UltrasonicWhisper+: 超音波によるヒアラブルデバイスへの可聴音生成攻撃手法の提案

渡邉 拓貴[1,a)]　寺田 努[2,b)]

概要：多くのイヤホン型デバイス（ヒアラブルデバイス）は，外部の音を取り込んでユーザに提示できる外部音取り込み機能を有している．一方で，近年の研究によりマイクの非線形性を利用することで，スマートスピーカ等へ超音波を用いて不正に音声入力できることが示されている．同様の攻撃が，外向きマイクを有するヒアラブルデバイスに適用可能と考えられる．つまり，攻撃者から超音波で送信された情報がヒアラブルデバイス内部で可聴音に復調され，内向きスピーカを通してユーザに情報提示できる可能性がある．この攻撃により，攻撃者はヒアラブルデバイスからの指示を装った虚偽の指示をユーザへ提示したり，空間からの音を装った虚偽の音情報を提示できる可能性がある．本研究では，提案手法による超音波攻撃の可能性を評価した．5 種類のヒアラブルデバイスにより提案手法で復調された音質を評価した結果，平均 MCD（Mel-Cepstral Distortion）は 7.90，MOS（Mean Opinion Score）は 2.53 であった．また，ヒアラブルデバイスからの指示を装った超音波攻撃により実験参加者がどの程度騙されるかを評価した結果，14.9%の虚偽の指示に従うことが確認された．さらに，周囲空間からの音を装った超音波攻撃により実験参加者がどの程度騙されるか評価した結果，4 種類の音源で平均 28.4%の虚偽の音を受容することがわかった．なお，上記実験では実験参加者は超音波攻撃があり得ること，それがどのような音かは事前に知っていた状態であり，超音波攻撃の危険性を軽減する対策が必要であることがわかった．

1. はじめに

近年，イヤホン型のウェアラブルデバイス（ヒアラブルデバイス）が急速に普及しつつある．調査会社の International Data Corporation によると，2024 年のウェアラブルデバイス全体の出荷数の 60%以上をヒアラブルデバイスが占めるとされている [1]．したがって，今後の社会ではより多くの人が，常にヒアラブルデバイスを装着する社会が訪れると考えられる．ヒアラブルデバイスを常時装着する社会では，ユーザは常にデバイスを介した音を聞くことになる．つまり，ヒアラブルデバイスはユーザの聴覚を代替することになり，ヒアラブルデバイス装着時特有の脅威が存在すると我々は考える．

先行研究では，超音波によってスマートフォンやスマートスピーカなどの機器に内蔵されたマイクに音声入力ができることが示されている [2], [3]．これはマイクの非線形性に基づいており，攻撃は超音波によって行われるため人間には聞こえず気づかないが，デバイス内部では超音波が可聴音に復調され，音声アシスタント（Alexa，Google Assistant，Siri 等）の操作が可能になる．外部音取り込み機能を有したヒアラブルデバイスは，外部の音を取り込むための外向きマイクと，音を提示するための内側スピーカを内蔵しているため，同様の攻撃によってユーザ聴覚へ攻撃できる可能性がある．図 1 に示すように，ヒアラブルデバイス外側のマイクが攻撃者から発信された超音波を取得し，マイクの非線形性によってヒアラブルデバイス内部で可聴音が生成され，可聴音は内向きスピーカからユーザに提示される．鋭い指向性を持つパラメトリックスピーカ [4], [5] とは異なり，ヒアラブルデバイスを装着していない人には攻撃音を聞くことができず，ヒアラブルユーザにだけ音を提示できる．したがって，周囲の人に気づかれずに，ヒアラブルユーザにだけ虚偽の情報を提示したり，聴覚を妨害することが可能になり得る．この超音波攻撃は，人々が常時ヒアラブルデバイスを身につけている環境において脅威となり得る．具体的には，以下のような脅威が想定される．

ヒアラブルデバイスからの音を装った虚偽情報提示：攻撃者は，ヒアラブルデバイスからの音を装って，ヒアラブルユーザ以外には気づかれず，非接触で虚偽の情報をヒアラブルユーザに提示でき，ユーザは虚偽情報をヒアラブルデバイスからの音情報と信じてしまう可能性がある．これ

1 公立はこだて未来大学
2 神戸大学
a) hwata@fun.ac.jp
b) tsutomu@eedept.kobe-u.ac.jp

が可能であった場合，例えばユーザがナビゲーションシステムを利用している際に，攻撃者が虚偽の案内を提示することでユーザが誤った方向や特定の場所に誘導され，犯罪に巻き込まれる可能性や（図 2a），工場等で作業中にヒアラブルデバイスからの指示を聞き作業を行う際に，攻撃者が虚偽の指示を出すことでユーザが間違った作業を行い，重大な結果を招く可能性が考えられる．さらに，意味のある音声を提示するだけでなく，単純なノイズを提示することでユーザの聴覚を妨害することが可能になり得る．

周囲空間からの音を装った虚偽情報提示：音の聞こえてくる方向と空間で反響する感覚を再現することで，超音波攻撃により耳の中に発生する音をヒアラブルデバイスからの音ではなく，外の空間で発生した音だと勘違いさせられる可能性がある（空間超音波攻撃）．音は実際にはヒアラブルデバイスから発生しているが，ユーザに周囲の空間からの音だと勘違いさせることができる．空間超音波攻撃によって，周囲には本来存在しない音を提示することでユーザの行動が操作され，思わぬ危険にさらされる可能性がある．例えば，スマートフォンなどの画面に集中して信号を確認せず，虚偽の音響信号の音を信じて横断歩道を渡ったり（図 2b），車のクラクションや人の声などを提示して注意を引くことで，スリなどの犯罪に巻き込むことが可能になり得る．

上記の脅威を念頭に，本研究ではヒアラブルデバイスからの音を装った際の超音波攻撃の性能と，空間からの音を装った際の空間超音波攻撃の性能を調査した．提案手法によって復調された音質を 5 種類のヒアラブルデバイスで調査したところ，メルケプストラム歪み（MCD: Mel-Cepstrum Distortion）は平均で 7.90，平均オピニオン評点（MOS: Mean Opinion Score）は 2.53 であった．提案手法によりヒアラブルデバイスからの音を装った指示をした場合に，実験参加者がどの程度虚偽の情報を受容するかを調査した結果，超音波攻撃があり得ることを知っている状態でも 14.9%の虚偽の指示を受け入れることが確認された．さらに，空間超音波攻撃の性能も調査した．空間超音波攻撃によってどの程度攻撃者の意図通りの位置に提示音を感じさせられるかを調査した．実験参加者周囲を 45° ずつ区切った 8 方向からの音を再現した結果，音像定位正解率は 32.0%であった．しかし，45° の回答誤差を許容すると，音像定位正解率は 65.5%であり，特に攻撃として脅威となり得る横から後ろの音像定位正解率は 80%以上であった．また，空間超音波攻撃によって提示された音と，環境から発生した音とを混同するかどうかを調査した結果，超音波攻撃が発生することと，その音がどのような音かを知っている状態でも，4 種類の音源で平均 28.4%，最も顕著な音源では平均 50.6%の攻撃音を受容してしまうことが確認された．

なお，本論文は国際会議 [6] で発表した内容に評価実験を追加し，さらに空間超音波攻撃とその評価実験を加えて手法を拡張したものである．

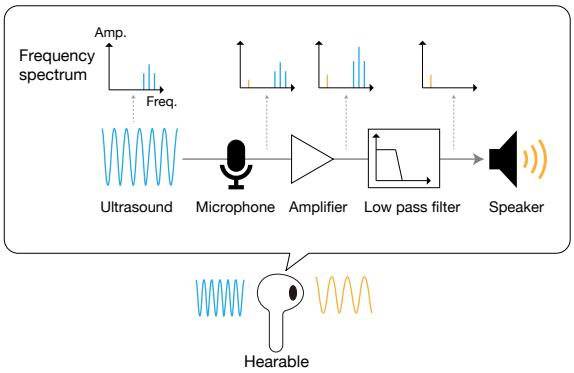


図 1: ヒアラブルデバイス内蔵マイクにおける非線形効果

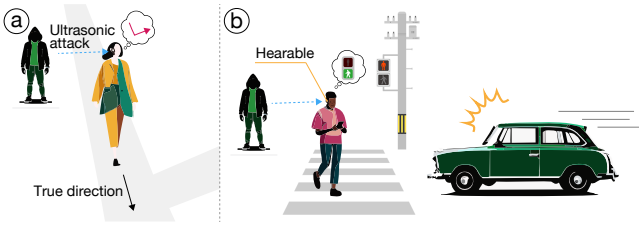


図 2: 超音波攻撃による脅威の例

2. 関連研究

2.1 超音波を利用した攻撃手法

先行研究において，マイクの非線形性を利用した手法が複数提案されている．BackDoor は，マイクの非線形性を利用して，超音波によって機器内蔵のマイクへ可聴音を生成する手法である [2]．同様に，DolphinAttack では，超音波で変調された音声指示によって，音声アシスタントが操作可能であることを示している [3]．InfoMasker は，スマートデバイスのマイクが盗聴された際に，マイクの非線形性を用いて超音波によりノイズを付加する盗聴防止システムである [7]．隠しマイクを含め，ユーザの周囲にあるマイクを無効化できるウェアラブルマイクジャマーも開発されている [8]．EchoAttack は，間接経路と直接経路を活用したヒアラブルデバイスへの超音波攻撃システムである [9]．Iijima らは，パラメトリックスピーカから発生する指向性のある音を利用した攻撃手法を提案した [10]．

これらの研究では，マイクの非線形性を利用した通信，攻撃，妨害が研究されている．本研究では，ヒアラブルデバイスの内蔵マイクに対して超音波攻撃を利用するが，この方法が機械（音声アシスタント）に及ぼす影響ではなく，人に及ぼす影響を調査している．

2.2 空間音響

空間音響に関して多くの研究がなされてきているが，近

年はイヤホンにその機能が搭載されつつある．Apple のイヤホンでは人の頭部をトラッキングすることで，頭の動きに合わせて音像定位される [11]．Sony のイヤホンでは，個人最適化によって空間オーディオの体験を向上できる [12]．これらの手法では音源はイヤホンから生じるものであり，ユーザの音楽や映画の没入感・鑑賞体験の向上を目的としている．

Moriga らは，パラメトリックスピーカによる音情報提示で方向提示する手法を提案している [4]．Kuratomo らは，視覚的に環境に影響を与えない群衆誘導管理手法として，パラメトリックスピーカを用いた立体音響による誘導音を検討した [5]．これらの研究は，パラメトリックスピーカにより音の方向を提示するという点では似ているが，ヒアラブルデバイスを装着していないユーザに焦点を当てている．本研究では，ヒアラブルデバイスのマイクの非線形効果によって超音波から発生する可聴音に着目し，これを用いて音像定位させ，生成音によってユーザを欺くことができるかどうかを調査した．

3. 提案手法

3.1 想定環境

本研究では，ユーザが外部音取り込み機能を有効にしたヒアラブルデバイスを装着する環境を想定する．提案手法の特徴は以下の通りである．

- 超音波攻撃はヒアラブルユーザにのみ聞こえ，デバイス非装着者は攻撃音が聞こえず気づくことができない．
- 攻撃者は物理的にヒアラブルデバイスへ接触する必要がなく，距離をとって攻撃可能である．
- 提案手法はヒアラブルデバイス内蔵マイクの特性を利用するため，ユーザのデバイスへの介入や，特別なデバイスを必要としない．

上記の特徴より，提案手法はヒアラブルデバイスの装着者以外には気づかれずに，ユーザのヒアラブルデバイスに情報を提示することができる．ヒアラブルデバイスの装着者は，提示された音をヒアラブルデバイスからの音と誤解し，その音の情報を信じてしまう可能性がある．また，外部音を装った音を提示することも可能であり，本来環境に存在しない音を提示することで，ユーザの行動を操作できる可能性がある．

3.2 マイクの非線形性

図 1 に，マイクの非線形性効果により超音波から生成される可聴音の概要を示す．提案手法はマイクの非線形性を利用して，超音波から可聴音を生成する．文献 [3] によれば，非線形性効果は以下の式により示される．

$$s_{out}(t) = As_{in}(t) + Bs_{in}^2(t), \tag{1}$$

ここで，$s_{in}(t)$ はマイクへの入力信号，$s_{out}(t)$ はマイクからの出力信号，A と B はそれぞれ入力信号と 2 次項 s_{in}^2 のゲインを示す．本来，マイクの出力としては，$As_{in}(t)$ のみが期待されるが，マイクの非線形性により $Bs_{in}^2(t)$ も生成される．

提案手法では，振幅変調信号を入力信号として用いる．復調してユーザに提示したい信号を $\cos(2\pi f_m t)$，変調の搬送波周波数を f_c とすると，振幅変調された入力信号は以下のように示される．

$$s_{in}(t) = \cos(2\pi f_m t)\cos(2\pi f_c t) + \cos(2\pi f_c t). \tag{2}$$

式 2 を式 1 に代入すると，入力信号には無い周波数成分（$f_c - f_m$，$f_c + f_m$，f_c，f_m，$2(f_c - f_m)$，$2(f_c + f_m)$，$2f_c$，$2f_c + f_m$，$2f_c - f_m$）が生成される．f_c がヒアラブルデバイス内部のローパスフィルタのカットオフ周波数より十分に高い周波数の場合，生成される多くの周波数成分はローパスフィルタによりカットされるが，ユーザに提示したい信号の周波数成分 f_m はローパスフィルタを通過し，ヒアラブルデバイスの内蔵スピーカを通じてユーザに提示される．

3.3 空間超音波攻撃

図 3 は空間超音波攻撃の概要を示す．人は，音の方向によって異なる頭部伝達関数（HRTF: Head-Related Transfer Function）によって音の方向を認識する．したがって，音の再現方向に対応する HRTF を適用することで，音の方向を提示することができる．本研究では，ダミーヘッド（KEMAR）で測定された HRTF を用いた [13]．スピーカから出力する信号は以下のように表される．

$$X_l(\omega) = S(\omega)H_l(\omega) \tag{3}$$

$$X_r(\omega) = S(\omega)H_r(\omega), \tag{4}$$

ここで，X はスピーカから送信する信号，S は原信号，H は HRTF，ω は周波数，添え字の l, r はそれぞれ左と右を示す．左耳，右耳に X_l, X_r の信号をそれぞれ送信することで，音の方向感を再現できる．

HRTF により方向感を与えることはできるが，利用した HRTF は無響室で測定されたものであり，部屋の壁や天井などの反響は考慮していない．この解決手法の一つとして，空間インパルス応答（RIR: Room Impulse Response）をさらに畳み込むということが考えられる．RIR は空間の音響特性を表すものであり，コンサートホールのシミュレーション等に用いられる．一般的な部屋の RIR を上記の HRTF とともに適用することで，音源の方向と空間で反響している感覚を再現できる．本研究では，公開されているデータセットのうち，本研究の実験環境に近いオフィスの RIR を用いた [14]．

3.2 節と 3.3 節で提案した要素を組み合わせることで，空

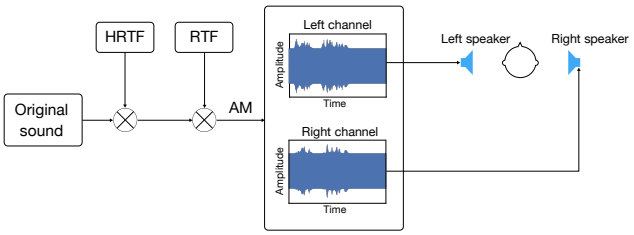


図 3: 空間超音波攻撃の概要

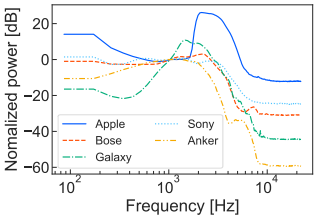


図 4: 各デバイス周波数特性

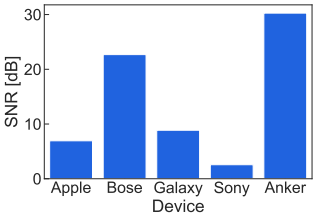


図 5: SNR

間超音波攻撃が実現できる．具体的には，3.2 節の $s_{in}(t)$ を 3.3 節で生成された信号とする．左右チャンネルに応じた信号をそれぞれユーザの左右のヒアラブルデバイスに入力することで，超音波で空間音響提示が可能となる．ここで，スピーカを正面に設置すると左スピーカから右耳に，右スピーカから左耳にも音が入力されるクロストーク音が生じるため，音像定位へ悪影響を与える．したがって，本研究では超音波スピーカをユーザの左右に設置した．超音波の指向性により左右に設置された超音波スピーカから逆側の耳への入力信号は，直接信号と比較して十分小さくなるため，クロストーク音を無視できる．

3.4 実装

システムは，PC（Apple MacBook Pro），アンプ（Fostex PC200USB-HR），超音波スピーカから構成される．超音波スピーカは，小型の超音波スピーカを並べたアレイから構成される．使用した小型超音波スピーカは UT1007-Z325R であり，200 個の小型超音波スピーカを 20cm×11cm の範囲に並べたものを使用する．なお，4.1 節，4.2 節のヒアラブルデバイスを装った攻撃実験の際には実験参加者の右側片方だけに，4.3 節，4.4 節の空間音響攻撃の実験の際には実験参加者の両側に超音波スピーカを設置した．PC から出力する超音波信号は，3 章に述べた手法に従って作成する．信号の作成には Python 3.9.7 を用いた．振幅変調の搬送波周波数は，使用した超音波スピーカの中心周波数である 40kHz とした．

4. 評価実験

4.1 基本性能

まず，提案手法の基本性能を評価した．評価では，超音波スピーカから振幅変調された超音波信号を送信し，復調された音を，ヒアラブルデバイスを装着した疑似耳を備えたバイノーラルマイク（DuoPop 2.0）を用いて録音した．ヒアラブルデバイスと超音波スピーカの距離は 1m で，超音波信号はバイノーラルマイクの右側から送信された．本論文では 5 種類のヒアラブルデバイス（Apple AirPods Pro，Bose QuietComfort Earbuds Ⅱ，Samsung Galaxy Buds Pro，Sony WF-1000XM4，Anker Soundcore Liberty Air 2 Pro）を用いた．以降，各デバイスのことを Apple，Bose，Galaxy，Sony，Anker と呼称する．

4.1.1 周波数特性

提案手法により復調される音の周波数特性を調査した．音源に 1Hz から 8kHz までを 10 秒間に線形に遷移するスイープ信号を用いた．本研究の環境ではサンプリング周波数 96kHz で，搬送波周波数を 40kHz としている．したがって，48kHz 以上はエイリアシングが発生するため，8kHz までのスイープ信号とした．

各デバイスの周波数特性を 1kHz の値で正規化したものを図 4 に示す．Bose, Sony, Anker に対しては，3kHz 付近まで比較的平坦な特性が見られた．一方で，Apple, Galaxy においては 200Hz–1kHz 付近で周波数特性が低下することが確認できた．

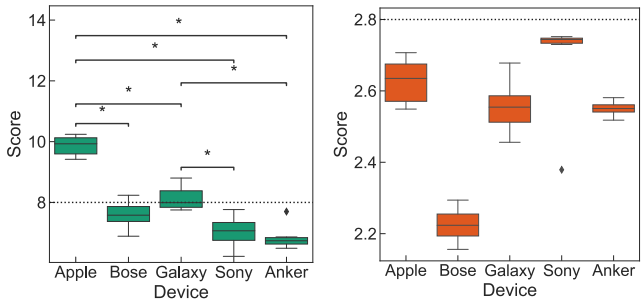
4.1.2 信号対雑音比

提案手法により復調される音の信号対雑音比（SNR: Signal-to-Noise Ratio）を調査した．音源として，3 秒のホワイトノイズを使用し，超音波信号送信時に復調された信号と，信号未送信時のノイズとの SNR を計算した．スピーカからの音圧は，ヒアラブルデバイス付近で約 95dB とした．

図 5 に各デバイスの SNR を示す．Bose, Anker では 20dB を超え，比較したデバイスの中では高い SNR が得られた．一方で Sony は SNR が 2.56 で最小となった．

4.1.3 音質

復調された音声の品質を調査した．3 種類の環境（飛行機，電車，ナビゲーション）を想定し，2 種類の声（女性，男性）で作成したアナウンス音声を使用した．評価には客観的指標である MCD と，主観的指標である MOS の 2 つの指標を用いた．MCD は音質の評価によく使われる客観的な指標である．これは 2 つのメル周波数ケプストラル係数間の歪みを数値化したもので，小さいほど歪みが小さいことを意味する．この評価では，復調された音と，変調前の原音をヒアラブルデバイスから再生し，バイノーラルマイクで録音したものとを比較した．MCD が 8 以下であれば，音声認識システムとして許容される [3]．MOS は，実験参加者が復調音を聴いて，その音質を 1（非常に悪い）から 5（非常に良い）までの 5 段階で評価する主観的な指標である．本評価では，先行研究 [15] で使用された畳み込みニューラルネットワークに基づく自動品質評価である MOSNet を使用した [16]．



(a) MCD　　(b) MOS

図 6: 復調音の音質

MCD の結果を図 6a に示す．図中の点線は，音声変換システムの評価の基準とされる 8 に引かれている．MCD に関しては，5 つ中 3 つのデバイス（Bose，Sony，Anker）で，音声変換の基準である 8 を下回った．分散分析の結果，デバイスによる有意差が確認できた ($p < .01$)．Bonferroni 法による多重比較の結果，Apple と他全てのデバイス ($p < .05$)，Galaxy と Sony ($p < .05$)，Galaxy と Anker($p < .05$) の間に有意差が確認できた．

MOS の結果を図 6b に示す．図中の点線は，元音源に対して MOSNet を適用した際に得られた平均値である 2.8 を示す．つまり，この線に近いほど劣化が少ないといえる．最も高い評価が得られたのは Sony であったが，分散分析の結果，デバイスによる MOS に有意差は確認されなかった．

上記の結果より，今回選定したデバイスでは，どのデバイスにおいても超音波による音声提示が一定の品質で可能であると考えられる．これらの評価結果を踏まえ，以降の評価では Anker Soundcore を用いた．

4.2　ヒアラブルデバイスからの音を装ったユーザ欺瞞

4.2.1　実験環境

提案手法を用いてヒアラブルデバイスからの音を装った際に，ユーザが騙されるのかを調査した．本実験では，工場等でのピッキング作業を模した環境を想定する．物品ごとに，作業員の装着したヒアラブルデバイス，もしくは外部から仕分け作業の指示が来て，作業員はその音声指示に従って，該当する箇所に物品を仕分けするような環境を想定した．そのような環境において，音声指示に超音波による虚偽の音声指示が混ざった際に，虚偽情報に従ってしまうかどうかを調査した．

実験参加者は，音声指示された番号と同じ番号のボタンをタブレット上で押下するよう指示された．音声は，ヒアラブルデバイスからの音声指示，外部スピーカからの音声，超音波攻撃により復調された音の 3 種類であり，実験参加者はヒアラブルデバイスからの音声と，外部スピーカからの音声にのみ従うように指示された．なお，実験参加者には超音波で虚偽情報が提示されることがあると教えており，

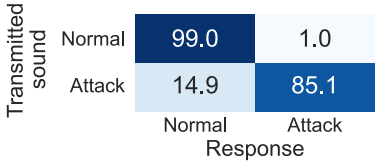


図 7: 回答結果の混同行列

虚偽情報には従ってはいけないと教示されている．指示のインターバルは 0.5 秒とし，全部で 30 個の指示（ヒアラブルデバイス 10 個，外部スピーカ 10 個，超音波スピーカ 10 個）を与えられる．ヒアラブルデバイスと外部スピーカからの音は，聴取点で約 55dB であった．超音波スピーカからは約 89dB と比較的高い音圧で音を発生し，可聴音を生成させる．世界保健機関によると，89dB の音圧を週に 5 時間以上聴くと，難聴の危険が高まる可能性がある [17]．本実験での音の聴取時間はこの基準よりも十分短かった．実験に参加したのは 10 人（男性：7 人，女性：3 人）で，平均年齢は 25.2 歳（標準偏差（SD）：4.3 歳）であった．実験を始める前に，実験参加者は実験手順に慣れるまで練習した．実験参加者はそれぞれ 5 セットの実験に参加した．結果として，1,500 個の音声指示に対する回答が得られた（30 指示 ×5 セット ×10 人）．

なお，この評価は，実験参加者が攻撃に対して注意を喚起されている環境で行われている．実験参加者は実験前に練習をしていたため，攻撃が起こりうることと，それがどのような音であるかを知っていた．

4.2.2　結果

結果を図 7 に示す．なお，途中で回答をスキップしたものは，総数から除外している．音声の指示数は通常音声と攻撃音声で総数が異なるため，各行で 100%に正規化されている．図 7 より，通常の音声が提示された場合には 99%正しく回答できている．一方で，攻撃音声が提示された場合，14.9%の回答で攻撃ではなく，通常の音声だと判断して回答していることがわかる．また，回答数は少ないが，通常の音声を攻撃と判定する回答（1.0%）も見受けられた．

4.3　空間超音波攻撃による音像定位

4.3.1　実験環境

空間超音波攻撃による音の聞こえる方向の知覚（音像定位）性能を評価した．実験参加者は 10 名（男性：8 人，女性 2 人）で，平均年齢 24.1 歳（SD：3.8 歳）である．3.3 節で提案した手法によりステレオ音源を用意する．この音源を 40kHz の超音波を搬送波として振幅変調し，図 8 に示すように左チャネルの音源を実験参加者の左耳に，右チャネルの音源を実験参加者の右耳に発信するように超音波スピーカを実験参加者の左右に設置した．実験参加者と超音波スピーカとの距離は 1m とした．送信信号には 3 秒のホワイトノイズを用い，実験参加者の周囲 45° ごとの方向（図

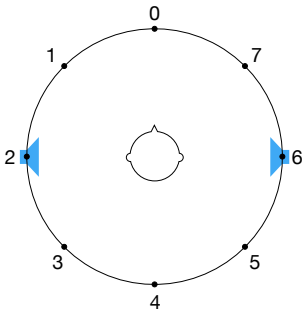


図 8: 実験環境

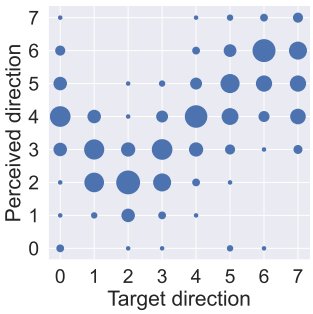


図 9: 音像定位の回答結果

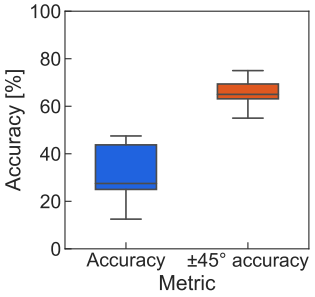


図 10: 音像定位正解率

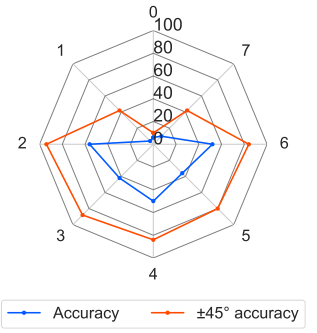


図 11: 各方向の音像定位正解率と ±45° 正解率

8 の数字の箇所）に対応する音源を計 8 種類を用意する．

実験参加者は上記の音源をヒアラブルデバイス（Soundcore Liberty Air 2 Pro）を装着し，外部音取り込み機能を有効にした状態でランダムな順番で聞き，知覚した音源方向を回答した．実験参加者は実験中に頭をできるだけ動かさないよう指示された．実験では合計で 400 個の回答が得られた（8 方向 ×5 回 ×10 人）．

評価指標として，全回答数に対する正答数を音像定位正解率と定義する．また，提案手法の利用シーンを想定した場合，おおよその方向を与えられるだけでも攻撃は可能と考える．そこで，回答の 1 つずれを許容する指標として ±45° 正解率を定義する．例えば，正解方向が 0 であった場合，0，1，7 の回答であれば正解とみなされる．

4.3.2 結果

実験結果を図 9 に示す．横軸は超音波スピーカから再現された音源方向であり，縦軸は実験参加者が感じた音像方向である．円の大きさは回答数に比例しており，この図の対角線上に回答が集まっていると，正しく方向の提示が可能であったといえる．概ね対角線上に回答が集まっているが，対角線上を中心として周囲に分散していることが確認できた．また，複数の前後混同が確認できた．つまり，方向 0 と 4，方向 1 と 3，方向 5 と 7 などの前後の取り違いが確認できた．これは，イヤホン装着状態での音像定位実験ではよく発生するものであり，避けられないものと考えられる [18], [19]．一方で，0 番方向からの音源を再現したにも関わらず，6 番と混同したり，2 番方向からの再現音源を 0 番と混同するような結果も見られた．この原因として，実験中のユーザの頭の動きによって，聞こえる左右の音量バランスが崩れたためと考えられる．

図 10 に示すように，音像定位正解率は 32%，±45° 正解率は 65.5%であった．なお，先行研究 [4] では裸耳に対してパラメトリックスピーカで音像定位実験を行っており，同程度の結果（音像定位正解率 33%，±45° 正解率 65%）が得られている．また，ヒアラブルデバイスを装着することで音像定位正解率は低下することが先行研究よりわかっている [18], [19]．したがって，ヒアラブルデバイスを経由した上での超音波情報提示としては妥当な結果だと考える．

音源方向ごとの正解率を図 11 に示す．±45° 正解率に関して分散分析を行った結果，方向による有意差が確認された ($p < .01$)．Bonferroni 法による多重比較の結果，方向 0 と 2–7 の間 ($p < .05$)，方向 1 と 2–6 の間 ($p < .05$)，方向 7 と 2–4，6 の間 ($p < .05$) にそれぞれ有意差が確認された．前方である方向 0, 1, 7 では音像定位正解率が比較的低く，特に正面である方向 0 では ±45° 誤差の許容の有無に関わらず低い値となった．この原因として，正面の音源は左右にずれた音源と比較すると両耳音量差が小さく，正しく回答することが困難だったと考えられる．一方で方向 2–6 の真横から背後にかけては正解率が比較的高く，±45° 正解率は 80%以上であった．ユーザの背後は視覚的に確認できないため，虚偽情報の提示という点で，前方よりも脅威と考えられる．これらの結果から，提案手法では特に攻撃として脅威となり得る背後からの音を装って情報提示が可能であると考える．

4.4 空間音を装ったユーザ欺瞞

前節の実験より，超音波攻撃によるヒアラブルデバイスへの方向を含んだ音情報提示が可能なことが示された．従って，本節では提案手法によって提示された音と，空間で実際に発生した音とを混同する可能性があるか調査した．

4.4.1 実験環境

実験参加者は前節と同様の 10 名である．実験参加者の周囲 8 方向（図 8 の数字の箇所）に，可聴音スピーカ (Fostex P650K) を設置し，実験参加者の左右に超音波スピーカを設置する．これらのスピーカから，どれか一つの可聴音スピーカからの音，もしくは超音波スピーカにより再現されたいずれかの方向の音が提示される．再生される可聴音スピーカの位置と超音波スピーカにより再現される位置はランダムに提示される．音源としては，実際の攻撃シナリオ等を考慮し以下の 4 種類を選定した．

- ホワイトノイズ（WN: White Noise）：全周波数帯域を含んでいるため最も特徴のある音が得られる．
- 人の呼びかけ声（Voice）：急に聞こえてきた人の声に反応し，ユーザの注意を引くことが可能になり得る．

- クラクション（Horn）：呼びかけ声同様，ユーザの注意を引くことができる．
- 音響信号機（ATS: Acoustic Traffic Signal）：街中でスマートフォン等に注視して信号機をみておらず，聞こえてきた音響信号機の音を信じて横断歩道を渡ってしまった場合に事故になり得る．

音は各方向から合計 5 回送信される．そのうち 3 回は可聴音スピーカからの音であり，2 回は超音波スピーカからの音である．1 回の音提示は約 3 秒間であり，その後 1 秒間の回答時間が与えられる．一つの音源につき 40 回の音提示がある．結果として，1,600 個の回答を得た（4 種類の音 ×40 回 ×10 人）．

実験参加者は，スマートフォンの実験用アプリ上で，知覚した音方向の番号を押下する．その際，聞こえてきた音が可聴音スピーカからの音ではなく，超音波スピーカからの攻撃であると知覚した場合，方向の番号に加えて Attack というボタンを押下するよう指示された．なお，実験参加者は実験前に超音波による攻撃があり得ることを知らされており，可聴音スピーカから聞こえる音と超音波により聞こえる音を事前に聞いていた．

評価指標として，±45° 正解率，攻撃を普通の音と勘違いした確率（FAR: False Acceptance Rate），普通の音を攻撃と勘違いしてしまった確率（FRR: False Rejection Rate）を用いた．

4.4.2 結果

図 12 に実験参加者の回答結果を示す．橙色は可聴音スピーカからの音の回答結果を，青色は超音波スピーカからの音の回答結果を示す．この図より，可聴音スピーカよりも超音波スピーカによる音の提示のほうが回答結果にばらつきがあることが確認できる．

図 13 に ±45° 正解率を示す．可聴音の平均 ±45° 正解率は 78.2%，超音波による平均 ±45° 正解率は 58.3%であった．音源の種類による違いに着目すると，可聴音，超音波ともに音源の種類によって音像定位正解率に大きな差はなかった．

図 14 に FAR と FRR を示す．音響信号機とクラクションではそれぞれ 50%程度の FAR，20%程度の FRR が確認された．一方，人の声，ホワイトノイズでは 10%を下回った．FAR について分散分析を行った結果，音源による違いに有意差が確認された ($p < .01$)．Bonferroni 法による多重比較の結果，ホワイノイズとクラクションの間 ($p < .05$)，ホワイトノイズと音響信号機の間（$p < .05$），人の声とクラクションの間（$p < .05$），人の声と音響信号機の間（$p < .05$）にそれぞれ有意差が確認された．FRR について分散分析を行った結果，音源の違いによる有意差が確認された ($p < .05$)．Bonferroni 法による多重比較の結果，ホワイトノイズとクラクションの間に有意差が確認された ($p < .05$)．

5. 議論

5.1 ヒアラブルデバイスを装った超音波攻撃

4.1 節に示すように，本論文で用いた全てのヒアラブルデバイスにおいて超音波攻撃が可能なことが示された．また，4.2 節に示すように実験参加者は虚偽情報を 14.9%信じてしまうことが確認された．実験参加者は事前に超音波攻撃があり得ることと，それがどのような音であるかを知っていたため，これらの事前知識がなければ，より攻撃の危険性が高まると考えられる．

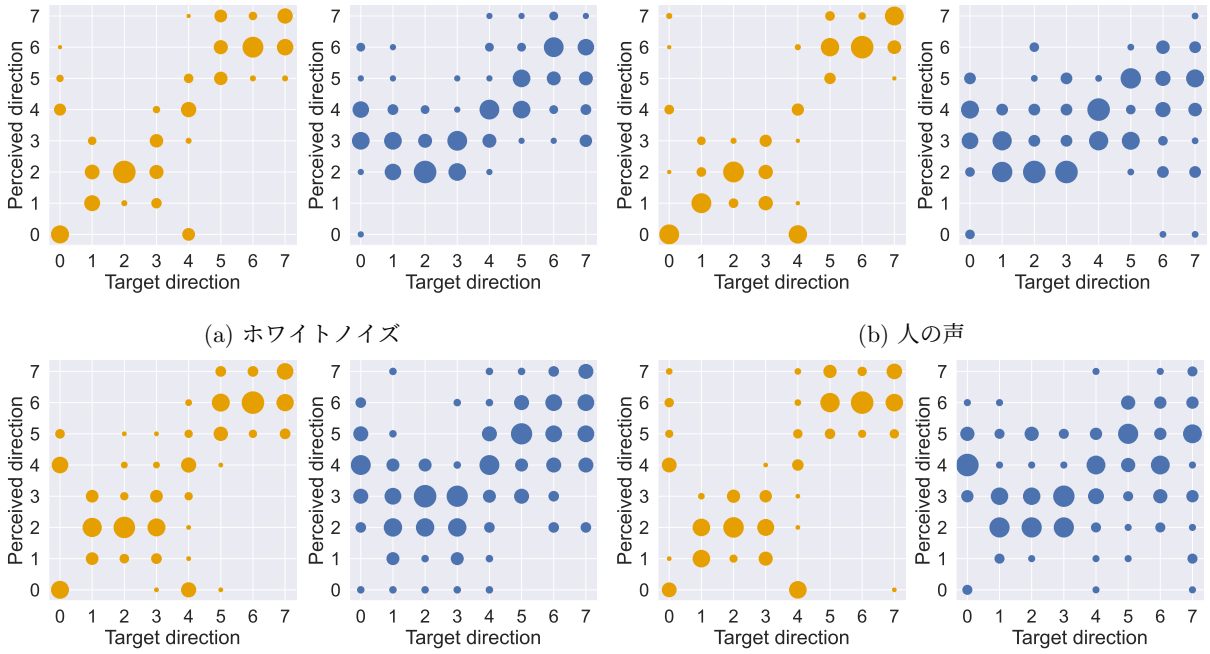
図 6b に示すように，提案手法で得られた MOS は平均 2.7 であり，これは音だけに集中していれば，本来のヒアラブルデバイスからの音であるか，超音波攻撃音か聞き分けられる音質だと考えられる．しかし，4.2 節に示すように，実験参加者は虚偽の情報を受容してしまっている．これらの結果から，作業中やスマートフォンの操作中など，ユーザが他のことに気を取られている時に，より超音波攻撃が効果的であると考えられる．

5.2 空間超音波攻撃

本稿の実験では，実験参加者は事前に超音波攻撃があり得ることと，その音がどのような音なのかを事前に知った上で実験に臨んだ．しかし，結果としてクラクション，音響信号機では 50%を超える FAR が確認されている．また，通常音なのに攻撃と判断してしまった FRR も 20%前後確認されており，実験参加者は通常音と攻撃音が区別できていないといえる．したがって，提案手法によって周囲空間からの音を装って虚偽の情報を提示することが可能であると考える．また，ホワイトノイズと人の声では他二つの音源と比較すると FAR/FRR が小さいが，5–8%程度の FAR が確認されている．今回の実験では実験参加者が事前情報を持っており，かつ連続的に提示される可聴音と超音波音を聴き比べられるような状態での実験だった．したがって，事前情報がなく，急に超音波による攻撃音が提示されれば，より攻撃音を受け入れてしまう可能性が高いといえる．以上より，提案手法により空間からの音を装ってヒアラブルデバイスに情報提示することが可能と考える．

5.3 音源による違い

4.4 節の実験では，音源によって FAR，FRR が大きく異なることが確認できた．これは，音源のもつ周波数成分によるものと考えられる．図 15 に各音源の周波数スペクトルを正規化したものを示す．ホワイトノイズは周波数全体に周波数成分が存在し，人の声は 8kHz 程度まで満遍なく周波数成分が存在する．一方で，音響信号機は 1kHz 付近に，クラクションは 3kHz 付近にピークが確認できる．図 4 に示すように，今回用いたデバイスでは 3kHz 以上で急激に周波数特性が低下する．従って，特に 3kHz 以上の周



(a) ホワイトノイズ　(b) 人の声

(c) クラクション　(d) 音響信号機

図 12: 各音源での回答結果（橙：可聴音，青：超音波）

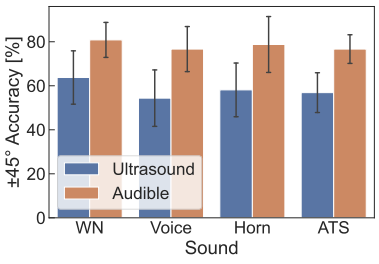


図 13: 各音源の ±45° 正解率

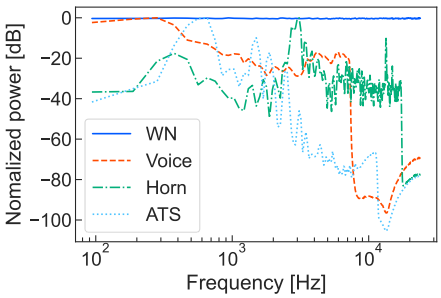
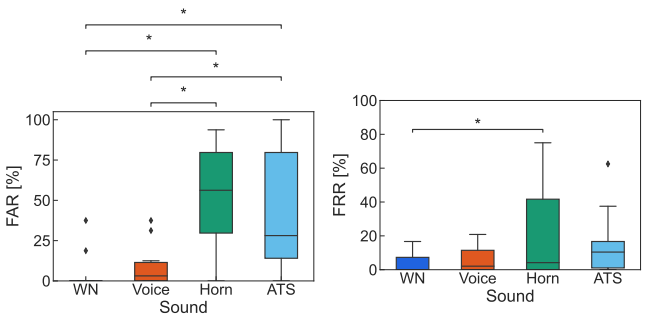


図 15: 各音源の周波数スペクトル



(a) FAR　(b) FRR

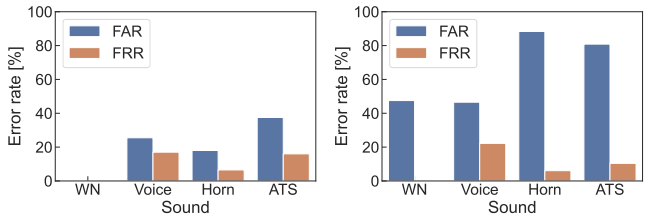
図 14: 各音源に対する FAR と FRR

波数にも成分をもつホワイトノイズと人の声は，元の音源と比較して聞き分けが容易であったと考えられる．4.4 節での評価は聞き比べであるため，より多くの周波数成分をもつ音源の方が聞き分けが容易であったが，攻撃音が突然提示された場合比較する音が存在しないため，攻撃される危険性が高くなると考える．

5.4 超音波攻撃への対策

この手法により攻撃された時に，どのように対策するかを議論する．最も簡易的な方法としては機械学習により攻撃音か否かを判別する手法が考えられる．そこで，4.4 節の環境で得られる音を用いて検証した．ダミーヘッドマイク（SAMAR Type 4700M）を 4.4 節の実験環境で実験参加者が位置する箇所に設置し，ヒアラブルデバイス (Anker Soundcore Liberty Air 2 Pro) を装着させる．このヒアラブルデバイスを経由してダミーヘッドで録音されたステレオ音源を用いてモデルを学習する．この環境でランダムに音を発生させ，得られた音源に対して，音響特徴量として一般的に用いられるメル周波数ケプストラム係数（MFCC: Mel-Frequency Cepstral Coefficients）を計算した．MFCC の次元数は 12 とし，両耳分で合計 24 次元の特徴量とした．モデル作成には自動機械学習 [20] を用いて，最も性能の良いモデルとパラメータを利用した．

音源ごとに学習した結果を図 16a に示す．ここでは，機



(a) 同音源の学習　　(b) LOSOCV

図 16: 機械学習による FAR と FRR

械は事前に各音の超音波攻撃による音と，可聴音スピーカからの音を学習しており，本稿の評価実験と同様の前提条件といえる．図よりホワイトノイズは完全に攻撃の検知が可能であるが，人の声，クラクションでは 20%前後の FAR，音響信号機では 37.5%の FAR が確認されており，4.4 節での評価実験と同様の傾向が見られる．図 16b に，leave-one-sound-out cross-validation（LOSOCV）で評価した結果，つまり，学習データとテストデータに同じ種類の音源が入っておらず，事前に攻撃音に関する知識がない状態での結果を示す．この結果はより実環境に近いものといえる．ホワイトノイズと人の声の FAR は約 45%であり，クラクションと音響信号機では 80%以上である．仮に攻撃音を事前に知っていたとしても，図 16a に示すように音源によっては 37.5%の FAR が確認されている．これらの結果から，単純な機械学習では攻撃音の検出が難しいことが確認された．

ハードウェア的な防御手法としては，マイク外側に物理的なカバーを設置する方法がある．提案手法では超音波を用いるが，超音波の特徴として可聴音よりも回折が生じにくい．したがって，マイク正面に物理的なカバーを設けることで，可聴音は音の回折によりマイクで取得可能だが，超音波を減衰可能である．予備的な検討として，マイク外側に 2cm 角のカバーを装着し，ダミーヘッド側方から音を提示した．カバー無しの場合とありの場合で，超音波では約 15.7dB の音圧減少が確認され，可聴音では変化が見られなかった．

5.5　アプリケーション

本研究では，ヒアラブルデバイスへの攻撃に着目したが，提案手法の特性を考慮すると，攻撃以外の応用も考えられる．例えば，美術館・博物館のような環境で音声ガイドとして用いることを想定する．展示の前に超音波情報提示用スピーカを置いておけば，ヒアラブルデバイス装着者にだけ音声ガイドが聞こえ，それ以外の聞きたくない人には聞こえない．ユーザは自分のヒアラブルデバイスを用いればよく，美術館側にとっては専用の機器を用意したり貸し出すコストが生じない．また，車通りの多い箇所で，ヒアラブルデバイスを装着していて車の接近に気づかない人だけに，超音波でクラクションによる注意喚起が可能と考えられる．提案手法を活用したアプリケーションの研究は今後の課題である．

5.6　制限

5.6.1　超音波スピーカからの出力音圧

本研究は，ヒアラブルデバイス内蔵マイクの非線形性を利用したが，非線形性はスピーカ側にも存在する [2]．つまり，スピーカ側の音量を上げすぎると，スピーカ側でも非線形性により可聴音が生じ，ヒアラブルデバイス装着ユーザ以外にも聞こえることになる．

音波が伝わる空気にも非線形性があり，パラメトリックスピーカとして復調されることが知られている [10]．本論文の実験においては，ヒアラブルデバイス非装着ユーザには音が聞こえないような音圧に設定して実験を行ったが，より長距離への攻撃を想定すると大音量でスピーカを鳴らす必要があり，上記の特性によりヒアラブルデバイス非装着者にも音が聞こえる可能性がある．提案手法の適用可能距離と可聴音の発生との関係の詳細な調査は，今後の課題である．

5.6.2　超音波スピーカの位置

本研究ではユーザ正面にスピーカを配置すると，クロストーク音が発生していることが確認できたため，ユーザの左右にスピーカを設置し，クロストーク音が発生しないようにした．本稿ではまずは空間超音波攻撃の概念実証のため，ユーザ左右にスピーカを設置することでクロストーク音を削除した．しかし，現実の攻撃シナリオを考えると，ユーザの左右にスピーカを設置するよりも正面や背後から二つのスピーカを用いて音を提示できる必要がある．類似技術として，既存のスピーカ等に適用されているサラウンドオーディオ技術を活用すればクロストーク音の削除は実現可能と思われる．例えば，左耳用のスピーカから右耳に入力されてしまう音を打ち消すため，これと逆位相の音波を右耳に入力することで擬似的に左右のスピーカからの音がそれぞれ左右の耳だけに入るようにすることができる．

5.6.3　ユーザの頭の動き

本稿の実験では，実験中に実験参加者の頭の位置を動かさないように指示した．これは，実験結果でも確認できたように，実験参加者が頭を動かすことによって得られる左右の音量差が生じ，意図した音像定位位置とは異なる場所に音像が生成されるためである．この解決として，攻撃対象者の頭の向きをトラッキングすることが考えられる．頭の向きをトラッキングし，耳に向けて指向性スピーカをモータ等で物理的に動かす，もしくはビームフォーミングで狙いを移動させると，攻撃対象者の頭の動きによらない音像定位が可能になると考えられる．

6. おわりに

本論文では，ヒアラブルデバイス内蔵マイクの非線形性を利用した超音波攻撃法を提案した．提案手法を 5 種類のヒアラブルデバイスに適用し，その性能を MCD と MOS の観点から評価した結果，平均 MCD は 7.90，MOS は 2.53 であった．また，提案手法を用いた音声提示によって，ユーザが騙されるかどうかを評価した．その結果，ユーザは注意喚起されても 14.9%の偽の指示に従うことが確認された．また，本研究では，超音波を用いたヒアラブルデバイス向け空間音響提示法を提案し，その実現可能性を検証した．評価の結果，ユーザの周囲 8 方向からの音を再現した場合，±45° 正解率は 65.5%であった．さらに，空間超音波攻撃によって提示された音をユーザが受容するかどうかを調べた結果，4 種類の音源で平均 FAR は 28.4%，最も影響のあった音源では平均 FAR は 50.6%であった．本研究によりヒアラブルデバイス常時装着環境の潜在的なリスクが明らかになり，その対策を講じる必要があることを明らかにしたといえる．

謝辞 本研究は JST さきがけ（JPMJPR2138），JSPS 科研費 21K11973 の助成を受けたものである．

参考文献

[1] IDC: Wearable Devices Market Insights, , available from ⟨https://www.idc.com/promo/wearablevendor⟩ (accessed 2024-10-28).

[2] Roy, N., Hassanieh, H. and Roy Choudhury, R.: BackDoor: Making Microphones Hear Inaudible Sounds, *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '17, New York, NY, USA, Association for Computing Machinery, pp. 2–14 (2017).

[3] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T. and Xu, W.: DolphinAttack: Inaudible Voice Commands, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, New York, NY, USA, Association for Computing Machinery, pp. 103–117 (2017).

[4] Moriga, M., Zempo, K., Mizutani, K., Wakatsuki, N. and Kawamoto, M.: Auditory lateralization steering independent of loudspeaker position using sound from ultrasound, *2016 IEEE 5th Global Conference on Consumer Electronics*, IEEE (2016).

[5] Kuratomo, N., Karic, B. and Kray, C.: Design and Effect of Guiding Sound for Pedestrians While Maintaining the Streetscape Perception, *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, SUI '23, No. Article 30, New York, NY, USA, Association for Computing Machinery, pp. 1–10 (2023).

[6] Watanabe, H. and Terada, T.: UltrasonicWhisper: Ultrasound Can Generate Audible Sound in Your Hearable, ISWC '23, New York, NY, USA, Association for Computing Machinery, p. 132–134 (online), DOI: 10.1145/3594738.3611379 (2023).

[7] Huang, P., Wei, Y., Cheng, P., Ba, Z., Lu, L., Lin, F., Zhang, F. and Ren, K.: InfoMasker: Preventing eavesdropping using phoneme-based noise, *Proceedings 2023 Network and Distributed System Security Symposium*, Reston, VA, Internet Society (2023).

[8] Chen, Y., Li, H., Teng, S.-Y., Nagels, S., Li, Z., Lopes, P., Zhao, B. Y. and Zheng, H.: Wearable Microphone Jamming, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, Association for Computing Machinery, pp. 1–12 (2020).

[9] Li, G., Cao, Z. and Li, T.: EchoAttack: Practical Inaudible Attacks To Smart Earbuds, *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '23, New York, NY, USA, Association for Computing Machinery, pp. 383–396 (2023).

[10] Iijima, R., Minami, S., Zhou, Y., Takehisa, T., Takahashi, T., Oikawa, Y. and Mori, T.: Audio Hotspot Attack: An Attack on Voice Assistance Systems Using Directional Sound Beams and its Feasibility, *IEEE Transactions on Emerging Topics in Computing*, Vol. 9, No. 4, pp. 2004–2018 (online), DOI: 10.1109/TETC.2019.2953041 (2021).

[11] Apple: Listen with spatial audio for AirPods and Beats, , available from ⟨https://support.apple.com/en-us/102469⟩ (accessed 2024-10-28).

[12] Sony: Sony 360 Reality Audio — So Immersive. So Real., , available from ⟨https://electronics.sony.com/360-reality-audio⟩ (accessed 2024-10-28).

[13] Watanabe, K., Iwaya, Y., Suzuki, Y., Takane, S. and Sato, S.: Dataset of head-related transfer functions measured with a circular loudspeaker array, *Acoustical Science and Technology*, Vol. 35, No. 3, pp. 159–165 (online), DOI: 10.1250/ast.35.159 (2014).

[14] Szöke, I., Skácel, M., Mošner, L., Paliesek, J. and Černocký, J.: Building and evaluation of a real room impulse response dataset, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 4, pp. 863–876 (online), DOI: 10.1109/JSTSP.2019.2917582 (2019).

[15] Liao, Q., Huang, Y., Huang, Y., Zhong, Y., Jin, H. and Wu, K.: MagEar: Eavesdropping via Audio Recovery using Magnetic Side Channel, *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pp. 371–383 (2022).

[16] Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y. and Wang, H.-M.: MOSNet: Deep Learning based Objective Assessment for Voice Conversion, *Proc. Interspeech 2019* (2019).

[17] WHO: Safe Listening Devices and Systems: a WHO-ITU Standard, , available from ⟨https://www.who.int/publications/i/item/9789241515276⟩ (accessed 2024-10-28).

[18] Marentakis, G. and Liepins, R.: Evaluation of hear-through sound localization, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, Association for Computing Machinery, pp. 267–270 (2014).

[19] Watanabe, H. and Terada, T.: Transparency Mode of Hearable Reduces Your Spatial Hearing: Evaluation and Cancelling Method to Restore Spatial Hearing, *IEEE Access*, Vol. 11, pp. 97952–97960 (online), DOI: 10.1109/ACCESS.2023.3312713 (2023).

[20] Le, T. T., Fu, W. and Moore, J. H.: Scaling tree-based automated machine learning to biomedical big data with a feature set selector, *Bioinformatics*, Vol. 36, No. 1, pp. 250–256 (2020).