5ZA-03

Audio Adversarial Examples に対する 動的リサンプリング法とノイズ除去法による防御

尾曲　晃忠†　　田村　慶一††　　橋田　修一††

広島市立大学情報科学部†　　広島市立大学大学院情報科学研究科††

1　はじめに

近年，一般家庭に設置された AI スピーカやスマートフォンなど，音声認識器の利活用が注目されている．今後，公共の場所で使用される機会も増えていくことが考えられ，音声認識器のセキュリティ対策が重要となっている．音声認識器に対する攻撃手法として，Audio Adversarial Examples[1]を使用した攻撃が示されている．本研究では Audio Adversarial Examples を対象とした防御手法を提案する．

2　Audio Adversarial Examples

Audio Adversarial Examples とは，音声に人工的に摂動を加えることで音声認識器の認識結果を変化させる手法である．摂動が加えられた前後の音声を比較しても異音に気付くことは安易では無く，異音に気付いたとしても認識結果に変化が生じることに気付くことは難しい．例えば，人間には挨拶に聞こえる音声が音声認識器には異なる文に認識され，知らない間に悪意のある攻撃を受けてしまう．なお，Audio Adversarial Examples は DeepSpeech に代表される深層学習を用いた Speech-to-Text システムを対象としている．また矢倉らの研究[2]によって，攻撃場所の反響やノイズの影響を考慮することで実世界において攻撃が可能であることが明らかになっている．

3　提案手法

本研究では Audio Adversarial Examples の防御手法として，周波数ベースで防御をする動的リサンプリング法，振幅ベースで防御をするノイズ除去法を提案する．Audio Adversarial Examples は人工的な摂動を加える手法であるため，元音声の振幅や周波数は小さく変化する．この点に着目し，防御手法を検討した．なお，提案手法は Carlini ら[1]の Audio Adversarial Examples を対象としているため，深層学習を用いた Speech-to-Text システムが防御対象である．動的リサンプリング法とノイズ除去法では，最初に Speech-to-Text システムにより認識結果を導出する．これと並行して動的リサンプリング後やノイズ除去後の音声に対して同様に Speech-to-Text システムを使用して認識結果を導出する．最後に，導出された 2 つの認識結果を比較することで人工的な摂動が加えられているか判断する．認識結果の比較には CER を使用する． CER が v 以下であれば正常な音声と判断をして受け入れ，v より大きければ攻撃と判断して拒否する．CER とは S_1, S_2 を比較対象とすると，CER = (編集数) / (S_1 の文字数)によって求められる二つの文章の編集距離のことである．図 1 に攻撃を検知した場合の例を示す．また，図 1 で使用する v は 0.4 とする．

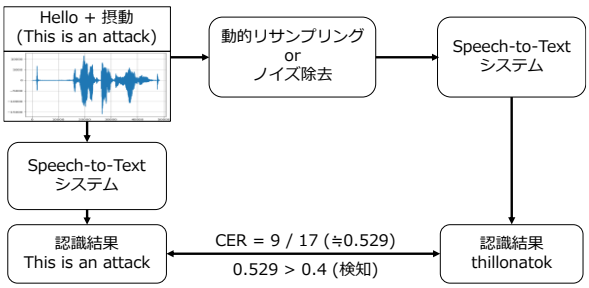

図 1：攻撃を検知した場合の例

3.1　動的リサンプリング法

動的リサンプリング法では，最初に入力音声を固定長の区間で n 個に分割し，各区間を f_i(Hz)でダウンサンプリングをする．次にダウンサンプリングした区間を結合し，入力音声とダウンサンプリング後の音声の認識結果から Audio Adversarial Example を検知する手法である．f_i は音声の入力毎に乱数によって決定するため，攻撃者はサンプリングレートを事前に知ることは困難で，サンプリングレートを考慮した摂動を加えるのは難しい．

3.2　ノイズ除去法

ノイズ除去法とは入力音声に加えられた摂動をノイズとみなして除去し，入力音声と除去後の音声の認識結果から Audio Adversarial Example を検知する手法である．よって，この手法はノイズの決定方法が重要である．音声認識器にマイクから命令を入力して録音するとその音声内には振幅が小さいほど，より多く収録されることになる．小さな音はコマンドを認識する際に重要ではないことが多い．しかし，Audio Adversarial Examples は既存の音声に対して小さな摂動を加えるものである．よって，攻撃が成

Protecting against Audio Adversarial Examples with Dynamic Resampling and Noise Removal Method
†Omagari Akitada, Hiroshima City University
††Tamura Keiichi, Hiroshima City University
††Hashida Shuichi, Hiroshima City University

功するには小さな音がとても重要になる．これに基づき提案するノイズ除去法のノイズの決定方法を考案した．ノイズ除去法のアルゴリズムを図 2 に示す．なお，振幅を 10 等分したときの範囲で n 番目に小さな範囲を r_n とする．また，r_n にプロットされている点の数を $C(r_n)$ とする．

```
音声波形の振幅を0から最大振幅の範囲で10等分(r1 … r10の生成)
while(C(r1) ≧ C(r2) × 2)
        r1 = r1 / 10

n = 10
while C(rn) − C(rn − 1) ≦ C(rn − 1) − C(rn − 2)
        n = n − 1

# ノイズ除去
rxの範囲内である振幅を0にする(1 ≦ x ≦ n - 1)
```

図 2：ノイズ除去法のアルゴリズム

4　評価実験

4.1 実験内容

評価実験では，動的リサンプリング法，ノイズ除去法の誤検知率と見逃し率の比較と実行時間の比較を行う．また DeepSpeech を音声認識器として使用する．評価実験に使用したデータセットは Mozilla Common Voice から 33 サンプル，LibriSpeech から 33 サンプル，筆者の声から録音した 34 サンプルの合計 100 サンプルを用意した．また用意したサンプルに Audio Adversarial Examples[1]で摂動を加えた音声データを用意した．

4.2 CER の基準値と性能の関係

本評価では CER の基準値を 0.1 刻みで変更させ，動的リサンプリング法，ノイズ除去法それぞれの誤検知率と見逃し率の比較を行った．動的リサンプリング法とノイズ除去法の評価実験の結果をそれぞれ表 1，表 2 に示す．

表 1：実験結果(動的リサンプリング法)

CER 基準値	0.4	0.5	0.6	0.7	0.8
誤検知率	0.111	0.077	0.048	0.044	0.04
見逃し率	0	0.003	0.004	0.009	0.026

表 2：実験結果(ノイズ除去法)

CER 基準値	0.4	0.5	0.6	0.7	0.8
誤検知率	0.15	0.15	0.12	0.11	0.11
見逃し率	0.03	0.04	0.04	0.04	0.04

実験結果より，動的リサンプリング法，ノイズ除去法が Audio Adversarial Example を用いた攻撃に対する防御として有効であり，特に動的リサンプリング法がより有効であることが分かる．ノイズ除去法については今回のノイズの決定方法のアルゴリズムが有効であることは分かるが，図 2 の $C(r_1) \geqq C(r_2) \times 2$ という部分をより最適にすることでより有効になる可能性がある．また，動的リサンプリング法において必ず攻撃を防ぐためには見逃し率が 0 になる(CER の基準値) ≦ 0.4 が望まれるが，円滑な使用感を目指す場合は誤検知率が小さい方が良いため CER の基準値が大きい方が良いことが分かる．また，動的リサンプリング法のパラメータである f_i と n を最適化することでより有効になる可能性がある．

4.3 実行時間計測

動的リサンプリング法とノイズ除去法の実行時間を計測した．なお，Speech-To-text システムの実行時間は含まない．計測環境として，OS は Ubuntu 18.04 LTS，CPU は Intel Xeon E3-1270 V5 を用いた．計測結果を図 3 に示す．なお，動的リサンプリング法の分割数は 10 とした．

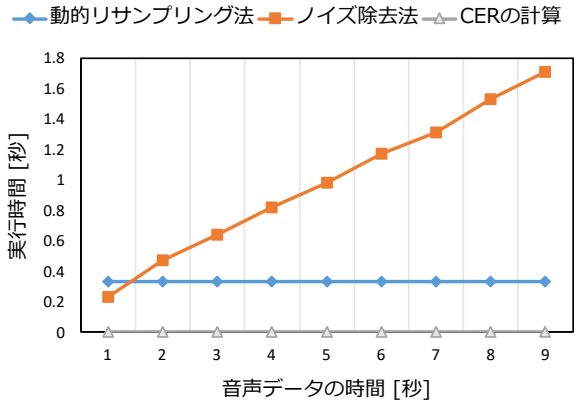

図 3:実験結果

図 3 より，動的リサンプリング法が比較的実行時間が小さく，実用的である．また，CER を導出する時間はとても短く，提案手法導入による使用感への影響は少ないことが分かる．

5　まとめ

本論文では，音声認識器に対する攻撃の一種である Audio Adversarial Examples に対する防御法として動的リサンプリング法とノイズ除去法を提案し，評価実験の結果，有効性を検証できた．特に，動的リサンプリング法の方が効果的であることが分かった．

謝辞

本研究の一部は，JSPS 科研費 JP18K11320，広島市立大学・特定研究費とサタケ技術振興財団の助成により行われた．

参考文献

[1] Nicholas Carlini and David Wagner，“Audio Adversarial Examples:Targeted Attacks on Speech-to-Text”，Deep Learning and Security Workshop，pp. 1-7，2018.

[2] 矢倉 大夢，佐久間 淳，“実世界でも攻撃可能な Audio Adversarial Example”，Computer Security Symposium，pp. 217-224，2018.