超音波の分離放射による音声認識機器への攻撃: ユーザスタディ評価と対策技術の提案

飯島 涼[1,2,a)]　南 翔汰[1]　シュウ インゴウ[1]　竹久 達也[2]　高橋 健志[2]　及川 靖広[1]　森 達哉[1]

概要：市中で販売されている音声認識装置の多くは，人が発生した肉声のみならず，ダイナミックスピーカや，超音波を用いた指向性スピーカから再生される音声にも反応することが知られている．肉声ではない音声信号に反応することを悪用し，なりすましやリプレイ攻撃等の様々な攻撃が可能となる．本論文では，超音波を用いた高度な音声攻撃を提案する．鍵となるアイディアは変調した超音波を搬送波と側帯波に分離して放射することであり，2 つの波を標的のマイクに一致する点で交差させることにより，可聴音が聞こえる範囲を極小化する．その攻撃を X(Cross)-Audio Attack と名付け，その特性と有効性を主観評価実験により評価する．また，このような音声認識デバイスを標的にした攻撃への汎用的な対策手法として，超音波から発生する音声，ダイナミックスピーカが発生する音声，および人間の肉声を見分ける識別器，Voice Liveness Detector を提案する．作成した識別器の性能評価を行い，これまでに提案された音声認識への攻撃全てを検知でき，かつ高精度で攻撃検知が可能であることを示した．

キーワード：音声認識，超音波，X-Audio Attack，Voice Liveness Detector

1. はじめに

音声認識技術の進展に伴い，人間とコンピューターがスムーズな音声会話によるインタラクションを行う世界が実現しつつある．スマートフォンはもとより，AI スピーカ，自動運転車，スマートウォッチ，冷蔵庫などの各種家電にまで音声認識が搭載され，それらは各種 IoT デバイスとも連携する．今後も対応する IoT デバイスが増え，さらに利用ケースが増えることが予想される．

音声認識技術が普及を始める一方で，音声認識をターゲットとした脅威の可能性も指摘され始めている．その中でもっとも危険性の高い攻撃が，ユーザに認識できないように加工した音声コマンドを入力するというものである．超音波を利用した DolphinAttack [10] や，特定の方向にしか音が届かないパラメトリックスピーカを使用した Audio Spotlight Attack [12] などの攻撃により，利用者に対して秘密裏に音声を入力し，音声認識デバイスを作動させることが可能となる．本論文では，パラメトリックスピーカの原理を応用し，特定の点でのみ音声を再生する技術を使用した攻撃を提案する．パラメトリックスピーカとは，超音波の特性を利用して特定の方向にしか聞こえない音を作り出すスピーカである．超音波を再生したい音声で変調し放射すると，空気中で元の音声が復調されるという現象を利用している．パラメトリックスピーカから出ている数種類の超音波を分離し，交差させて放射することで聞こえるポイントを極小化した後，音声認識デバイスのマイクに入力する．本論文では，この攻撃を X(Cross)-Audio Attack と名付け，その攻撃特性を評価する（図 1 参照）．この攻撃によって，DolphinAttack よりも距離が遠く，かつただ単にパラメトリックスピーカを使用した場合に比べユーザに聞こえにくい攻撃を実現することが可能となる．攻撃の特性を実験により測定したあと，ユーザの認知率を主観評価実験により測定する．

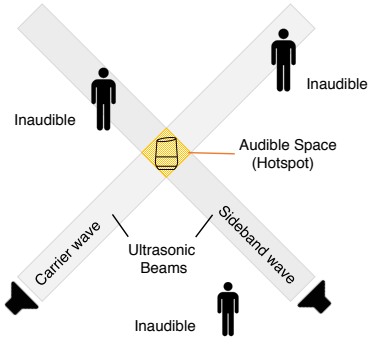

図 1　X-Audio attack の概要

最後に，音声認識デバイスへの攻撃を防ぐ対策手法を提

[1] 早稲田大学
[2] 情報通信研究機構
[a)] ryo@nsl.cs.waseda.ac.jp

案する．人間の肉声，パラメトリックスピーカから放射された声，およびダイナミックスピーカから再生された声を識別する識別器を作成する．この識別モデルの提案により，超音波を使用した攻撃や，録音・合成した音声を利用した攻撃を全て防ぐ音声認識デバイスの設計が可能になる．作成した識別器を Voice Liveness Detector と名付け，その識別精度を評価する．評価の結果，これまでに提案された音声認識への攻撃全てを検知でき，かつ 95%以上の高精度で攻撃検知が可能な対策モデルであることを示した．

本論文の貢献は以下の通りである．

- 超音波の特性を利用した新しい攻撃（X-Audio Attack）の可能性を指摘した（3 章）．
- 攻撃の特性を主観評価，客観評価の双方で行い，脅威となり得ることを示した（4 章）．
- 音声認識を標的にした攻撃の対策ができる識別器のモデル（Voice Liveness Detector）を提案した．性能評価を行い，音声認識への攻撃を防ぐモデルとして有効であることを示した（5 章）．

2. 研究背景

2.1 音声アシスタント機能

スマートスピーカやスマートフォンは音声認識技術および音声合成技術により，人間とのスムーズなインタラクションを可能とした音声アシスタント機能を実現している．音声アシスタント機能を実装したデバイスを操作する場合，**起動** (Activation) と**命令** (Command) から成る 2 つのフェーズが存在する．起動はユーザが特定の単語を発話することにより実現される．起動時の音声信号処理はデバイス内部の音声認識機能を用いる．起動後にユーザは命令を含む文を発話することにより，その命令がデバイスによって実行される．命令を含む音声データは，ネットワークを経由してクラウド上の音声認識サービスで処理され，その結果がやはりネットワークを経由してデバイスにフィードバックされる．起動時の音声認識処理と命令時の音声認識処理は異なるため，音声認識成功率において異なる特性を示すことがわかっている．また，デバイスの種別によって音声認識が可能な距離が異なるため，デバイスに応じた音声認識特性の調整が行われていると推察される．

デバイスを起動するための特定の単語として， Google Home や Android OS で使われる Google Assistant では，“OK Google” あるいは「ねぇ Google」などを, Amazon Echo では “Alexa” や “Computer” などの単語を， Apple 社の iOS， macOS， watchOS，tvOS 等で使われる Siri では “Hey Siri” が使われる．命令によって実行できるサービスは様々であり，実行可能なサービスをカスタマイズすることが可能なプラットフォームも存在する（Amazon Skills など）．サービスは電話をかける，スケジュールの確認をする，メッセージを送るなど，様々な操作が可能であり，複数の命令を組み合わせることも可能である．

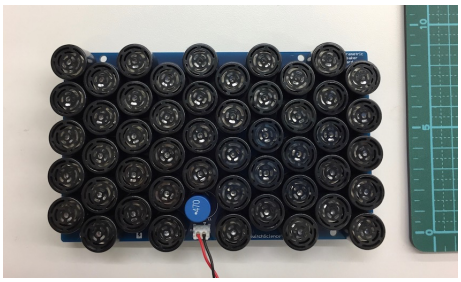

図 2 パラメトリックスピーカ

スマートスピーカは家族など，複数のユーザが共有して利用するケースが多いため，音声データにもとづいて個人を識別するマルチユーザ機能が実装されている．各ユーザはセットアップ時に自身の音声を何度か入力することによって，話者識別を行うための訓練データを供出する．ユーザが使用中の声を学習し徐々に適応する機能も存在する．

2.2 パラメトリックスピーカの原理

パラメトリックスピーカは超音波を利用して指向性のある音を生成する．パラメトリックスピーカは多数の超音波素子を並列に並べた形状をしている．パラメトリックスピーカの例を図 2 に示す．

周波数が近接した 2 つの有限振幅音波（周波数を $f_1 > f_2$ とする）を同方向に放射すると，空気の非線形現象によりその周波数の和音や差音 ($f_1 \pm f_2$) が生成される [8]．パラメトリックスピーカを駆動する際，可聴音で振幅変調した超音波を空気中に放射すると，変調波内の 2 つの周波数成分の差音が可聴音として復調される．この現象をパラメトリック現象という． X-Audio attack が応用するパラメトリック現象について簡単に述べる．

パラメトリックスピーカから出る音波を $p = p(x, t)$ とする．ここで，x はスピーカからの距離， t は時間である．指向性をもたせたい音は振幅変調して放射されるため，パラメトリックスピーカからは主に 3 種類の周波数の音波が出ている．搬送波として使われる超音波の周波数を f_c, その両サイドに現れる側帯波の周波数を f_{s_-},f_{s_+} とする．

f_m を元信号のもつ周波数，すなわち攻撃者が入力したい信号の周波数とすると，周波数間には $f_{s_-} = f_c - f_m$, $f_{s_+} = f_c + f_m$ の関係がある．まとめると，パラメトリックスピーカから放射される音波 p は

$$p = p_c \sin\left(2\pi f_c t'\right) + p_{s_-} \sin\left(2\pi f_{s_-} t'\right) + p_{s_+} \sin\left(2\pi f_{s_+} t'\right) \tag{1}$$

で表される．

ここで，t は $t' = t - x/c_0$ によって定義される遅延時間である．c_0 は音速，p_c, p_{s_-},p_{s_+} は搬送波，側帯波の振幅を表す．

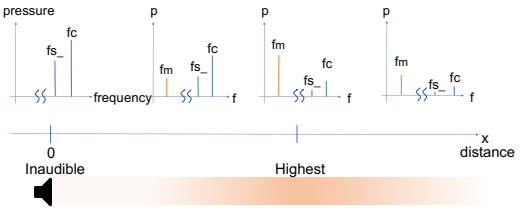

図 3 Demodulation in the air.

パラメトリックスピーカから放射された音波はバーガース方程式（Burger's equation）という流体モデルにしたがって伝搬する [4]．バーガース方程式は音が流体を伝搬する過程で生じる非線形性を表す式であり，

$$\frac{\partial p}{\partial x} = \frac{\beta}{\rho_0 c_0^3} p \frac{\partial p}{\partial t'} + \frac{\delta}{2c_0^3} \frac{\partial^2 p}{\partial t'^2}, \tag{2}$$

で示される．β は非線形係数，ρ_0 は空気密度，c_0 は音速である．(2) 式の右辺第 1 項が非線形性を表している．この項に (1) 式を代入すると，

$$\begin{aligned} p\frac{\partial p}{\partial t'} &\underset{f_c, f_{s\_}}{\approx} p_c \sin(2\pi f_c t') p_s 2\pi f_{s\_} \cos(2\pi f_{s\_} t') \\ &\quad + p_c 2\pi f_c \cos(2\pi f_c t') p_s \sin(2\pi f_{s\_} t') \\ &\underset{f_c - f_{s\_}}{\approx} \pi p_c p_{s\_} (f_c - f_{s\_}) \sin(2\pi (f_c - f_{s\_}) t') \\ &= \pi p_c p_{s\_} f_m \sin(2\pi f_m t'), \end{aligned} \tag{3}$$

が得られる．$\underset{a}{\approx}$ は，（表現を簡潔にするため）a とは関係のない項を取り除いていることを意味する．(3) 式の結果を (2) 式に代入すると

$$\frac{\partial p}{\partial x} \underset{f_m}{\approx} \frac{\beta \pi p_c p_{s\_} f_m}{\rho_0 {c_0}^3} \sin(2\pi f_m t') \tag{4}$$

が得られる．(4) 式より，振幅変調した音波が伝搬する過程で生じる非線形性によって，元信号を含む音波が現れることがわかる．図 3 はパラメトリック現象の概要を示している．この現象は，超音波素子それぞれから放射される音波が同位相であるという仮定の元成立する．

2.3 超音波を利用した音声認識機器への攻撃

音声認識機器を標的にした既存の攻撃のうち，超音波に関連する攻撃は 2 つ存在する．DolphinAttack は，音声認識機器への入力の過程での非線形性を利用して，超音波で音声コマンドを入力してしまうというものである．Audio Spotlight Attack は，特定の方向にしか音が届かないパラメトリックスピーカを使用して，ユーザに気づかれないよう音声コマンドを入力する攻撃である．パラメトリックスピーカは，音声で振幅変調された超音波の側帯波，搬送波を，同じ地点から放射することで指向性を実現する．どちらの攻撃も，現状使用されている音声認識機器が，人の口から出た音声と，スピーカ/超音波素子からでた音波を見分けられないことを前提とした攻撃である．

3 章で，DolphinAttack よりも到達する距離が長く，かつ Audio Spotlight Attack よりもさらにきこえる範囲を狭めた X-Audio Attack について述べたあと，5 章で関連する攻撃を防ぐ対策モデルを示す．

3. X-Audio Attack

3.1 攻撃の概要

X-Audio Attack は， 2.2 節で示したパラメトリック現象を応用した攻撃である．基本的なアイディアは，パラメトリックスピーカが同方向に放射していた音波を分離し，標的のデバイスが存在する地点で交差させることで，1 点でのみ音声が聞こえるようにするというものである．音声が聞こえるポイントを Hotspot と呼ぶ．

まず，通常のパラメトリックスピーカと同様に音声コマンドで超音波を振幅変調したあと，側帯波と搬送波を MATLAB を用いて分離させる．次に 2 つのパラメトリックスピーカを用意し，一方のパラメトリックスピーカからは側帯波を，もう一方からは搬送波を同時に放射する．2 つのパラメトリックスピーカから出る音波を，標的の位置で交差させると，2.2 節で示した非線形性により，交差した点でのみ元の音声信号が復調される (図 1 参照)．Hotspot を音声認識機器の搭載するマイクに合わせることで，任意の命令が実行可能となる．

搬送波と側帯波を分離するアイディアは，過去の研究で考案されており [11]，実際に 1 点でのみ音圧が大きくなることが示されている．ここで，実際に A 特性の音圧レベル（dBA）を測定した結果を図 4 右に示す．スピーカの位置関係は，実験時のセットアップ図 5 と同様である．参考として，ダイナミックスピーカ，パラメトリックスピーカを部屋の中央に置いた場合の音圧と比較する．ダイナミックスピーカは同心円状に，パラメトリックスピーカは直線状に音波が伝わっているのに対し，X-Audio Attack のモデルでは音波の交差する中央部分の音圧が大きくなっていることがわかる．2 つの音波の交差部分以外にも超音波は存在し，厳密に中央部分のみが高い音圧となっていないことに注意されたい．

攻撃対象として，Google Home, Amazon Echo を使用する．実験時に放射する音声は距離やノイズ以外で認識率に影響を与えうる要因を排除するため，起動で使用する単語は，“OK, Google”，または “Alexa”，命令で使用する単語は，“What's on my next schedule?” に固定した． 起動では，“OK, Google”，“Alexa’ と入力した後，デバイスが待機状態になった場合を成功，命令では，アカウントに登録されている次の予定を答えた場合を成功とする．また，変調する前のコマンドの音声は実験の再現性を考慮して，Amazon Polly [1] の Ivy という合成音声を利用して生成する．今回は入力音声の言語として英語を選択する．

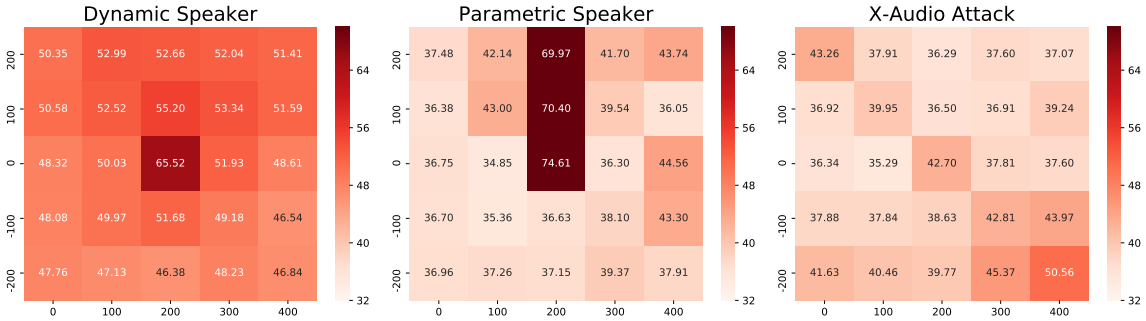

図 4 ダイナミックスピーカ (左)，パラメトリックスピーカ (中央)，X-Audio Attack(右) の音圧測定結果

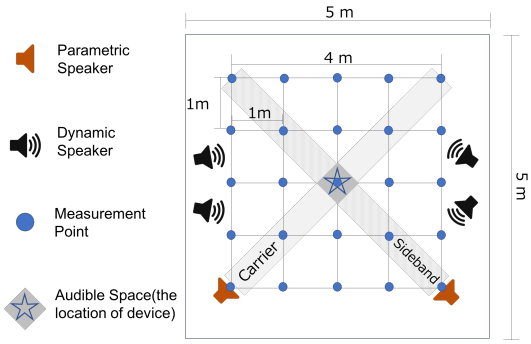

図 5 X-Audio Attack の実験セットアップ

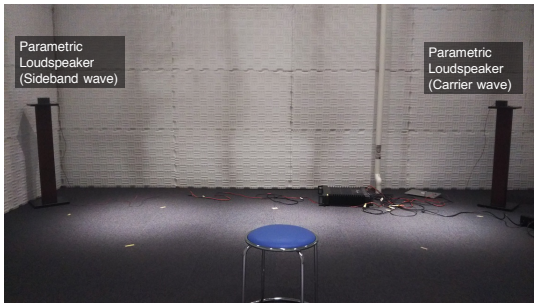

図 6 主観評価実験セットアップ

3.2 攻撃条件の仮定

攻撃対象の音声認識機器は，使用者の声の特徴を学習している場合と学習していない場合に分けられる．学習していない場合は，どのような声を使用しても問題ないが，学習している場合は，声の種類によっては起動フェーズを突破できない可能性がある．機器が利用者の声を学習済みの場合は，過去の論文にある，利用者の声を録音して合成し直す技術を利用し [6]，個人認証の壁については突破できている状態であると仮定して話を進める．

3.3 攻撃成功率の測定

実験のセットアップ図を図 5 に示す．部屋の形による影響を取り除くため，早稲田大学の音響室を使用して実験を行った．音響室の大きさは 5×5 m で，壁，天井にはそれぞれ吸音材が設置されている．搬送波，側帯波がそれぞれ部屋の中央で交差するようにパラメトリックスピーカを部屋の隅に設置し，音波を放射する．交差する角度は 90 度に固定し，音声認識機器の位置を移動させて各位置での認識成功回数を測定する．音響室の 4×4 m 四方内を 縦横それぞれ 1 m ずつに区切り，その交点に音声認識機器を置き（図 5 青点部），各地点で 10 回ずつ起動，命令コマンドをそれぞれ試す．正しい反応を返した回数を認識成功回数として記録する．部屋は通常よりも雑音の少ない部屋となっているため，ダイナミックスピーカを使用して平均 50 dB 前後の雑音を付加して実験を行うことにする．

3.4 主観評価実験

ユーザにどのように聞こえるかを評価するため，主観評価実験を行った．測定点は図 5 とそろえ，実験参加者にはそれぞれの地点に座ってもらう（図 6 参照）．聞こえるかどうかを試すために用いる音声として，ランダムな単語を 100 個用意する．実験を始める前にスピーカの高さと実験参加者の耳の高さを合わせ，事前に聞こえるかどうかを試すテストを行ったのちに実験を始める．それぞれの地点では，同じ単語が 2 度再生され，聞き取れた場合は聞き取れた単語を，聞き取れなかった場合には，「単語は聞き取れなかったが何らかの音がしたことはわかる」場合には△を，「まったく聞き取れなかった」場合には x を記入してもらう．実験参加者数は 20 人で，年齢は 19–27 歳である．その後，回答結果と本来の解答間で jaccard 係数（$= j$ とする）を計算し，以下の基準に基づいてスコアを算出する．

- 4 点: $j = 1$ （正解と完全に一致）
- 3 点: $j \geq 0.5$ （正解と概ね一致）
- 2 点: $j < 0.5$ （正解の一部と一致）
- 1 点: $j = 0$ or △ を記入
- 0 点: x を記入

4. 攻撃の実現可能性評価

4.1 各地点での攻撃成功率

各地点での攻撃成功回数を示した図を，図 7 に示す．どの結果も，音波を交差させた中央の点でのみ認識が成功していることがわかる．起動コマンドは，Google Home,

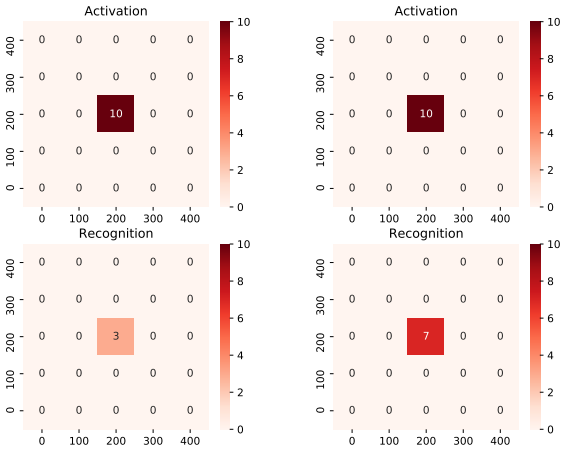

図 7 X-Audio Attack の攻撃成功回数. 上: activation, 下:recognition. 左:Google Home, 右: Amazon Echo を示している.

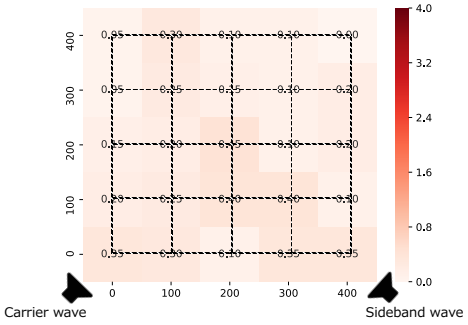

図 8 主観評価実験の平均スコア (400 × 400 cm).

Amazon Echo 共に 100%の認識率となった. 命令コマンドは Amazon Echo で 70%, Google Home で 30%の認識率となっており, 同様に攻撃コマンドが入力されてしまう可能性が示されている. 2.5 × 2.5 m の範囲で追加検証を行なった結果, Google Home, Amazon Echo, スマートフォン (SHARP SHV37) の全てで 100%起動, 命令が成功することを確認し, AI スピーカ以外の音声認識デバイスでも認識が成功することがわかる.

4.2 主観評価実験

主観評価実験の結果を図 8 に示す. すべての計測点で平均スコアは 1 を下回っており, ユーザは音声が再生されたことに気づくのも難しい状況であることがわかる. この実験は音声が再生されること, その音声を聞き取ることに集中している状態で行われていることから, 実環境ではより気づきにくい音声となることが考えられる. 音声が復調され, 攻撃が成功する Hotspot においても, 単語を正答できた人は 0 人で, 回答した単語の一部が解答と一致した人 (2 点獲得した人) が 1 人であった. このように, 攻撃が成功する地点でも気づくことが難しい攻撃であることがわかる.

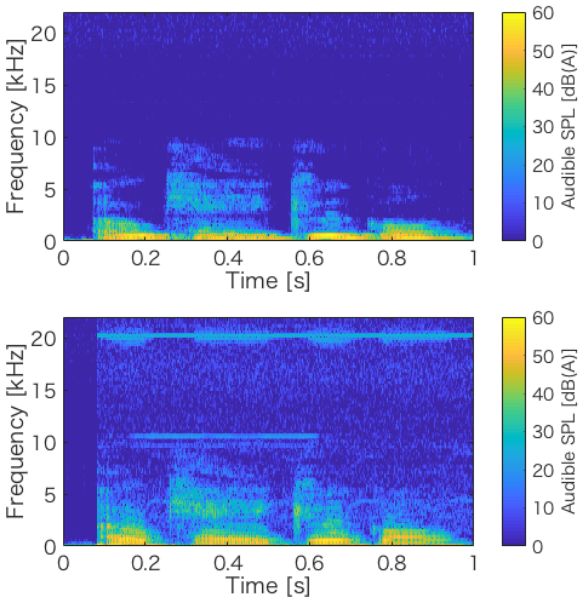

図 9 ダイナミックスピーカ (上)/パラメトリックスピーカ (下) から収録した音声スペクトログラムの比較

5. 対策

5.1 対策の概要

この章では, 本論文で示した X-Audio Attack, 2.3 章, 7 章で示されている各種攻撃を防ぐ対策手法を提案する. これらの攻撃すべて, **現状使用されている機器の多くが, 人の口から入力された音声と, スピーカから入力された音声を見分けることができず, スピーカからの音声にも応答してしまう**という事実に依存している. 我々は, 音声がどのように入力されたかを識別する識別器を生成する. この識別器の作成により, 古典的ななりすまし攻撃をはじめ, 超音波を用いた攻撃, 最新の Audio Adversarial Example までを含む攻撃を全て防ぐことが可能となる. 我々はこの識別器を Voice Liveness Detector と名付け, その性能を評価する.

図 9 はダイナミックスピーカから放射した音声と, パラメトリックスピーカから放射した音声をサンプリング周波数 80 kHz で収録した波形のスペクトログラムである. パラメトリックスピーカから収録した音声は折り返し雑音として現れる周波数に特徴があるが, 搬送波の周波数によって折り返し雑音として現れる周波数帯が変化してしまうため, 特定の周波数を監視するという手段での監視はできない. また, パラメトリックスピーカで用いられる搬送波の周波数は, 一般的に約 40 kHz 付近が使用される. 通常のマイクで収録できる音波の帯域は 10 Hz – 25 kHz となっていること, 音声認識用に収録される音声のサンプリング周波数は 16 kHz や 20 kHz に設定され, (サンプリング定理により) 受け取れる波形の周波数成分は音声帯域として 8

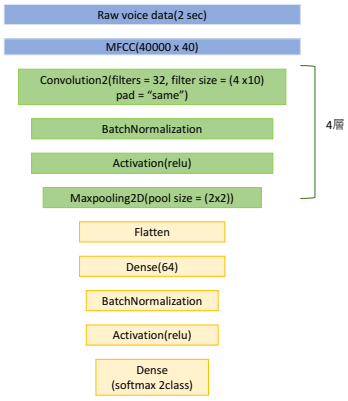

図 10　2 次元畳み込みネットワーク

kHz 以下や 10 kHz 以下に限定されることを考えると，超音波を検知するというアプローチは難しい．

そこで，機械学習技術を用いて「人の口から出てマイクに入力された音声」と「超音波から復調されてマイクに入力された音声」を分類する識別器（ultrasound voice detector）/「人の口から出てマイクに入力された音声」と「ダイナミックスピーカから放射されマイクに入力された音声」を分類する識別器（dynamic speaker voice detector）をそれぞれ作成する．これらの識別器を総称して Voice Liveness Detector と呼ぶことにする．画像の畳み込みネットワークを参考にモデルを構築し，話者認識を行う手前の処理として識別器を挿入することで，音声認識の処理が行われる前に悪質な音声を受け付けないようにする．

5.2　学習

特徴量は音声認識でよく使われるメル周波数ケプストラム係数（MFCC, Mel Frequency Cepstram Coefficients）を使用する．音声 1 フレームあたりの次元数は 40 とする．メル周波数ケプストラム係数の横軸は時間軸と同じ次元を持ち，発言内容によって特徴量の次元が変化してしまうため，音声から初めの 2 秒間を切り取って使用することにする．本研究では音声認識で使用されるサンプリング周波数として 20 kHz の音声を使用するため（ 5.3 節参照），最終的な入力の次元は 40000 × 40 となる．MFCC を取得後，各パラメタを [−1, 1] の範囲に収めるため， 訓練データ，テストデータはそれぞれ平均，分散を取り正規化を行う．

学習に使うモデルは，Keras の 2 次元畳み込みネットワークを使用する．作成したモデルを図 10 に示す．畳み込み層，プーリング層等を 4 段配置して次元削減を行ったあと，softmax 関数で識別を行う．損失関数はクロスエントロピーを使用する．学習時には 10-fold 交差検証を行い，評価を行うことにする．本モデルは，Keras で提供されていた MFCC 向けの学習モデルを元に構築している [5]．

5.3　データ

学習，テストに用いるデータは WAVE 形式の音声とし，人の口からマイクに入力された音声と，超音波から復調された音声をそれぞれ用意する．人の口からマイクに入力された音声として，株式会社 ATR-Promotions が提供する「デジタル音声データベース–セット B」を使用する [7]．音声認識，音声合成の実験に使用される音声データベースで，各音素がバランスよく含まれた文章が選定されている．データは男性声，女性声がバランスよく含まれ，アナウンサーや声優が発音している．ダイナミックスピーカから再生し，録音した音声を「ダイナミックスピーカからマイクに入力された音声」，パラメトリックスピーカから再生し，録音した音声を「超音波から復調されてマイクに入力された音声」として保存する．上記のデータのうち，80%を交差検証に用いる学習データとし，20%を交差検証後のテストに使用する．学習データ数 3076，テストデータ数 840 となった．

人の口からマイクに入力された音声には “raw voice”，ダイナミックスピーカから入力された音声には “dynamic voice”，超音波から復調された音声には “ultrasound voice” とラベル付けし，簡単のため以後ラベル名で表記することにする．録音に用いるマイクとして RION の NL-52 を使用し，元の音声データと音圧レベルが揃うように収録を行う．パラメトリックスピーカからの音声を録音する場合の距離は，復調される可聴音がもっとも大きい地点に固定する．raw voice の音声データはサンプリング周波数 20 kHz，量子化 bit 数 16（signed）となっているため，dynamic voice, ultrasound voice もサンプリング周波数と量子化ビット数を raw voice と揃えた上で録音を行うことにする．対策モデルは実環境に近いデータを使用するため，50 dB 程度の平均的な雑音を有する部屋で収録を行う．上記のデータを使用して学習にかかった時間は平均して 10–20 分程度 (iMac 2013 年モデル，CPU は Intel®Core™i5-4570R, 2.70 GHz) である．

5.4　結果

交差検証で作成したモデルそれぞれの性能評価をテストデータで行った結果，ultrasound voice detector, dynamic speaker voice detector の全てのモデルで 100%の識別率となった．テスト音声の読み込みからテスト結果を受け取るまでに要した時間は，840 個のデータに対して 32 秒であった（1 音声あたり 0.038 秒）．実際に音声認識の手前に Voice Liveness Detector を搭載する場合には，1 度の認識あたり 1 つの音声のみを検知するため，リアルタイムでの検知が可能である．また，追加のセンサなどハードウェアを使用せずに検知が可能であり，すでに販売されている商品についてもソフトウェアアップデートのみで Voice Liveness Detector の搭載が可能である．本モデルは音声認

表 1 Voice Liveness Detector の追加検証評価 (accuracy score)

	Ultrasound	Dynamic
最高精度	1.00	0.95
平均値	0.94	0.53
中央値	0.99	0.67
標準偏差	0.07	0.27

識アルゴリズムとは独立に設計されており，本モデルの適用にあたって使用する音声認識アルゴリズム（GMM-HMM, DNN-HMM 等）の種類を問わず拡張することができる．

学習データとテストデータをそれぞれ同じデータベース（ATR503）内で分けて使用しており，過学習している可能性を考え，追加で用意した実験参加者（男女 2 名ずつ）の音声を使用して性能評価を行うことにする．実用性の評価を兼ねて，収録する音声は，音声認識デバイスの起動コマンドである “OK, Google”，“Alexa” をはじめとする音声コマンドを使用した．起動後の命令をそれぞれ変え，198 個のテストデータを収集した．

追加で行なった性能評価の結果を表 1 に示す．ここで，平均や標準偏差は cross validation で作成した 10 個のモデルで評価した値の平均値である．Ultrasound voice は超音波から生じる雑音という共通の特徴を含むため，声質の違いによらず 100%の精度で検知可能であった． dynamic speaker voice は raw voice と似通っていることから精度にばらつきがあり，最高精度は 95%，平均 53%となった．モデルのパラメタは，学習に使用した話者の声質の違いにも左右されやすいと考えられる．今回は，学習に使用したアナウンサー，声優等の年齢から離れた 10 代女性の被験者を含んでいるため，平均の精度が低くなっている．実利用の際には，利用するユーザ単位で声を収集し，学習データを作成することで精度を向上させる事が可能である．

6. 議論

6.1 音声アシスタントシステムに対する攻撃シナリオ

今回の実験では，攻撃に使用するデバイスは Google 社，Amazon 社が提供しているものに限定し，使用するコマンドはすべて “What’s on my next schedule?” で統一した．このコマンドでは，利用者の予定を聞き出すことが可能である．その他に考え得る攻撃シナリオとして，

- 音量を下げるコマンドを使用して攻撃を隠蔽する
- 攻撃者の電話番号に電話をかけて盗聴する
- メッセージとして悪性の URL を送信し，拡散させる
- 特定の web サイトにアクセスするよう命令する（Amazon Echo Spot 等画面付きデバイス）
- スマートスピーカと連携した IoT 機器を不正に操作する
- 車載スピーカの音量を最大にして運転を妨害する

などが考えられる．Amazon Echo や Google Home では，ユーザがコマンドを自作することも可能である．また，様々なサービス間の連携を行う IFTTT に登録することで，各種 SNS や，クラウドサービスなど既存のサービスを操作するコマンドが増加する．さらに，Google Home では一度に複数のコマンドを実行することも可能であり，今後コマンドが増えるほど，可能な攻撃の組み合わせと自由度は増える．

表 2 特性評価，制限事項の比較

特性	Dolphin	Audio Spotlight	X-Audio
距離	×	◎	○
気づきにくさ	△	△	◎
壁の反射	△	△	◎
セットアップの負担	○	○	△

6.2 従来の攻撃と X-Audio Attack の比較

従来提案されてきた攻撃と本研究で示した攻撃 X-Audio Attack の比較を表 2 に示す．DolphinAttack，Audio Spotlight Attack は 1 つの超音波素子アレイを用いるため 2.2 章のパラメトリック現象により，鋭い指向生を持ちスピーカの正面方向には音声が聞こえてしまうという問題点があった． X-Audio Attack では，複数の超音波素子アレイを用い搬送波と側帯波を別々の超音波素子アレイから放射させ，交差する点でのみ音声が聞こえるため，従来よりも聞こえにくい攻撃となっている．また，DolphinAttack や Auido Spotlight Attack の場合，壁や障害物から反射した音波が聞こえてしまうという問題があった．X-Audio Attack では，搬送波，側帯波は分離して放射しているため，反射した地点において音声が復調されるという現象は起こらない．2 つのパラメトリックスピーカを使用するため，これまでの攻撃よりもセットアップの手間はかかってしまうが，総合的に見て従来の攻撃よりも認知されにくい攻撃が可能であり，強力である．

6.3 制限事項

本研究で提案した X-Audio Attack，Voice Livenss Detector の制限事項を述べる．まず X-Audio Attack は，2 つに分けた波形を交差させるのに十分な大きさ，形状の部屋である必要がある．細長い形状の部屋や廊下では交差させるのが難しく，通常のパラメトリックスピーカを併用する必要がある．また，壁や障害物の多い環境では，音声の到達が難しくなる可能性がある．壁，障害物を回避するアプローチとして，天井にパラメトリックスピーカを複数設置して分離放射するというアプローチが考えられる．また，音声認識機器が学習している場合には標的の音声を録音・合成した音声を使用して攻撃を行う必要がある．学習している場合には， 7 章で示す攻撃と組み合わせた上で攻撃を行う．Voice Liveness Detector は，スピーカからの音声や超音波の音声を録音する際には毎回同じ距離で，同じ機材

を使用している．今後は攻撃者がより音質の良いマイクで収録を行い攻撃を仕掛ける可能性を考え，機材の違い，距離の違い，部屋の雑音や形状の違いにも強いかどうかなど，より汎化性能を考慮した検証を行う必要があり，その検証は今後の課題である．

6.4 研究倫理

本研究は音声アシスタントシステムへの攻撃手法とその実現性の評価，ならびに対策方法を示した．その攻撃方法はいわゆるソフトウェアの脆弱性に基づく攻撃ではないが，JPCERT/CC への報告と関連する企業との調整を進め，2018 年 1 月 27 日に Google Home を発売する Google 日本法人，LINE Clova を発売する LINE 社への報告が完了した．我々の狙いは，音声認識技術を実装したシステムに潜む脆弱性を早期に発見し，警鐘を鳴らすことである．本研究以外にも音声アシスタントシステムに対する攻撃手法が報告されている（2.3 章，7 章参照）．本研究の成果が活性剤となり，このような脅威への対策を意識した音声アシスタントシステムの設計検討が進むことを期待したい．

また，攻撃の成功条件を利用者の立場から評価するため，主観評価実験を行なった．人間を対象とした実験の実施においては実験参加者に対する細心の配慮がなされる必要がある．我々は，早稲田大学が設置する研究倫理オフィスが定める「人を対象とする研究に関する倫理規程」に則り，実験参加者の聴覚や心理状態に負荷を与えないよう慎重に実験を設計した．具体的には，復調される音圧レベル，雑音として利用したピンクノイズの音圧レベルが通常の生活で経験する騒音レベルを越えないよう設計した．実験参加者からは予めインフォームドコンセントを得た上で 問題が出たらすぐに実験を中止できることを伝えた．また，10–15 分に一度，2–3 分の休憩をはさみながら主観評価実験を行った．

7. 関連研究

2.3 節で示した超音波による攻撃以外のアプローチの攻撃を紹介する．古典的な攻撃のアプローチとして，Replay Attack，Voice Conversion Attack があげられる．Replay Attack はユーザが音声コマンドを入力する声をあらかじめ録音しておき，ダイナミックスピーカを利用して機器を操作しようとするものである．Voice Conversion Attack は，音声合成を利用して利用者の声を模倣した音声コマンドを作成する攻撃である．さらに発展させた攻撃として，雑音を混ぜ合わせて利用者に気づかれないように入力を試みる Hidden Voice Commands [2]，人間には音楽や害のない音声にか聞こえないが，音声認識機器には攻撃コマンドとして認識されてしまう音声版の Audio Adversarial Examples [3,9] がある．これらの攻撃は，いずれもダイナミックスピーカや超音波素子，パラメトリックスピーカを使用しての攻撃が前提となっているため，今回我々が提案した対策手法で全ての攻撃を防ぐことが可能である．

8. まとめ

超音波が持つ伝搬特性を応用した新しい攻撃 X-Audio Attack を提案した．攻撃の実現可能性を評価し，主観評価実験により攻撃が認知されないことを確認した．また，スピーカからの攻撃を検知する識別器 Voice Liveness Detector を提案し，性能評価を行なった．今後多くの IoT 機器や自動運転車，各種家電に音声認識が搭載され，これまでに示された音声攻撃の脅威が深刻化する可能性がある．本研究で提案した対策技術は，汎用的かつ高精度に攻撃検知が可能であり，音声インタフェースを実装したシステムに広く利用される可能性がある．本研究で示した成果をきっかけとして，音声認識に関するセキュリティ・プライバシー技術の開発・研究が促進されることを期待したい．

謝辞 本研究の一部は JSPS 科研費 18K19789 の助成を受けたものです．

参考文献

[1] Amazon: Amazon Polly, `https://console.aws.amazon.com/polly/home/SynthesizeSpeech`.
[2] Carlini, N. et al.: Hidden Voice Commands, *Proc. of USENIX Security Symposium*, pp. 513–530 (2016).
[3] Carlini, N. and Wagner, D. A.: Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, *2018 IEEE Security and Privacy Workshops*, pp. 1–7 (2018).
[4] Gurbatov, S. N. et al.: *Waves and Structures in Nonlinear Nondispersive Media [electronic resource] : General Theory and Applications to Nonlinear Acoustics*, Springer, Berlin, Heidelberg, 2nd. edition (2012).
[5] Mahmood, Z.: Beginner's Guide to Audio Data, `https://www.kaggle.com/fizzbuzz/beginner-s-guide-to-audio-data`.
[6] Mukhopadhyay, D. et al.: All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines, *Computer Security - ESORICS 2015*, pp. 599–621 (2015).
[7] Promotions, A.: ATR 音声データベース B セット, `https://www.atr-p.com/products/dbpdf/bset_spec.pdf`.
[8] Yoneyama, M., Fujimoto, J., Kawano, Y. and Sasabe, S.: The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design, Vol. 73, No. 5, pp. 1532–1536 (1983).
[9] Yuan, X. et al.: CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition, *27th USENIX Security Symposium*, pp. 49–64 (2018).
[10] Zhang, G. et al.: DolphinAttack: Inaudible Voice Commands, *Proceedings of the 2017 ACM SIGSAC Conference on CCS*, pp. 103–117 (2017).
[11] 松井唯，生藤大典，中山雅人，西浦敬信：キャリア波と側帯波の分離放射によるオーディオスポット形成，電子情報通信学会論文誌 (A)，pp. 304–312 (2014).
[12] 飯島涼， 南翔太，シュウインゴウ，及川靖広， 森達哉：パラメトリックスピーカを用いた音声認識機器への攻撃と評価，暗号と情報セキュリティシンポジウム 2018 論文集 (USB).