SecHack365

漸進的音声認識システムに対する実時間敵対的攻撃

研究駆動コース 二又航介

❶ 背景

音声認識システムに対する敵対的攻撃

システムにより認識される結果を任意の内容に変更する摂動(ノイズ)を付与する攻撃手法[1]

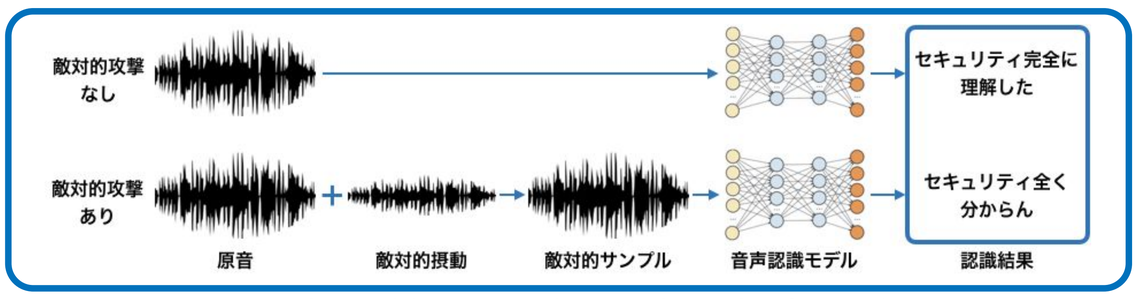

図 1. 敵対的攻撃により認識結果を任意の内容に変更する例

[1] Nicholas Carlini and David Wagner, Audio Adversarial Examples, 2018.

漸進的音声認識システム(ISR)

使用者の発話内容を漸進的に文字として書き起こす音声認識システム(ASR)[2]

- 長い音声波形を小さなブロック単位で音声認識

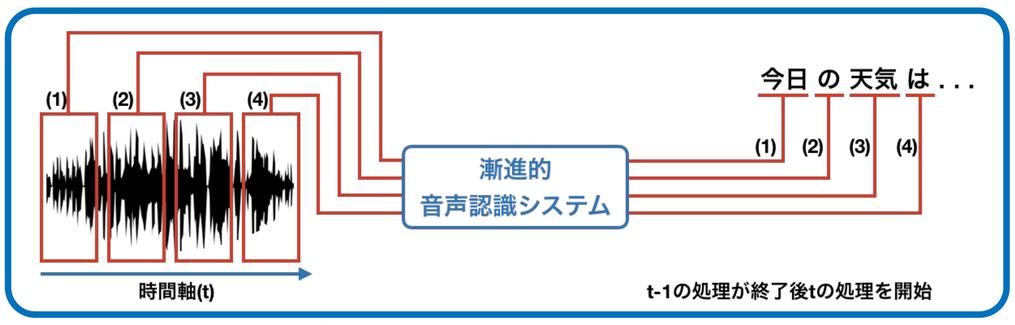

図 2. 漸進的音声認識システムによる音声波形の書き起こし例

[2] K. Hwang and W. Sung, Character-level incremental speech recognition with recurrent neural networks, 2016.

本研究の目的

ISRに対する実時間敵対的攻撃手法の提案

- 既存研究ではASRの短系列のみ対象
- ISRの認識結果は即座に出力されるため短系列として長系列に対する攻撃手法を提案

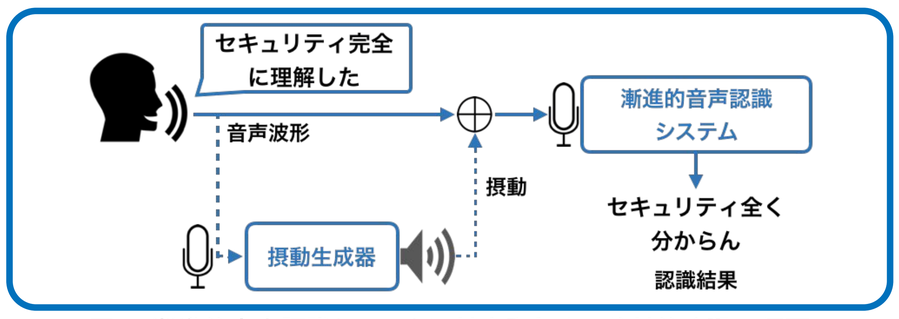

図 3. 漸進的音声認識システムに対する実時間敵対的攻撃の概要

❷ 関連研究

実時間敵対的攻撃

未観測の系列データに摂動を付与することで実時間で敵対的攻撃を行う[3]

- 非実時間の敵対的攻撃手法を教師として模倣学習によって実時間摂動生成器を作成

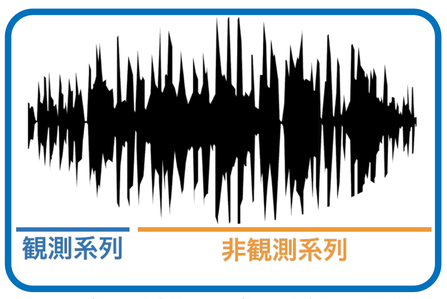

図 4. 観測系列と非観測系列の関係

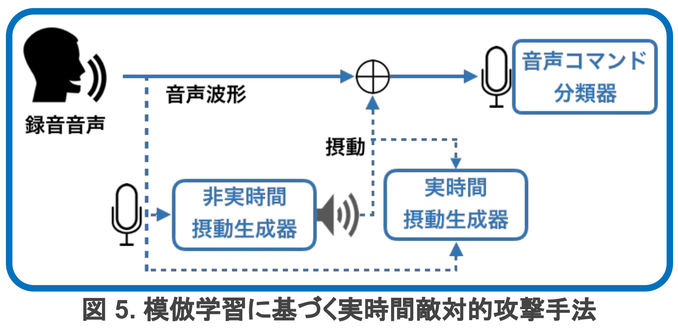

図 5. 模倣学習に基づく実時間敵対的攻撃手法

[3] K. Hwang and W. Sung, Character-level incremental speech recognition with recurrent neural networks, 2016.

非実時間摂動生成器の作成

One pixel attack[4]により非実時間摂動を作成

- 計算量の問題から数サンプルのみノイズを付与

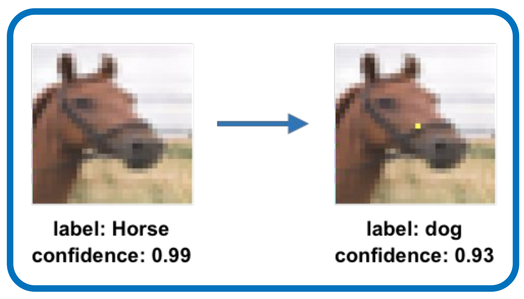

図 6. One pixel attackの例

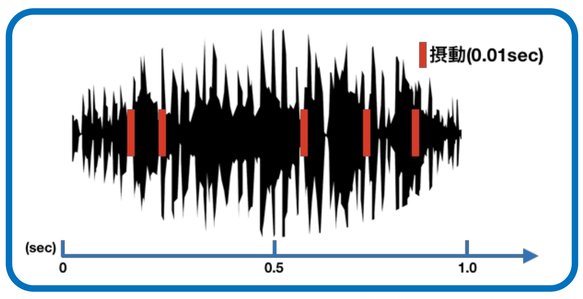

図 7. One pixel attackの音声波形への適用

[4] Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi, One pixel attack for fooling deep neural networks, 2019

❸ 提案手法

非実時間摂動生成器によるデータ作成

ISRの各ブロックにOne pixel attackを適用

- ISRの各ブロックの出力により時系列を考慮
- 音声波形($\mathbf{o}^{st}$)と出力系列($\mathbf{o}^{it-1}$)から摂動($\mathbf{a}^{t}$)を付与するタイミングを推定

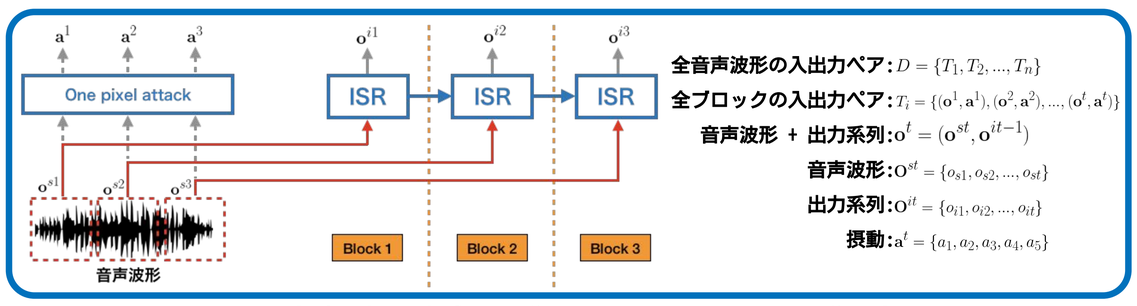

図 8. 実時間摂動生成器の学習に用いる入出力ペアの学習方法

❹ 実験

実験設定

ISRに対するCharacter Error Rate(CER)によって実時間敵対的攻撃の成功率を計測

- ISRの認識結果および出力確率のみ参照可
- ノイズよって変更される文字列は問わず
- Wall Street Journal(WSJ)[5]コーパスを使用

[5] Douglas B. Paul, Janet M. Baker, The Design for the Wall Street Journal-based CSR Corpus, 1992

実験結果

提案手法によりISRに対するCERが有意に上昇

- Random noiseでは約9%のCER上昇(**14.43%**)
- 提案手法では14%のCER上昇(**23.30%**)

表1. WSJコーパスの統計情報

Data type	Utterance	Hours
Train	37,318	82.00
Val	393	0.93
Test	393	0.93

表2. 各モデルに対するCER

Model	CER
ISR(original)	9.06
+ One pixel attack(non-real time)	52.30
+ Random noise(real time)	14.43
+ Adversarial noise(real time)	**23.30**

表3. 提案手法によるISRに対する認識誤り例

Original	ameritrust said it would finance the repurchase through internal funds, short-term borrowings and some of the proceeds from the sale of its investment in central bancorp.
+ Adversarial noise	american says it would fine the reputation through internal fund_, short_____ borrowing_ and some of the proceed_ from the sale of its investment in central bancorp.

NICT 国立研究開発法人 情報通信研究機構