L-006

ブラックボックス音声認識モデルへの決定境界に基づく敵対的サンプル攻撃

Adversarial Black-Box Attacks for Audio Recognition Based on Decision Boundary

山本 恭敬 * 福永 拓郎 † Peter Fulla ‡
Yasutaka Yamamoto Takuro Fukunaga

概 要

機械学習による自動認識システムへの攻撃手法の 1 つに敵対的サンプル攻撃がある．これは，人間の認識と認識モデルの判断が異なるようなサンプルを生成し，そのサンプルを用いてシステムを攻撃する手法のことである．

本研究では音声認識モデルに対する敵対的サンプル生成の新たな手法として，決定境界に近いサンプルを生成する手法を提案する．この手法は音声認識モデルの詳細について知らなくてもサンプルを生成できるブラックボックス攻撃であり, かつ音声認識モデルが攻撃者の意図通りの認識をするようなサンプルを生成する標的型攻撃に分類される．提案手法の有効性を確かめるために, Whisper や ReazonSpeech などの音声認識モデルに対して敵対的サンプルを生成し, 既存手法を用いて生成するよりも質の高いサンプルが生成できることを確かめた．

キーワード: AI セキュリティ, 敵対的サンプル, ブラックボックス攻撃, 音声認識.

1 はじめに

近年, 深層学習技術の発展は目覚ましく, 生活の様々な場面で機械学習の技術が利用されるようになっている．一方で, 機械学習特有の脆弱性の存在も指摘されている．そのような脆弱性を突いた攻撃の 1 つに**敵対的サンプル攻撃**がある．これは，人間の認識と認識モデルの判断が異なるようなサンプルを生成し，そのサンプルを用いてシステムを攻撃する手法のことである．例えば, 画像認識 AI において人間がパンダと認識する画像をテナガザルに分類させたり, 音声認識 AI において人間が「こんにちは」と聞こえる音声を「さようなら」と文字起こしさせたりといった具合である．

敵対的サンプルの生成手法は**ホワイトボックス攻撃**, **グレーボックス攻撃**, **ブラックボックス攻撃**の 3 つに分類することができる．ホワイトボックス攻撃は攻撃対象のモデルの中身が明らかである状況を仮定した攻撃であり, 入出力以外にモデルの勾配や中間層の値など, モデルから得られる情報すべてをサンプルの生成に用いることができる．グレーボックス攻撃は攻撃者が攻撃対象のモデルに部分的にアクセスできることを仮定した攻撃である．例えば入出力以外にモデルの損失関数の情報をサンプル生成に用いることができることなどを仮定するのが一般的である．ブラックボックス攻撃は攻撃者が攻撃対象のモデルに完全にアクセスできないときの攻撃であり, 入出力のみをサンプル生成に用いることができる．

また, 敵対的サンプルの生成手法には**標的型**か**非標的型**かというもう 1 つの分類基準もある．標的型は攻撃者がクラスを指定し, 認識モデルによって指定したクラスに分類されるようなサンプルを生成する手法のことである．非標的型は人間が認識する自然なクラスとは異なるクラスに分類されるようなサンプルを生成する手法のことである．

今日に至るまで, 音声認識モデルへの敵対的サンプル攻撃に関する研究が数多く行われてきた．例えば, Carlini, Wagner [3] は音声認識モデルに対する標的型ホワイトボックス攻撃を提案した．この提案手法はソフトウェア上で敵対的サンプルを音声認識モデルに直接入力することを前提としていた．そのため敵対的サンプルを発生する音源とマイクが物理的に離れていることが想定されておらず, 残響や再生環境から生じるノイズにより攻撃の威力が落ちることが課題であった．そこで, Yakura ら [12] は物理的な残響やノイズに左右されない標的型ホワイトボックス攻撃を提案した．また, Esmaeilpour ら [5] は音声認識モデルに対する標的

*中央大学 理工学研究科 Graduate School of Science and Engineering, Chuo University, a19.mk66@g.chuo-u.ac.jp
†中央大学 理工学部 Faculty of Science and Engineering, Chuo University, fukunaga.07s@g.chuo-u.ac.jp
‡中央大学 理工学部 Faculty of Science and Engineering, Chuo University, fulla@ise.chuo-u.ac.jp

型・非標的型両方のホワイトボックス攻撃を提案した．

これらのホワイトボックス攻撃の研究に加え, グレーボックス攻撃やブラックボックス攻撃に関する研究も進められてきた．Taori ら [11] は標的型グレーボックス攻撃を提案した．Khare ら [7] は遺伝的アルゴリズムと多目的最適化を用いた標的型・非標的型両方のブラックボックス攻撃を提案した．Abdullah ら [1] は離散フーリエ変換及び特異スペクトル解析を利用した非標的型ブラックボックス攻撃を提案した．Shi ら [9] はスパイク雑音の検出及び低減に着目した非標的型ブラックボックス攻撃を提案した．

本研究では画像認識モデルに対する決定境界攻撃 [2] を参考に, 音声認識モデルの決定境界に近いサンプルを生成する敵対的サンプル攻撃手法を提案する．この手法は標的型ブラックボックス攻撃に分類される．以降では, 敵対的サンプル生成に用いる音声において, 人間に聞こえてほしい音声を**オリジナル音声**と呼び, 音声認識モデルに書き起こさせるテキストを**標的テキスト**と呼び, 標的テキストに対応する音声を**標的音声**と呼ぶ．従来の音声認識モデルに対する標的型ブラックボックス攻撃ではオリジナル音声と標的テキストを入力し, オリジナル音声に適切な摂動を加えることで敵対的サンプルを生成してきた．しかし, この手法では攻撃者の意図通りのテキストに書き起こさせるための摂動を探索により発見しなければならないため, 摂動を得るまでに時間がかかることや場合によってはそのような摂動が見つけられないという課題がある．そこで本研究ではオリジナル音声及び標的音声を入力し, 標的音声から少しずつ変更を加えることで徐々にオリジナル音声に近づけるようにして敵対的サンプルを生成する．

本論文の構成は次の通りである．第 2 章で本論文で用いる記号を定義する．第 3 章では本研究と特に関連が強い先行研究を紹介する．第 4 章で提案手法を導入し, 第 5 章で実験による評価の結果を述べる．第 6 章でまとめを述べる．

2 準備

A をベクトルあるいは行列としたときに, A の L2 ノルムを $|A|$ とあらわす．

音声データの 1 サンプルの取りうる値の集合を $\mathcal{S}$ とおく．本研究では, $\mathcal{S} = \{s \in \mathbb{R} | -1 \leq s \leq 1\}$ とする．以降では特に断りがない限り, 本研究で用いる全ての音声データのサンプリング周波数を $r = 16000$[Hz], 収録時間を T 秒とする．

音声を入力してメル周波数ケプストラム係数を出力する関数を MFCC とおく．ここで, 本研究ではメル周波数ケプストラム係数の, 周波数から求まる成分に対応する次元の数 n_{MFCC} を 20 とする．また, 短時間フーリエ変換におけるフレームシフト n_{fs} を 512 とする．これはサンプリング周波数 $r = 16000$[Hz] のもとで時間に直すと 32 ミリ秒に相当する．メル周波数ケプストラム係数がとり得る値の集合は $\mathbb{R}^{n_{\mathrm{MFCC}} \times \lfloor \frac{\lfloor rT \rfloor}{n_{\mathrm{fs}}} \rfloor}$ となる．

2 つのテキストを入力して編集距離（レーベンシュタイン距離）を出力する関数を d とおく．なお, 本研究では編集距離の計算にあたり, あらかじめテキストからは句読点を取り除く．

音声データを入力して文字起こししたテキストを出力する音声認識モデルを ASR とおく．

3 関連研究

3.1 画像認識モデルに対する境界攻撃 [2]

本節では画像の敵対的サンプル生成の手法として知られている境界攻撃での画像生成アルゴリズムについて説明する．この手法は標的型ブラックボックス攻撃に分類される．

まず, 画像データの 1 ピクセルの取りうる値の集合を $\mathcal{P}$ とおく．例えば, RGB 画像では 1 ピクセルが 0 以上 1 以下の実数をとるので, $\mathcal{P} = \{p \in \mathbb{R} | 0 \leq p \leq 1\}^3$ となる．

アルゴリズムの入力は $n \times m$ ピクセルからなる画像 $s \in \mathcal{P}^{m \times n}$ と $o \in \mathcal{P}^{m \times n}$ である．s を標的画像, o をオリジナル画像と呼ぶ．出力は s や o と同じサイズの画像 $a \in \mathcal{P}^{m \times n}$ である．画像 a に関して, 標的画像に似ているがオリジナル画像と同じカテゴリに分類されるような画像となることが求められる．

境界攻撃のアルゴリズムをアルゴリズム 1 に示す．ここで, 画像を入力して予測ラベルを出力する画像認識モデルを Q とおく．アルゴリズム中で使用される M_{B} は最大反復回数を表わすパラメータである．以下では本アルゴリズムで用いられている関数 ORTHOGONAL について述べる．

ORTHOGONAL は画像 $a \in \mathcal{P}^{m \times n}$, 画像 $o \in \mathcal{P}^{m \times n}$, ハイパーパラメータ $\delta \in \mathbb{R}$ を入力し, 画像 $p \in \mathcal{P}^{m \times n}$ を出力する関数である．o を中心とし $|a - o|$ を半径とする球 $\{x \in \mathcal{P}^{m \times n}, |x - o| = |a - o|\}$ を考え, a からランダムな方向におおよそ $\delta|a - o|$ の大きさだけ球上を移動して得られる画像が出力画像 p である．

アルゴリズム 1 境界攻撃

入力 標的画像 $s \in \mathcal{P}^{m\times n}$, オリジナル画像 $o \in \mathcal{P}^{m\times n}$
出力 $a \in \mathcal{P}^{m\times n}$

$a \leftarrow s$
$i \leftarrow 0$
while $i < M_{\mathrm{B}}$ **do**
  $p \leftarrow \mathrm{ORTHOGONAL}(a, o, \delta)$
  if $Q(p) = Q(o)$ **then**
    $a \leftarrow p$
    $\delta > 0$ の範囲で δ を少し大きくする
  else
    $\delta > 0$ の範囲で δ を少し小さくする
    continue
  end if
  $p \leftarrow \epsilon o + (1-\epsilon)a$
  if $Q(p) = Q(o)$ **then**
    $a \leftarrow p$
    $\epsilon > 0$ の範囲で ϵ を少し大きくする
    $i \leftarrow i + 1$
  else
    $\epsilon > 0$ の範囲で ϵ を少し小さくする
  end if
end while
return a

3.2 多目的最適化を用いた攻撃 [7]

音声認識モデルに対する標的型ブラックボックス攻撃の例として, Khare, Aralikatte, Mani [7] による研究を紹介する. 入力は音声 $o \in \mathcal{S}^{\lfloor rT \rfloor}$, テキスト $t \in \mathcal{C}^n$ である. o をオリジナル音声, t をターゲットテキストと呼ぶ. 出力は o と同じサイズの音声 $a \in \mathcal{S}^{\lfloor rT \rfloor}$ である. 音声 a は, オリジナル音声に似ているがターゲットテキストに文字起こしされるような音声となることが求められる.

この手法では音声 a と音声 o のメル周波数ケプストラムにおける距離 $|\mathrm{MFCC}(a) - \mathrm{MFCC}(o)|$ と, 音声認識モデルを用いて a を文字起こしして得られるテキスト $\mathrm{ASR}(a)$ とテキスト t の編集距離 $d(\mathrm{ASR}(a), t)$ の 2 つを目的関数とし, 両方同時に最小化することを試みる. この 2 つの目的関数は互いにトレードオフの関係にあり, 変数 a に関して一方を小さくしようとすると他方が大きくなる. この手法ではこの 2 つの目的関数を同時に最適化する解を遺伝的アルゴリズムで計算する.

4 提案手法

4.1 概要

サーベイ論文 [4] によると, 当該論文執筆時点に至るまでに発表された音声認識モデルに対する標的型ブラックボックス攻撃は Khare ら [7] による多目的最適化を用いた手法のみである. そこで本研究では, Brendel ら [2] による境界攻撃のアプローチを音声の敵対的サンプル生成に利用する音声境界攻撃 (Audio Boundary Attack, 以下 ABA）を提案する.

この手法の入力は n 次元ベクトルからなる音声 $s \in \mathcal{S}^n$ と $o \in \mathcal{S}^n$ である. s を標的音声, o をオリジナル音声と呼ぶ. 出力は s や o と同じサイズの音声 $a \in \mathcal{S}^n$ である. 音声 a に関して, 標的音声に似ているがオリジナル音声と同じテキストに文字起こしを行うような音声となることが求められる. これを次のような最適化問題として表現する.

$$\begin{aligned} \text{minimize} \quad & |a - o| \\ \text{subject to} \quad & \mathrm{ASR}(a) = \mathrm{ASR}(s), \\ & a \in \mathcal{S}^n. \end{aligned}$$

ABA は, この問題を解くヒューリスティックアルゴリズムの一種である.

アルゴリズム 1 を画像ではなく音声にそのまま適用しても問題なく動作する. しかし, 実際に音声に適用したところ, 画像と比較して音声の場合は 2 つの入力音声が混ざって聞こえる音声になり, 質の良い敵対的サンプルを生成することができなかった. この課題に対応するため, 本研究で提案する ABA アルゴリズムでは**同時更新**と**1 点更新**と呼ぶ新たなサンプルの更新方法を用いる.

4.2 音声の文字起こしに関する判定アルゴリズム

提案手法では音声 a を音声認識モデルに書き起こさせて得られるテキストが t に一致するか否かを調べる関数 JUDGE を用いる. JUDGE の出力は True もしくは False である. N を音声認識モデルによる書き起こしの回数を表すパラメータとする. 一部の音声認識モデルでは音声を書き起こしたテキストが実行の度に変わることがある. そのため, N 回書き起こしを行い, 全て t に一致した場合にのみ音声 a を音声認識モデルに書き起こさせて得られるテキストが t に一致すると判定し, JUDGE は True を出力する.

4.3 同時更新

本研究で用いる同時更新のアルゴリズムを SA とおく．SA の入力は音声データ $s \in \mathcal{S}^{\lfloor rT \rfloor}$, 音声データ $o \in \mathcal{S}^{\lfloor rT \rfloor}$, 出力は音声データ $a \in \mathcal{S}^{\lfloor rT \rfloor}$ である．ここで, JUDGE を呼び出す最大回数を M_{SA} とおく．

これは前述した境界攻撃とほぼ同じアルゴリズムである．

4.4 1 点更新

1 点更新のアルゴリズムをアルゴリズム 2 に示す．入力は音声データ $s \in \mathcal{S}^{\lfloor rT \rfloor}$, 音声データ $o \in \mathcal{S}^{\lfloor rT \rfloor}$, 出力は音声データ $a \in \mathcal{S}^{\lfloor rT \rfloor}$, 更新回数 $u \in \mathbb{N}$ である．ここで, JUDGE を呼び出す最大回数を M_{OP} とおく．また, R は $0 \leq R \leq 1$ を満たすハイパーパラメータである．以降, R をサンプル変動割合と呼ぶ．

アルゴリズム 2 1 点更新 OP

入力 $s \in \mathcal{S}^{\lfloor rT \rfloor}, o \in \mathcal{S}^{\lfloor rT \rfloor}, R \in [0,1]$
出力 $a \in \mathcal{S}^{\lfloor rT \rfloor}, u \in \mathbb{N}$

$a \leftarrow s$
$|c_i - o_i|$ の降順に $i = 1, 2, ..., \lfloor rT \rfloor$ を並べた列を計算し $(t_1, t_2, ..., t_{\lfloor rT \rfloor})$ とする．
$T \leftarrow \mathrm{ASR}(o)$
$u \leftarrow 0$
$m \leftarrow 0$
$j \leftarrow 1$
while $m < M_{\mathrm{OP}}$ かつ $j \leq \lfloor rT \rfloor$ **do**
    $p \leftarrow c$
    $p[t_j] \leftarrow p[t_j] + R(o[t_j] - p[t_j])$
    if $\mathrm{JUDGE}(p, T) = \mathrm{True}$ **then**
        $a \leftarrow p$
        $u \leftarrow u + 1$
    end if
    $j \leftarrow j + 1$
    $m \leftarrow m + 1$
end while
return a, u

4.5 音声境界攻撃の全体像

音声境界攻撃 ABA の全体像をアルゴリズム 3 に示す．ここで, M_{ABA} は最大反復回数を表わすパラメータである．同時更新を行う関数を SA, 1 点更新を行う関数を OP と記す．

ABA では, 数列 $R = (R_1, ..., R_k)$ と実数 L を用いてサンプル変動割合を調節する．R をサンプル変動割合リストとする．1 回のイテレーションに対して更新回数が 0 になった場合にサンプル変動割合を変更することで更新回数の上昇を狙う．これにより, 更新率を固定する場合と比較して, より o に近い敵対的サンプルを生成することが可能である．

アルゴリズム 3 音声境界攻撃 ABA

入力 標的音声 $s \in \mathcal{S}^{\lfloor rT \rfloor}$, オリジナル音声 $o \in \mathcal{S}^{\lfloor rT \rfloor}$, $k \in \mathbb{N}$, $R \in [0,1]^k$
出力 $a \in \mathcal{S}^{\lfloor rT \rfloor}$

$a \leftarrow s$
$m \leftarrow 0$
$i \leftarrow 0$
while $m < M_{\mathrm{ABA}}$ かつ $i < k$ **do**
    $a \leftarrow \mathrm{SA}(a, o)$
    $a, u \leftarrow \mathrm{OP}(a, o, R_i)$
    if $u = 0$ **then**
        $i \leftarrow i + 1$
    end if
    $m \leftarrow m + 1$
end while
return a

5 実験

音声認識モデルに対する敵対的サンプル攻撃の既存研究では実験において音声認識モデルに Deepspeech [6] を用いることが多かった．しかし, Deepspeech は本研究開始時点においてライブラリの提供が終了していたため, 代わりに本研究では Whisper [8] および ReazonSpeech [13] を用いた．Whisper はバージョン 20231107, モデル large を用いた．ReazonSpeech はモデル reazonspeech-nemo-v2 を使用した．本研究では JSUT コーパス [10] に含まれる日本語音声及び音声合成技術により生成した音声を用いて実験を行った．

5.1 ハイパーパラメータの影響

ハイパーパラメータが及ぼす影響を調査するために, ABA アルゴリズムにおいて用いたハイパーパラメータ R を様々な値に設定して評価を行う．ABA において R の値を変えたときの実験結果を表 1, 2 に示す．標的音声 s, オリジナル音声 o は ABA の入力音声を, a は ABA の出力音声を表す．音声 o のサンプル数（ベクトル o の要素数）を $L(o)$ とおく．設定した最大反復回数 M_{ABA} に対して, 実際の反復回数を N_{LOOP} とおく．基本のパラメータは $M_{\mathrm{SA}} = 0.01 \times L(o)$, $M_{\mathrm{OP}} = 0.1 \times L(o)$,

$M_{\mathrm{ABA}} = 100$ または 200, $R = [0.5]$, $N = 2$ である．ここで, ABA における音声認識モデルへのアクセス回数の上限は $M_{\mathrm{ABA}} \times (M_{\mathrm{SA}} + M_{\mathrm{OP}}) \times N$ で表される．本調査では音声合成により生成した 4 つの音声（こんにちは, さようなら, ありがとう, すみません）及び JSUT コーパスに含まれる音声（他のものは何もいらない, 立ち退きの予告を受けた）を用いた．表 1, 2 では原則 $N = 2$ のときの結果を掲載しているが, 一部の実験では $\mathrm{ASR}(a)$ が $\mathrm{ASR}(s)$ と一致しなかったため, その場合は 2 つのテキストが一致した $N = 8$ の実験結果を記載している．

基本のパラメータにおける実行時間は, 「こんにちは」「さようなら」「ありがとう」「すみません」の音声を用いた場合, 50 イテレーションで終了し, およそ 30 時間であった．「他のものは何もいらない」「立ち退きの予告を受けた」の音声を用いた場合は, 40 イテレーションで終了し, およそ 70 時間であった．

表 1: R の値を変えたときの ABA の実験結果（音声認識モデルが Whisper の場合）

$\mathrm{ASR}(s)$	$\mathrm{ASR}(o)$	R	N_{LOOP}	N	$\lvert a - o \rvert$
こんにちは	さようなら	[0.2]	56	2	8.19
こんにちは	さようなら	[0.5]	52	2	5.17
こんにちは	さようなら	[1.0]	24	2	4.29
こんにちは	さようなら	[1.0, 0.9,..., 0.1]	66	2	3.95
ありがとう	すみません	[0.2]	64	2	6.82
ありがとう	すみません	[0.5]	49	2	4.15
ありがとう	すみません	[1.0]	18	2	2.80
ありがとう	すみません	[1.0, 0.9,..., 0.1]	103	8	2.70
他のものは何もいらない	立ち退きの予告を受けた	[0.2]	41	2	2.581
他のものは何もいらない	立ち退きの予告を受けた	[0.5]	41	2	1.623
他のものは何もいらない	立ち退きの予告を受けた	[1.0]	50	2	0.991
他のものは何もいらない	立ち退きの予告を受けた	[1.0, 0.9,..., 0.1]	171	2	0.948

表 2: R の値を変えたときの ABA の実験結果（音声認識モデルが ReazonSpeech の場合）．N_{LOOP} に*がついているものは, 反復回数が設定した上限に達したので打ち切ったことを意味する．

$\mathrm{ASR}(s)$	$\mathrm{ASR}(o)$	R	N_{LOOP}	$\lvert a - o \rvert$
こんにちは	さようなら	[0.2]	100*	2.81
こんにちは	さようなら	[0.5]	100*	2.00
こんにちは	さようなら	[1.0]	91	2.93
こんにちは	さようなら	[1.0, 0.9,..., 0.1]	196	2.48
ありがとう	すみません	[0.2]	20	19.40
ありがとう	すみません	[0.5]	54	3.66
ありがとう	すみません	[1.0]	100*	3.09
ありがとう	すみません	[1.0, 0.9,..., 0.1]	200*	2.98
他のものは何もいらない	立ち退きの予告を受けた	[0.2]	96	3.92
他のものは何もいらない	立ち退きの予告を受けた	[0.5]	85	3.04
他のものは何もいらない	立ち退きの予告を受けた	[1.0]	146	1.98
他のものは何もいらない	立ち退きの予告を受けた	[1.0, 0.9,..., 0.1]	200*	1.98

基本的に $R = [1.0, 0.9, ..., 0.1]$ のときにオリジナル音声に近い敵対的サンプルを生成することができた．

5.2 既存手法との比較

提案手法の評価のため, 3.2 節で紹介した Khare, Aralikatte, Mani [7] のアルゴリズム（以下, 既存手法と呼ぶ）との比較を行った． 既存手法は論文の記述をもとに我々で実装したプログラムを用いた． 提案手法のパラメータは $M_{\mathrm{SA}} = 0.01 \times L(o)$, $M_{\mathrm{OP}} = 0.1 \times L(o)$, $M_{\mathrm{ABA}} = 200$, $R = [1.0, 0.9, ..., 0.1]$ を用いた．実験結果を表 3, 4 に示す．この実験結果によると, 出力音声とオリジナル音声の MFCC の L2 ノルム $|\mathrm{MFCC}(a) - \mathrm{MFCC}(o)|$ において既存手法と提案手法のどちらが優れているかは音声によって異なっているが, 生の音声波形の L2 ノルム $|a - o|$ は音声によらず提案手法のほうが優れていることが分かる．既存研究とは異なり, 提案手法では生の音声波形の L2 ノルムを目的関数としていることが理由であると考えられる．

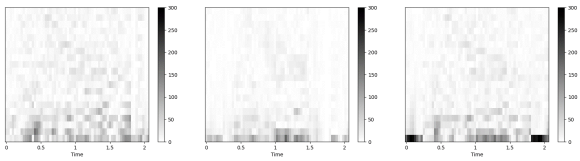
表 3: 既存手法と提案手法の比較の実験結果（音声認識モデルが Whisper の場合）

$\mathrm{ASR}(s)$	$\mathrm{ASR}(o)$	手法	$\lvert a - o \rvert$	$\lvert \mathrm{MFCC}(a) - \mathrm{MFCC}(o) \rvert$
こんにちは	さようなら	Khare+19	24.73	514.84
こんにちは	さようなら	提案手法	3.95	492.07
ありがとう	すみません	Khare+19	14.82	513.44
ありがとう	すみません	提案手法	2.70	558.98
他のものは何もいらない	立ち退きの予告を受けた	Khare+19	10.16	1290.57
他のものは何もいらない	立ち退きの予告を受けた	提案手法	0.948	729.55

表 4: 既存手法と提案手法の比較の実験結果（音声認識モデルが ReazonSpeech の場合）

$\mathrm{ASR}(s)$	$\mathrm{ASR}(o)$	手法	$\lvert a - o \rvert$	$\lvert \mathrm{MFCC}(a) - \mathrm{MFCC}(o) \rvert$
こんにちは	さようなら	Khare+19	24.94	415.47
こんにちは	さようなら	提案手法	2.48	338.79
ありがとう	すみません	Khare+19	25.89	672.82
ありがとう	すみません	提案手法	2.98	626.73
他のものは何もいらない	立ち退きの予告を受けた	Khare+19	8.72	1201.97
他のものは何もいらない	立ち退きの予告を受けた	提案手法	1.98	1295.67

MFCC に関する標的音声との差を可視化した図を図 1 に示す．図の各セルは MFCC の要素に対応しており, 標的音声と出力音声の差が大きい要素のセルの色が濃く表示される．



(a) オリジナル音声と標的音声 **(b)** オリジナル音声と提案手法により生成した敵対的サンプル **(c)** オリジナル音声と Khare+19 により生成した敵対的サンプル

図 1: MFCC の差の各要素に絶対値をとって得られる 2 次元マップの可視化．標的音声に「他のものは何もいらない」, オリジナル音声に「立ち退きの予告を受けた」を用いた．敵対的サンプル生成には whisper を用いた．

6 おわりに

本研究では音声認識モデルに対する敵対的サンプル生成手法として, 新たな標的型ブラックボックス攻撃の手法を提案した.

今後の課題に関して述べる. ABA に限らず, ブラックボックス攻撃はモデルへのアクセス回数を多く必要とし, その分検知されやすいという欠点がある. また, アクセス回数が多いほど敵対的サンプルの生成に時間がかかり, 実用的な攻撃とは言えない. そのため, 敵対的サンプル生成の精度を保ったままアクセス回数を減らすための仕組みが必要である. また, 生成された音声においてボツボツと発生するスパイクノイズが多く観察された. 人間の耳はホワイトノイズよりもスパイクノイズに敏感であるといわれており, 改善が必要である.

謝辞

本研究は JSPS 科研費 JP21H03397, JP24K14844 の助成を受けたものです.

参考文献

[1] H. Abdullah, M. Sajidur Rahman, W. Garcia, L. Blue, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor. Hear “No Evil”, See “Kenansville”: Efficient and Transferable Black-Box Attacks on Speech Recognition and Voice Identification Systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 712–729, 2021.

[2] W. Brendel, J. Rauber, and M. Bethge. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[3] N. Carlini and D. Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7, 2018.

[4] P. Cheng and U. Roedig. Personal Voice Assistant Security and Privacy—A Survey. *Proc. IEEE*, 110(4):476–507, 2022.

[5] M. Esmaeilpour, P. Cardinal, and A. L. Koerich. Towards Robust Speech-to-Text Adversarial Attack. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2869–2873, 2022.

[6] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech: Scaling up end-to-end speech recognition, 2014. arXiv:1412.5567.

[7] S. Khare, R. Aralikatte, and S. Mani. Adversarial Black-Box Attacks on Automatic Speech Recognition Systems Using Multi-Objective Evolutionary Optimization. In *Proc. Interspeech 2019*, pages 3208–3212, 2019.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proc. the 40th International Conference on Machine Learning*, 2023.

[9] X. Shi, C. Song, X. Huang, and Y. Wu. Research on Speech Adversarial Sample Attack Based on APA. In *Proc. the 4th International Conference on Big Data Analytics for Cyber-Physical System in Smart City - Volume 2*, pages 329–336, 2023.

[10] R. Sonobe, S. Takamichi, and H. Saruwatari. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, 2017. arXiv:1711.00354.

[11] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri. Targeted Adversarial Examples for Black Box Audio Systems. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 15–20, 2019.

[12] H. Yakura and J. Sakuma. Robust Audio Adversarial Example for a Physical Attack. In *Proc. the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5334–5341, 2019.

[13] Y. Yue, D. Mori, and S. Fujimoto. ReazonSpeech: A Free and Massive Corpus for Japanese ASR. In **言語処理学会年次大会発表論文集**, volume 29, 2023.