The Twenty-Fifth International Symposium on Artificial Life and Robotics 2020 (AROB 25th 2020),
B-Con Plaza, Beppu, Japan, January 22-24, 2020

Robust adversarial example generation in speech recognition using evolutionary multi-objective optimization

Shoma Ishida[1], Satoshi Ono[1]

Abstract

In recent years, Automatic Speech Recognition (ASR) systems are widely used in many products such as personal assistants of smartphones (e.g. Amazon Alexa and Apple Siri), voice command technologies in cars, and so on. On the other hand, deep learning methods are known to be vulnerable to adversarial examples, a small perturbation added to a target sample. It is essential that ASR systems have high security because ASR systems perform various tasks which may require user personal data. Therefore, studies have been conducted to generate adversarial examples to evaluate and improve the robustness of ASR systems. A few studies attempted to attack the ASR systems under black-box condition where classification result (class labels) and its confidence are available but internal information is not [1]. The black-box attack can be applicable consumer ASR systems and it is expected to find the vulnerabilities of the consumer systems. On the other hand, to discover more serious vulnerabilities in the real world, it is indispensable to design robust perturbations against environmental changes, time gap between the target speech and perturbation, and so on.

Therefore, this paper proposes a method for generating adversarial examples to ASR systems that are robust against time difference because, in the actual environment, it is difficult to play the perturbation noise accurately in time with the target speech. In the proposed method, robust adversarial example design is formulated as a multi-objective optimization problem, and an evolutionary multi-objective optimization algorithm solves the problem. The proposed method can be applicable commercial systems because it assumes the black box setting. Experiments using Speech Commands classification model [6] showed the effectiveness of the proposed method compared to the conventional method in terms of the robustness against the time difference

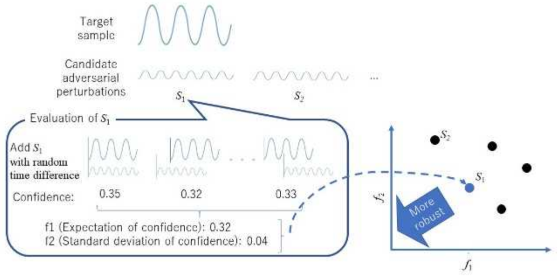
Target sample
Candidate adversarial perturbations
S_1
S_2
...
Evaluation of S_1
Add S_1 with random time difference
Confidence:
0.35
0.32
0.33
f1 (Expectation of confidence): 0.32
f2 (Standard deviation of confidence): 0.04
S_2
S_1
More robust
f_1
f_2

Figure 1. Robust adversarial example generation by multi-objective optimization

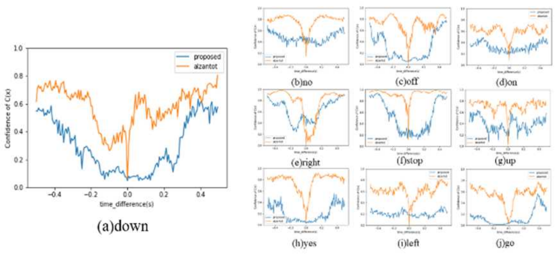

Figure 2. Comparison on the robustness against timing lag.

References

1. M. Alzantot, B. Balaji, M. Srivastava, Did you hear that? adversarial examples against automatic speech recognition, arXiv preprint arXiv:1801.00554 (2018)
2. T. Sainath, C. Parada. Convolutional neural networks for small-footprint keyword spotting, Sixteenth Annual Conference of the International Speech Communication Association (2015).

[1] Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University, 890-0065, Kagoshima, Japan