Each audio is the same data in subsection III-B. According to Fig. 3, our system can detect the attack with the sentence audio. This is because, the RMSE with the sentence audio is much larger than the RMSE with the normal audio. On the other hand, the detection accuracy of the attack audio with the word audio is about 30%. This is because it is difficult to make a difference in the word audio. In addition, according to Fig. 3, by setting the threshold to 1.7 times the average value of the RMSE with normal audio, all normal audio can be determined as normal audio.

The attack audio is mixed with audio from TV, radio, and video streaming services. In this case, the attack audio is generated based on the sentence audio. Therefore, our detection system is effective as a countermeasure against audio adversarial attack. On the other hand, our system cannot detect some attack audio with the word audio. Differences in feature values of word audio are not large. For this reason, the differences in the attack audio generated based on word audio is very small. In order to detect these attack audio, another metric is required. The solution of this problem is future work.

IV. Conclusion

In this paper, we proposed an audio adversarial attack detection system. Our method detected the attack audio based on the difference between the feature values and the reconstructing feature values. We showed that our method detects the attack audio generated based on sentence audio. However, some attack data generated based on word audio data cannot be detected. Therefore, the solution of this problem is future work.

References

[1] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proceedings of 2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, IEEE, 2018.

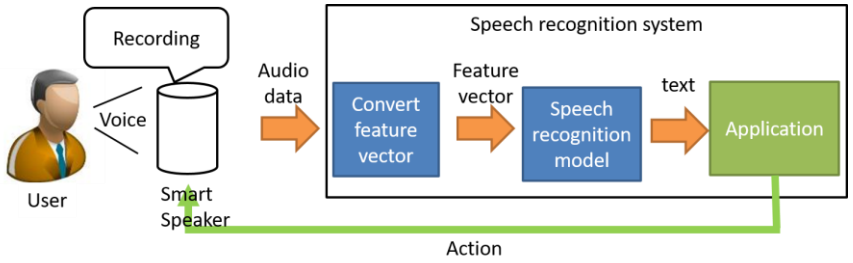
Propose a Defense Method against Audio Adversarial Attack

Yuya Tarutani
Okayama University

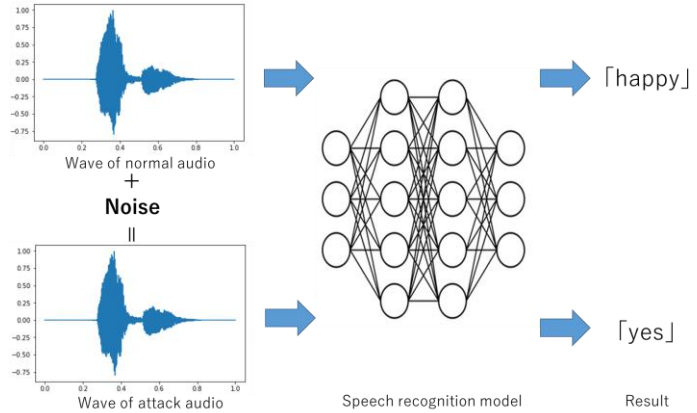
Kensuke Ueda
Mitsubishi Electric

Yoshiaki Kato
Mitsubishi Electric

Security Risk of Speech Recognition System



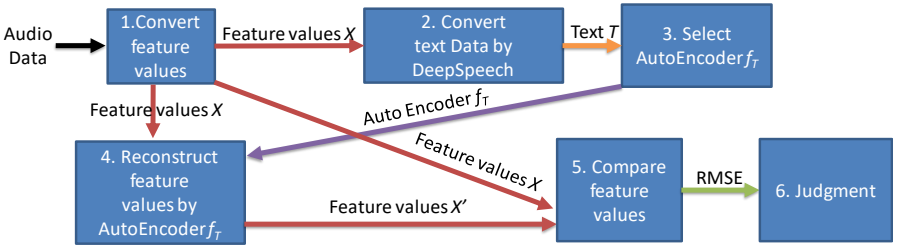
- Smart speaker are spreading as devices using speech recognition system
- Smart speaker have vulnerable in term of security
  - Various attack methods are proposed



- Audio Adversarial Attack causes misrecognition of speech recognition system
  - Attack audio is played without user's knowledge by a video streaming

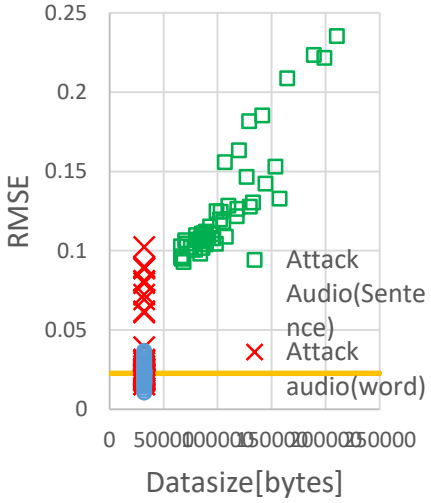
Goal **Propose a defense method against Audio Adversarial Attack**

Audio Adversarial Attack Detection System

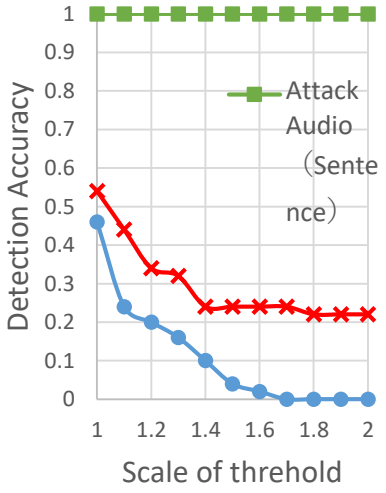


- Our method use the abnormality detection method with AutoEncoder
- Detect the attack audio based on RMSE between input and output of AutoEncoder
  - Threshold is average of RMSE of normal audio

Result of our method



- Relationship RMSE and data size



- Detection accuracy