EarHover：ヒアラブルデバイスにおける音漏れ信号を用いた空中ジェスチャ認識

鈴木俊汰[1,a)]　雨坂宇宙[1,b)]　渡邉拓貴[2,c)]　志築文太郎[3,d)]　杉浦裕太[1,e)]

概要：本研究では，ヒアラブルデバイスへの空中ジェスチャ入力を可能にする新しいジェスチャ認識手法 “EarHover” を提案する．本手法は，ヒアラブルデバイス特有の音漏れ現象に着目し，ヒアラブルデバイスと親和性の高いスピーカと外部マイクをジェスチャ認識に利用している．本研究では，有効な 27 種類のジェスチャの中から，信号の識別性とユーザの受容性の観点から，7 種類のジェスチャを選択した（N=10）．その後，これらの 7 種類のジェスチャとニュートラル状態の計 8 状態のデータをインイヤー型/オーバーイヤー型の 2 つのプロトタイプデバイスを用いて収集し，深層学習による認識性能の評価を行った（N=13）．実験の結果，インイヤー型/オーバーイヤー型デバイスのそれぞれの F-score は 78.7%/73.4% であった．さらに，5 種類のジェスチャに絞ることで，F-score は 6 状態で 86.2%/82.5% であった．

1. はじめに

近年の技術発展に伴い，イヤホン型のウェアラブル端末であるヒアラブルデバイスが注目を集めている．ヒアラブルデバイスは，スマートフォンと接続することで，音声アシスタントやアプリ操作，通話，音楽再生など多くの機能を使用することができ利便性が高い．ヒアラブルデバイスの操作は，スマートフォンから間接的に操作する方法と，ヒアラブル本体に直接触れて操作する方法の 2 種類がある．スマートフォンによる操作は，画面を見ながら操作する必要があり，ユーザビリティが低下する．そのため，デバイス本体のみで操作が完結するのが理想的である．これを実現するために，現在販売されているヒアラブルデバイスには，静電容量センサや感圧センサ，物理ボタンが搭載されていることが多い．静電容量センサは，感圧センサと同じように指先で操作することができるが，センサ部分に手で直接触れる必要があるため，手袋をしたままでは操作できないという制約がある．感圧センサや物理ボタンは，センサ部分を指で押下する必要があり，筐体サイズが小さい場合に押下が難しい．また，押下によって耳に物理的な負担がかかり，ノイズが発生する問題もある．また，これらのセンサは，押下する時間や回数を変化させることで操作数を定義できるが，その数には限界がある．例えば，AirPods [1] の操作数は 1〜3 回のタップ，長押し，スワイプの 5 種類に制限されている．そのため，例えば，音楽を聴きながら時間やスケジュールを確認するようなアプリケーションを実装するためには，さらに操作方法を追加する必要がある．

これらの問題を解決するために，ヒアラブルデバイスに対してハンドジェスチャ入力を可能にする研究が多く行われてきた．Xu ら [2] は，ヒアラブルデバイスに内蔵されたマイクを使用して，顔や耳付近のタップやスライドジェスチャを認識するシステムを提案している．菊地ら [3] は，ヒアラブルデバイスに取り付けた 4 つのフォトリフレクタを用いて，耳を引っ張ったときに生じる耳介甲の変形を測定するジェスチャ入力システムを提案している．しかし，これらのシステムは耳や顔に触れる必要があるため，調理中や園芸時などで手が汚れている時や，手術中など手指を清潔に保つ必要がある時に利用することが難しい．

デバイスに触れる必要のないハンドジェスチャ入力方法として，赤外線センサ [4] やカメラ [5] を使った研究がある．しかし，これらの研究では追加のセンサを導入する必要があり，実装コストがかかるうえ，小さなデバイスにセンサを実装する必要があるため，外観上のデザイン制約がある．また，ヒアラブルデバイスでは Siri [6] や Google アシスタント [7] などといった音声認識によるコマンド入力が可能な機能がある．しかし，文化的な背景として公共の場での使用を躊躇してしまうという欠点がある [8].

[1] 慶應義塾大学
[2] 北海道大学
[3] 筑波大学
[a)] shunta130904@keio.jp
[b)] amesaka@keio.jp
[c)] hiroki.watanabe@ist.hokudai.ac.jp
[d)] shizuki@cs.tsukuba.ac.jp
[e)] sugiura@keio.jp

図 1 EarHover の利用イメージ

これらの，デバイスに対する物理的な接触や追加センサの問題，社会的受容性，生活動作との混同に対する解決策として，超音波信号によるドップラーシフトを利用した，空中ジェスチャ認識手法 “EarHover” を提案する（図 1）．EarHover はヒアラブルデバイスから再生される信号の内，外側に漏れる信号（音漏れ信号）を利用し，ユーザの空中ジェスチャを認識する．音楽鑑賞や音声録音に使用できる，デバイス内蔵のスピーカと外部マイクを利用しているため，既存手法と比較してヒアラブルデバイスとの親和性が高い．具体的には，ヒアラブルデバイスから正弦波を再生することで，正弦波の音漏れ信号を発生させる．この時，ユーザがデバイス付近で空中ジェスチャを行うと，音漏れ信号がジェスチャを行う手に反射し，かつ，ドップラーシフトが発生する．空中ジェスチャは種類によって手の速度や角度，形状が異なるため，発生するドップラーシフトもそれぞれ異なる．このドップラーシフトをスペクトログラム画像で表現し，深層学習モデルによる学習を行うことで，空中ジェスチャの分類を行う．本研究では 13 人を対象に 2 種類のプロトタイプデバイスを用いた空中ジェスチャ認識実験を行い，ジェスチャ認識率の調査及び性能評価を行った．

本研究の貢献は以下の通りである：

- 音漏れ信号を利用した空中ジェスチャ認識手法を提案する．本手法はヒアラブルデバイスと親和性の高い内蔵スピーカと外部マイクのみで実装可能である．
- インイヤー型/オープンイヤー型の 2 種類のプロトタイプデバイスを用いてジェスチャ認識実験を行った．実験の結果，8 状態（7 ジェスチャ＋ニュートラル状態）の f スコアはそれぞれ 78.7% と 73.4% であることを確認した．また，認識ジェスチャを 5 種類に絞ることで f スコアは，それぞれ 86.2% と 82.5% に向上することを確認した．

2. 関連研究

2.1 ヒアラブルデバイスへの入力

ヒアラブルデバイスによるジェスチャ入力は，手を用いるハンズ入力 [2], [3], [9], [10] と頭部の動作を用いるハンズフリー入力 [11], [12], [13], [14], [15] に分かれる．

2.1.1 ハンズ入力

Xu ら [2] はデバイス内蔵のマイクを使用して顔や耳近くのタップやスライディングジェスチャの認識を行っている．考案された 27 のジェスチャから，ジェスチャの特徴量やユーザビリティアンケートを基準に 8 つのジェスチャを選定し，95.3% の認識率を達成している．菊地ら [3] はヒアラブルデバイスに取り付けられた 4 つのフォトリフレクタを利用して，耳を引っ張ることによって生じる変形を計測し，ジェスチャ認識を行っている．認識率は 88.2% を達成している．Lissermann ら [9] は，静電容量センサが搭載された耳掛け型デバイスを開発し，デバイスへのインプットエリアを拡張している．玉城ら [10] は，手の輪郭や爪の位置，指関節の角度をヒアラブルデバイスに搭載したカメラから推定するシステムを提案している．システムはデータベースから類似の画像を検索することでデバイスに触れないホバージェスチャを認識するアプリケーション例も提案している．

デバイスや耳などをタッチする必要がある既存手法は手が濡れていたり，汚れている場面，手を清潔に保つ必要のある場面では利用が困難である．また，カメラを利用した空中ジェスチャ認識手法は，センサコストやプライバシーを考慮すると現時点で実装される可能性は低い．EarHover はスピーカと外部マイクを用いて空中ジェスチャの認識が可能であり，音楽鑑賞や音声録音としても利用できる点でヒアラブルデバイスとの親和性が高く，実装が現実的である．

2.1.2 ハンズフリー入力

真鍋ら [11] はヒアラブルデバイスのイヤーチップに導電性ゴムを導入する事で眼球運動を測定しており，アイジェスチャ入力を実現している．また，Denys ら [12] は，ヒアラブルデバイスに耳道内電極を取り付け，顔の筋肉の動きから表情ジェスチャを認識するシステムを提案している．Tobias ら [13] は，スピーカと圧力センサが実装されたイヤープラグを有するプロトタイプデバイスを作製し，鼓膜の耳鳴りのような音（Tensor Tympani）を測定することで，ハンズフリー入力を実現するシステムを提案している．Wei ら [14] はヒアラブルデバイスの IMU センサと両耳の後ろに実装したマイクを利用し，顎の動きと音声を測定することで，13 種類のハンズフリージェスチャを認識するシステムを提案している．

ハンズフリー入力はハンズ入力と比較して，デバイスやデバイス周辺をタップする必要がないというメリットがある．しかし，上に述べたような顔の筋肉の動きや Tensor Tympani，歯の音などの小さな動きは日常生活の動作と混同するという課題がある．またハンズ入力と同様に，追加のデバイスが必要である．本研究では外部マイクを実装しているが，ヒアラブルデバイス内蔵のマイク性能の向上次第で，センサを追加せずに実装可能であると考えられる．

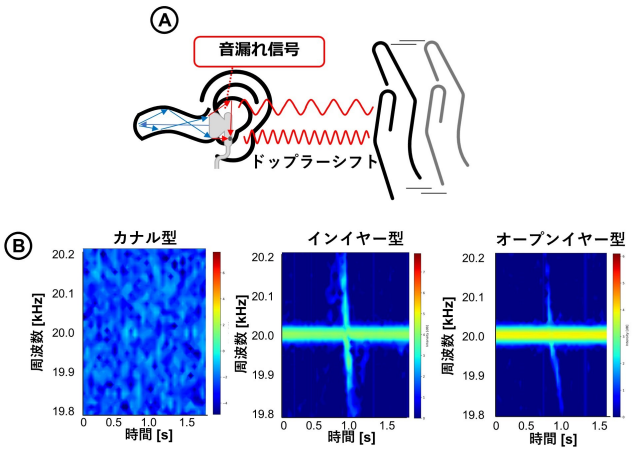

図 2 提案手法の原理．A：ジェスチャ時の音漏れとドップラーシフトの様子，B：デバイスごとの音漏れとドップラーシフトの様子

2.2 ドップラーシフトを用いたジェスチャ認識手法

本研究では空中ジェスチャを認識する手法として，音響信号のドップラーシフトを利用している．ドップラーシフトを利用した空中ジェスチャ認識は広く研究されている．Yongpan ら [16] は，スマートウォッチとスマートフォンから 19 kHz の音声信号を使用し，空中で描いた数字や英字を認識するシステムを提案している．Mingshi ら [17] は，数字の空中ジェスチャを認識するシステムを提案している．このシステムはスマートウォッチ上で実装されており，データ処理段階で 5 kHz-15 kHz のウェーブレットフィルター処理を行っている．Yang ら [18] は，スマートフォンから 20 kHz 周波数信号を再生し，口と舌の動きによって発生するドップラーシフトを 2 つのマイクで計測することで，単語認識を可能にしている．Yingcheng ら [19] はヘッドホン外部に搭載したスピーカから超音波トーン信号を再生し，相手の手話動作を認識するシステムを提案している．

我々が提案する EarHover は，音漏れ信号に着目することで市販のヒアラブルデバイスのセットアップに近いスピーカと外部マイクによる空中ジェスチャ認識システムを提案している．

3. 提案手法

3.1 認識原理とデバイスの音漏れ特性

本研究では，ヒアラブルデバイスからの音漏れ信号のドップラーシフトを利用して空中ジェスチャを認識する．ドップラーシフトは，マイクとスピーカの位置が相対的に変化することで，測定される音響信号の周波数が変化する現象である．ユーザが耳の周りで手を動かすと，音漏れ信号が手に反射する（図 2A）．このとき，手の移動する速度や角度に応じて音漏れ信号の反射特性が変化するため，ドップラーシフトのシフト幅や録音される信号の音量も変化する．この変化パターンが行う空中ジェスチャによって異なる特性を利用し，認識を行う．

ヒアラブルデバイスには主にカナル型，インイヤー型，オープンイヤー型の 3 つのデバイスタイプがあり，この順に音漏れ量が増えていく．カナル型デバイス [20], [21] は，耳道開口部を覆うイヤーチップが実装され，高い気密性があるため，Sony の wf-1000xm4 [20] や Apple の AirPods Pro [21] など，ノイズキャンセリング機能を持つデバイスに多く採用されている．インイヤー型デバイス [1], [22] はカナル型と同様に耳道開口部に挿入する必要があるが，イヤーチップが無いためカナル型よりも気密性が低く，閉塞感が軽減される特徴がある．Apple の AirPods [1] や EarPods [22] などが製品として挙げられる．オープンイヤー型デバイス [23], [24] は耳介部分にかけることで耳道開口部を塞がないため，最も気密性が低い．Victor の HA-NP35T [23] や Shokz の OpenFit [24] が製品として挙げられる．

図 2B は各デバイスタイプからの音漏れ信号とジェスチャ中のドップラーシフトの様子をまとめている．この図に示されているように，カナル型では音漏れとドップラーシフトの発生を確認できなかった．一方，インイヤー型とオープンイヤー型では音漏れおよびドップラーシフトを確認することができた．この結果に基づき，本研究ではインイヤー型とオープンイヤー型を使用して空中ジェスチャ認識実験を行う．

3.2 EarHover

図 3 に EarHover の概要を示す．一般的に市販のヒアラブルデバイスは，ハイレゾ対応の製品でない限り，再生可能な周波数帯域は 20 Hz–20 kHz であることが多い．そのため，本研究では上記の周波数帯域の内，人間に聞こえにくいとされている 20 kHz の正弦波を利用する．認識システムは空中ジェスチャ時の音漏れ信号を録音し，スペクトログラム画像に変換して，Convolutional Neural Network (CNN) による機械学習を行う．

3.2.1 データ前処理

録音した音響データに対し，ドップラーシフトのみを抽出するため 20 kHz 信号を除去するバンドストップフィルタ処理を行った．その後，ウィンドウサイズが 100 ミリ秒のハニング窓を用いて，100 サンプル分の長さで短時間フーリエ変換（STFT）を行い，スペクトログラム画像を生成する．画像に表示する周波数範囲は 19.800 kHz から 20.200 kHz とし，20 kHz 付近の情報のみを表示している．

3.2.2 データ拡張

空中ジェスチャは，ユーザが触覚フィードバックを得ることや目視での動作確認ができない．したがって，同じジェスチャでも手の位置や角度にばらつきが発生し，得られる音響データも一定でない．これらの動作ノイズに対応した CNN モデルを生成するため，スペクトログラム画像に対して 2 種類のデータ拡張方法を適用した．

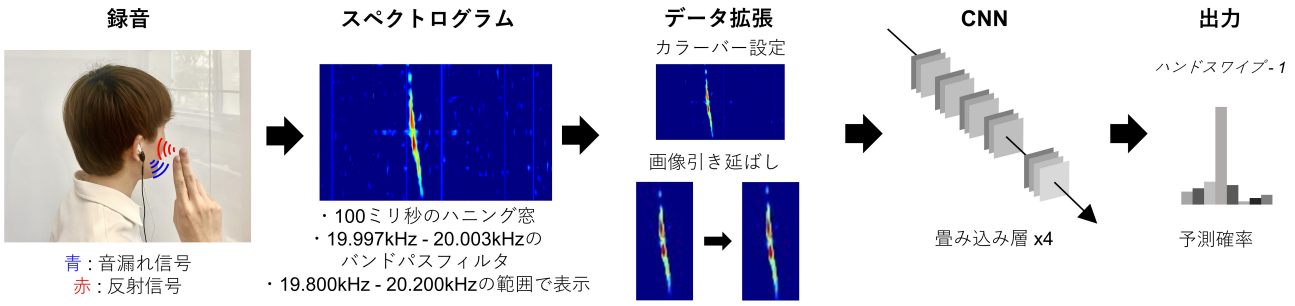

図 3 EarHover の概要

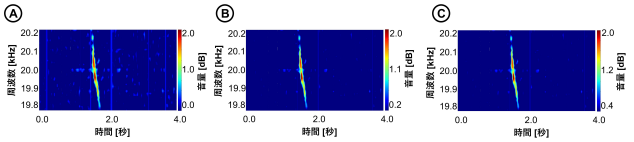

図 4 カラーバーごとのスペクトログラム画像の様子．A：0.0 dB-2.0 dB，B:0.2 dB-2.0 dB，C：0.4 dB-2.0 dB

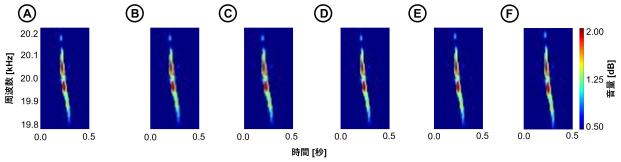

図 5 画像引き延ばし倍率ごとのスペクトログラム画像の様子．A：幅 1.0 倍（元画像），B：幅 1.05 倍，C：幅：1.1 倍，D: 高さ 1.0 倍（元画像），E: 高さ 1.05 倍，F：高さ 1.1 倍

1 つ目は，スペクトログラム画像のカラーバー調整である．本研究で用いたカラーバーは音量を RGB で表しており，音量の最小値と最大値を決定することで，最小値以下の音量は青色で，最大値以上の値は赤色で表現される．一定の音量で再生される超音波信号のドップラーシフト部分の音量は，手とデバイスの距離に応じて変化すると考えられる．したがって，生成されたスペクトログラム画像のカラーバーを調整することは，ドップラーシフトの音量変化に対するデータ拡張として効果的であると考えられる．カラーバーのパラメータ決定についての詳細は 5.2.2.1 節で述べる．図 4 では，最大音量を 2.0 dB に固定し，最小音量を 0.0 dB から 0.4 dB まで変化させた時のスペクトログラム画像の様子を示す．

2 つ目は，スペクトログラム画像の引き伸ばしである．ジェスチャの完了時間は各試行で異なると考えられる．スペクトログラムの横軸は時間の推移を表すため，横方向に伸ばすことでジェスチャの完了時間を擬似的に引き延ばす．また，スペクトログラムの縦軸は周波数を表すため，縦方向に伸ばすことでジェスチャの速度を擬似的に速めている．引き延ばしのパラメータ決定についての詳細は 5.2.2.2 節で述べる．図 5 に元画像と横方向と縦方向を 1.05 倍，1.1 倍に引き伸ばしたスペクトログラム画像を示す．

3.2.3 CNN モデル

元のスペクトログラム画像およびデータ拡張したスペクトログラム画像を利用して，CNN による深層学習を行った．バッチサイズは 32，エポックは 200 とし，学習中の val_loss が 20 エポック間で改善が見られない場合に学習を切り上げる early stopping を採用した．実装した 4 つの畳み込み層のフィルタ数は 32，64，128，256 に設定し，それぞれに Relu 関数と MaxPooling 関数を実装した．畳み込み層の後には平滑化層と 2 つの全結合層を接続した．

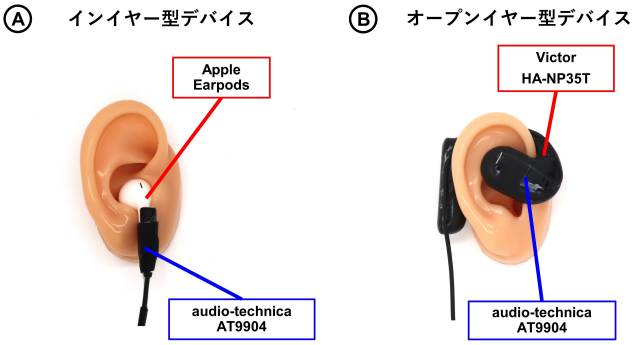

図 6 プロトタイプデバイス．A：インイヤー型デバイス，B: オープンイヤー型デバイス

4. 実装

図 6 に実装した 2 つのプロトタイプデバイスを示す．インイヤー型/オープンイヤー型デバイスの市販製品における内蔵マイクは一般に，高周波数帯域の録音ができない仕様になっている．これはマイクの利用目的が会話の録音であり，声の周波数に合わせているためである [25]．そこで，我々はヒアラブルデバイスに高周波数帯域の録音が可能な別マイクを取り付けて，EarHover のジェスチャ認識性能を調査した．本研究では，インイヤー型デバイスに Apple の EarPods [22]，オープンイヤー型に Victor の HA-NP35T [23] を利用する．超音波の周波数範囲をサポートしているマイクとして audio-technica の AT9904 [26] をデバイス外部に取り付け，音漏れ信号を録音した．音源の増幅のためにマイクアンプ (audio-technica: AT-MA2 [27]) を利用した．マイクとヒアラブルデバイスは PC（ASUS Zenbook 14 UM425QA [28]）に接続し，録音/再生を行う．この時の AD/DA 変換はオーディオインタフェース (Steinberg: UR22C [29]) を介して行った．

ヒアラブルデバイスの使用における音量設定の目安として，世界保健機関は成人が週 40 時間使用する場合の音量制限を 80 dB までと推奨している [30]．本研究では，健康リスクを考慮し，健康問題を引き起こさない 40〜45 dB の範囲で実験を行った．バンドストップフィルタでは，インイヤー型デバイスは 19.997 kHz から 20.003 kHz の範囲のフィルタを適用し，オープンイヤー型デバイスでは 19.980 kHz から 20.020 kHz の範囲のフィルタを適用した．２つのデバイスでフィルタの範囲が異なるのは，スペクトログラム画像に表れる 20 kHz 周辺部分がインイヤー型デバイスでは小さく，オープンイヤー型デバイスでは大きく表れたためである．これはインイヤー型デバイスは有線接続で信号再生が安定している一方，オープンイヤー型デバイスは無線接続で信号再生が比較的不安定であり，録音時には 20 kHz 付近に信号が大きく乱れてしまっていることが原因だと考えられる．信号録音プログラムは Python3.10 で実装し，サンプリングレートは 44.1 kHz，量子化ビットレートは 16 ビットに設定した．

5. 評価実験

5.1 実験 1：ジェスチャ選定

本実験では，Yu-chun ら [31] が提案したジェスチャデザインを参考にし，ドップラーシフトが発生すると考えられる 27 種類の空中ジェスチャ（図 7）を考案した．これらのジェスチャから，認識実験を行う空中ジェスチャセットを決定する．選定は，Xu ら [2] の手法を参考に，３種の選定過程を経て決定した．

5.1.1 実験手順

人の出入りが少なく，生活音の小さな研究室において，椅子に座った参加者に 27 種類のジェスチャを 5 セット行ってもらった．外部マイクをインイヤー型デバイスの右耳側に実装し，ジェスチャを行った際の音漏れ信号を記録した．実験は参加者の右側に物が何もない状態で行い，20kHz の正弦波のみを再生した．参加者は 11 人（男性：8 人，女性：3 人）で，全員が右利きであり，平均年齢は 27.9 歳，標準偏差は 10.7 歳であった．ヒアラブルデバイスの使用頻度は週に平均 4.13 日で，標準偏差は 2.84 日であった．参加者が日常的に使用しているデバイスは，カナル型，インイヤー型，オープンイヤー型，耳掛け型であった．この実験は 1〜2 時間程度の所要時間で，報酬は 3,000 円である．なお，この実験は慶應義塾大学理工学部の生命倫理委員会による実験承認を得ている（承認番号：2023-076）．データ記録終了後，参加者に以下の 3 つに関するアンケートを行った．評価を 1〜7 の尺度で設定し，7 を最高評価とした．（1：非常に悪い，7：非常に良い）

- 簡単さ：ジェスチャを正確に行うのはどれくらい簡単か？
- 社会的受容性：このジェスチャは社会的にどの程度受容性があると感じるか？
- 疲労感：ジェスチャーはどれくらい疲れないか？（注意：尺度は分析のために反転している）

5.1.2 ドップラーシフト領域によるジェスチャ選定

EarHover では，ドップラーシフトによるスペクトルの変化を利用しているため，ドップラーシフトの大きいジェスチャを選定する．取得した録音データは 3 節で述べた処理を行い，スペクトログラム画像に表示する音量カラーバーは最小値を 0.5 dB，最大値を 2.0 dB とした．このスペクトログラム画像を白黒画像に変換し（図 8），ジェスチャごとに画像全体に対するドップラーシフトとして現れる白部分の面積の割合を算出した（図 9）．この値が 1.5 より小さいものを削除対象とした．この条件により，以下の 9 種類のジェスチャが削除された：ハンドスワイプ - 1，近づく - 1，近づく - 2，近づく - 掌，離れる - 1，離れる- 2，離れる - 掌，指スワイプ - 上，X を描く．

5.1.3 類似性によるジェスチャ選定

残ったジェスチャセットの内，ドップラーシフトの分布が似ているジェスチャ同士は互いに誤認識する可能性がある．よって，ドップラーシフトの構造を比較することで似ているジェスチャ同士の内，一方を削除する．構造比較は，取得した各ジェスチャ 55 枚からランダムに 5 枚抽出し，比較する 2 つのジェスチャにおいて各画像を左から１ピクセルずつスライドさせ，白色部分の重なりを計算していく．重なり部分の全体に対する割合の最大値を比較し，値が大きいものを似ているジェスチャペアとした．図 10 に各ジェスチャの重なり面積の平均割合を計算し，まとめたものを示す．代表値として，算出した最大値の平均をとる．この代表値が大きいペア同士から順にアンケート調査のスコアが低い方を削除した（図 11）．この条件により，以下の 6 種類のジェスチャが削除された：ハンドスワイプ - 掌，ハンドスワイプ - 拳，ツイスト - 掌，ツイスト - 拳，離れる - 拳，グリップ - 開く．

5.1.4 ユーザビリティによるジェスチャ選定

参加者からの評価が低いジェスチャは，ユーザビリティの観点から認識しやすいジェスチャであっても避けるべきである．1〜7 の評価で行った 3 指標の平均は，簡単さは 4.89，社会的受容性は 3.96，疲労感は 4.63 であった．これを基に，簡単さが 4 または社会的受容性が 4 または疲労感が 3.5 より低いジェスチャを削除の対象とした．この条件の結果，以下の 5 種類のジェスチャが削除された：近づく - 拳，絞る - 開く，前方に円を描く，後方に円を描く，電話 - 上．

これらの 3 条件による選定の結果，以下の 7 種類のジェスチャが選定された：ハンドスワイプ - 2，ツイスト - 1，ツイスト - 2，指スワイプ - 下，グリップ - 閉じる，絞る - 閉じる，電話 - 下．図 12 に選定された 7 ジェスチャのそれぞれのドップラーシフトの様子をまとめたもの示す．

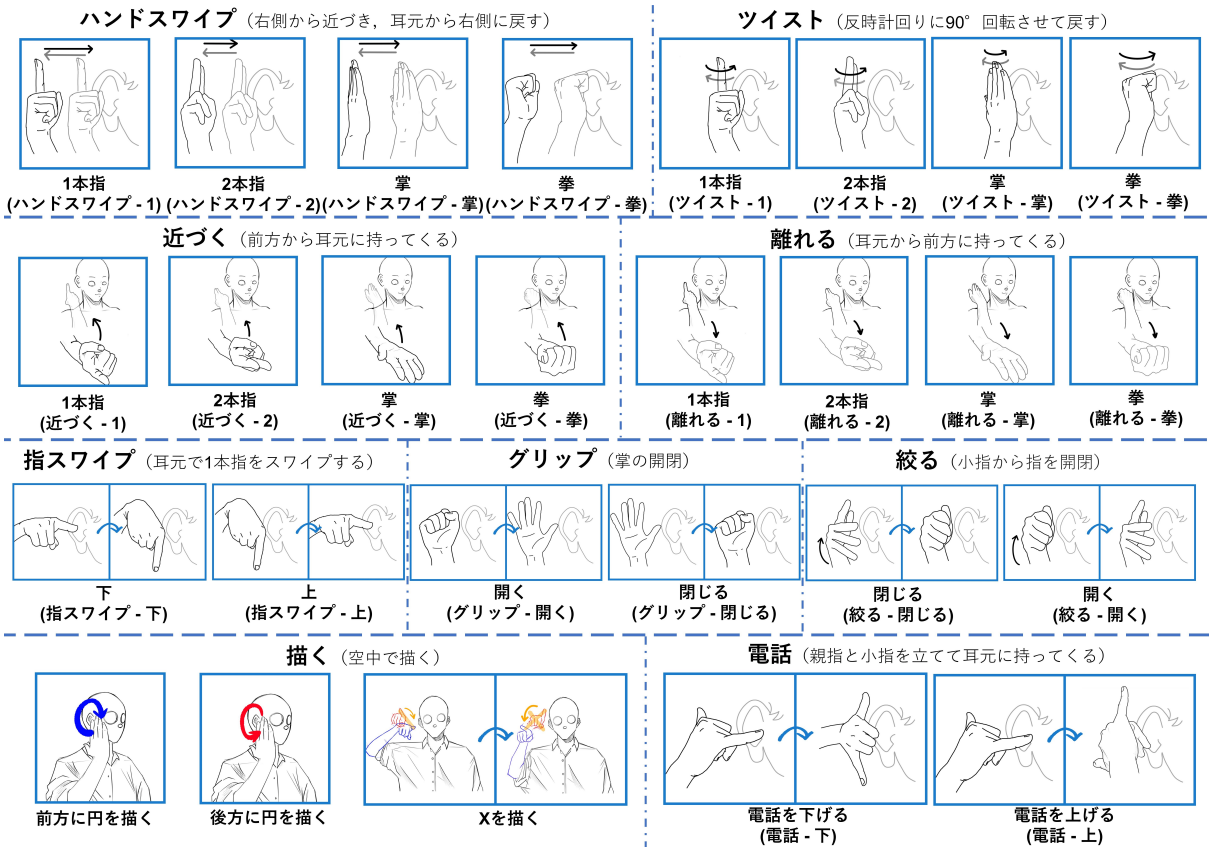

図 7　考案した 27 種類のジェスチャ一覧

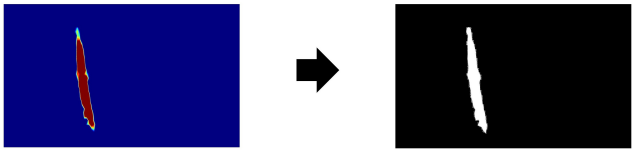

図 8　RGB 画像から変換した白黒画像

5.2　実験 2：ジェスチャ認識

5.1 節にて選定した 7 ジェスチャと何もしない状態（ニュートラル状態）の計 8 状態における提案手法の評価実験を行った．参加者は 13 人（男性：8 人，女性：5 人）で，全員が右利きであり，平均年齢は 27.5 歳，標準偏差は 9.93 歳であった．ヒアラブルデバイスの使用頻度は週に平均 4.27 日で，標準偏差は 2.64 日であった．実験は，参加者の周りに人や壁がなく人の出入りが少ない，生活音の小さい研究室で行った．この実験の所要時間は 1〜2 時間程度で，報酬は 3,000 円である．なお，この実験は慶應義塾大学理工学部の生命倫理委員会による実験承認を得ている（承認番号：2023-076）．

5.2.1　手順

安定した環境（以下，Stable）において，2 種類のデバイスでそれぞれ 160 回分（20 回× 8 状態）のデータを収集した．また，異なる環境での空中ジェスチャ入力を想定し，右手に手袋を装着した環境（以下，Gloves），参加者の右側の 50 cm の距離に人が座っている環境（以下，Neighbor），参加者が歩いている環境（以下，Walking）で，それぞれ計 32 回の試行（4 回× 8 状態）のデータ収集を行った．録音時は 20kHz の正弦波のみを再生した．

5.2.2　データ拡張デザイン

3.2.2 で述べたように，カラーバー設定と画像引き伸ばしの 2 種類のデータ拡張についての最適なパラメータを Stable のデータを基に調査する．取得した 20 回分のデータを，テストデータを 1 回分，残りの 19 回分のデータを 16 回分の訓練データと 3 回分の検証データにランダムに分割する．そして 20 分割交差検証結果の平均値を算出し，これを 3 人の参加者のデータで評価した．

5.2.2.1　カラーバー設定

インイヤー型デバイスのデータを使用して 3 つのデータ拡張方法の精度を比較した．1 つ目は，最小値が 0.00 dB から 0.50 dB まで 0.05 dB 刻みで増加するもので，0.00-0.50 と定義する．2 つ目は，最小値が 0.20 dB から 0.70 dB まで 0.05 dB 刻みで増加するもので，0.20-0.70 と定義する．3 つ目は，最小値が 0.40 dB から 0.90 dB まで 0.05 dB 刻みで増加するもので，0.40-0.90 と定義する．これにより，11 倍のデータ拡張となる．

評価としては，各交差検証で各ジェスチャについて 11 枚の結果が出るが，それぞれの画像のジェスチャ予測確率を足し合わせて，その最大値をそのジェスチャの予測ジェスチャとした．各カラーバーのニュートラルを除いた 7 ジェ

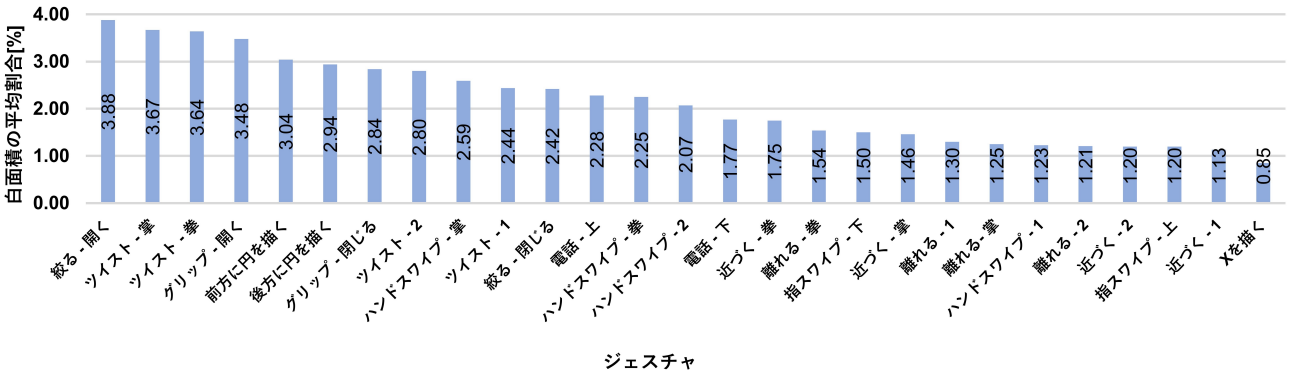

図 9　27 ジェスチャの画像全体に対する白面積割合の平均

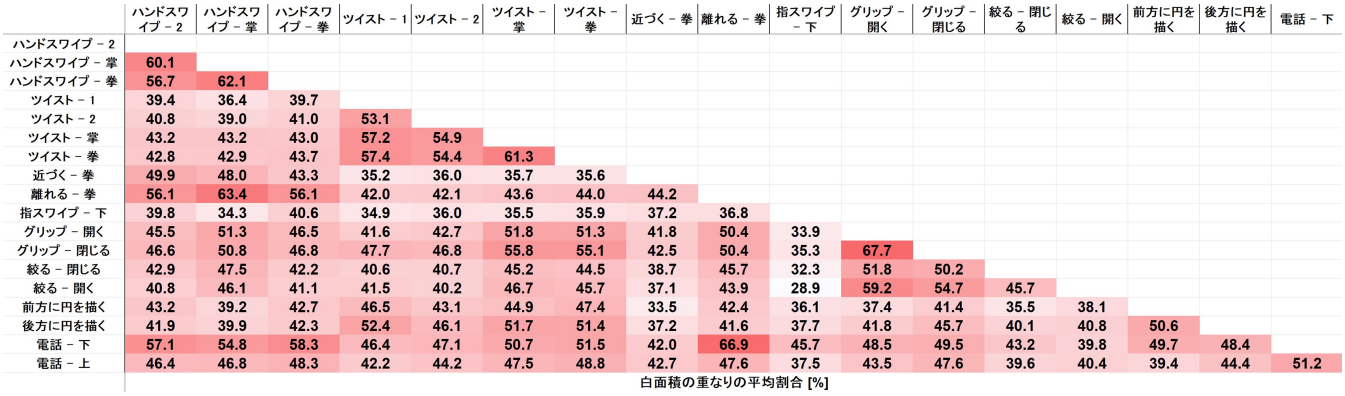

図 10　白面積の重なり面積の平均割合

スチャの認識率を調査し，平均値が最も高いものを最適なパラメータとする．図 13A に各参加者の 3 種類のカラーバー設定における 7 ジェスチャの認識率と平均認識率を示す．実験の結果から，0.00-0.50 の 73.1 % が最も高い認識率となり，以降このカラーバー設定を使用する．

5.2.2.2　画像引き伸ばし率

画像の引き伸ばしの調査は上述のカラーバー設定を使用しながら，2 つのデータ拡張方法の認識率を比較した．1 つ目は，元の画像を水平および垂直に 1.000 倍，1.025 倍，1.050 倍に引き伸ばすもので，0.025x と定義する．2 つ目は，元の画像を水平および垂直に 1.00 倍，1.05 倍，1.10 倍に引き伸ばすもので，0.050x と定義する．これにより 9 倍のデータ拡張となる．カラーバーと同様に，ニュートラルを除いた 7 ジェスチャの認識率を調査し，平均値が最も高いものをパラメータとする．図 13B に各参加者の 2 種類の引き伸ばし方法における 7 ジェスチャの認識率と平均認識率を示す．実験の結果から，0.050x の 73.8% が最も高い認識率となり，以降この画像引き延ばし率を使用する．

5.2.3　個人モデル - 7 ジェスチャ

各参加者ごとに，Stable で収集された自身のデータのみを利用して分類モデルを生成する個人モデルの認識率を調査した．2 つのデータ拡張とデータの分割は 5.2.2 節と同様である．図 14 に，各デバイスにおける全参加者の Stable の認識率をまとめた．実験の結果，インイヤー型デバイス/オープンイヤー型デバイスの平均認識率はそれぞれ 78.7%/73.4% であった．ニュートラル状態を除いた 7 ジェスチャのみの平均認識率は，それぞれ 75.8%/70.2% であった．図 16 に，各デバイスのジェスチャの分布を示す混同行列をまとめた．「ハンドスワイプ - 2」の認識率がいずれのデバイスでも最も高かった．また，「ツイスト - 1」と「ツイスト - 2」の誤認識が多く，認識率も低かった．

次に，テスト環境でのジェスチャ認識率評価を行った．認識には，先述の Stable データの交差検証から生成された 20 個の分類モデルに対して，テスト環境におけるデータから得られたスペクトログラム画像を各分類モデルに適用する．各分類モデルの予測の平均を算出し，最も平均が高いラベルを最終的な予測ラベルとして使用した．図 15 に，各デバイスにおける全参加者のテスト環境の認識率をまとめた．実験の結果，インイヤー型デバイス/オープンイヤー型デバイスの 8 状態の認識率は各テスト環境 (Gloves, Neighbor, Walking) にて，それぞれ 54.7%/58.9%, 60.3%/56.5%, 48.5%/42.9% であった．

5.2.4　個人モデル - 5 ジェスチャ

一般にヒアラブルデバイスで音楽再生アプリケーションを操作する際は，5 つのコマンド（再生/停止，音量を上げる，音量を下げる，次の曲に進む，前の曲に戻る）を用いて操作される．そこで，7 ジェスチャから 5 ジェスチャを選択し，認識率を再調査した．図 16 より，「ツイスト -

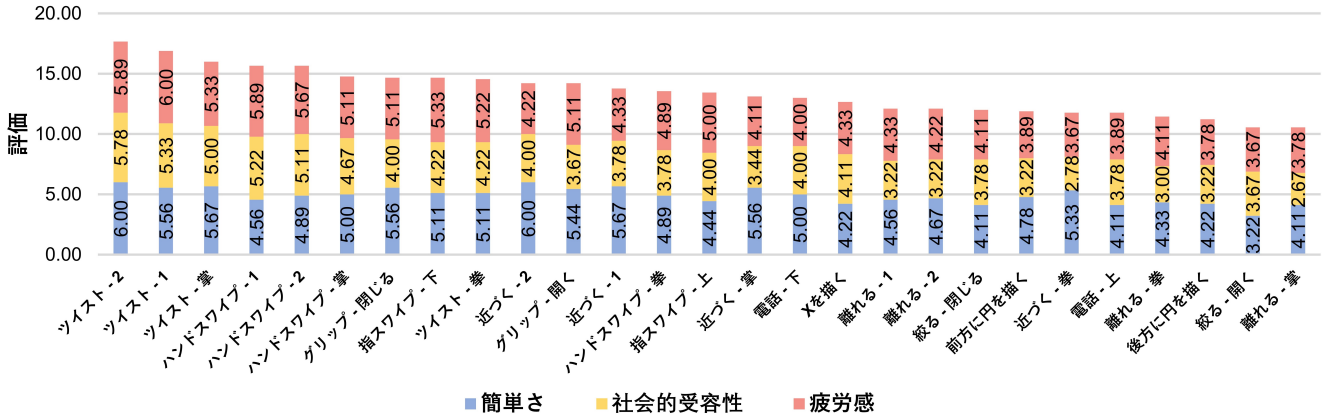

図 11 全参加者によるジェスチャ評価の平均

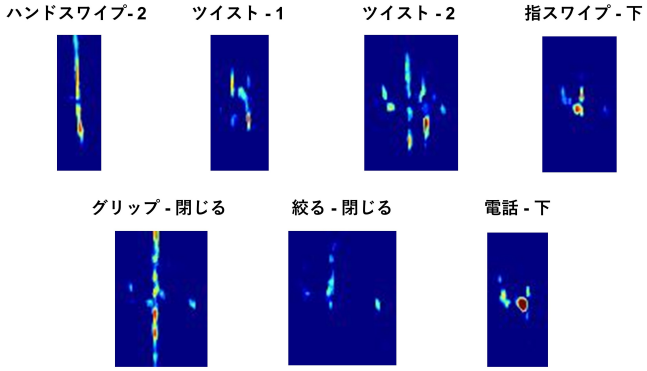

図 12 選択された 7 ジェスチャのドップラーシフトの様子

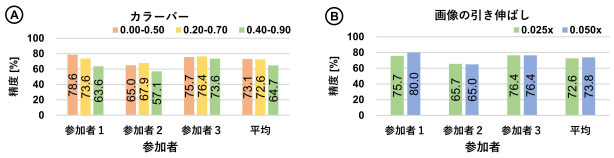

図 13 データ拡張を使用したときの認識率．A：カラーバー，B: 画像引き延ばし

1」と「ツイスト - 2」,「グリップ - 閉じる」と「絞る - 閉じる」間で混同が多かったため，各ペアで白領域（図 9）が少ない方の「ツイスト - 1」と「絞る - 閉じる」を削除し，再度認識率を調査した．図 17 に各デバイスの全ての環境の平均認識率をまとめた．実験の結果，インイヤー型デバイス/オープンイヤー型デバイスの 6 状態の認識率は 86.2%/82.5% であった．また，5 ジェスチャのみの認識率は，それぞれ 83.6%/79.2% であった．

6. 議論

6.1 システム改善

実験では，固定時間でジェスチャデータを収集し，その固定時間のスペクトログラム画像を生成して CNN で検証を行った．しかし，ジェスチャ時間は固定時間内の一部であるためデータが冗長であった．今後は，ジェスチャ時間の分布を調査し，最適なジェスチャ抽出時間を決定した上で，再度認識率の調査を行いたいと考えている．

6.2 調査範囲の拡大

本研究では調査していないが，周囲にノイズがある場合やシステムを使用しながら壁に近づく場合など，様々な実用環境での調査を行う必要がある．Neighbor と Walking での認識率を調査し，Stable よりも精度が低いことを確認したがこれは，ユーザが壁に向かって歩くような状況では精度がさらに低下する可能性があることを示唆している．データ拡張や処理を改善する（ノイズの追加など）ことで精度を向上させる方法を模索する必要がある．また，実験では正弦波のみ再生したため，同時に音楽を再生してより実環境に近い場面での精度調査を行う必要がある．

6.3 認識率の改善とユーザビリティ調査

5.2 節で述べた実験では，参加者によって認識率が大きく異なっており，認識率の高いデバイスは異なる結果となった．これは，ユーザのジェスチャ動作が不安定であり，安定したデータを得られなかったためであると考えられる．特に，ヒアラブルデバイスにおける空中ジェスチャの実行は，直接の目視確認やタッチフィードバックを得られない．安定したジェスチャ動作を得るためには，ジェスチャの練習が必要だと考えられる．また，提案手法ではユーザがジェスチャを入力したときに音声フィードバックなどを与えることも考えられる．今後は，ユーザがジェスチャの練習を重ねた場合やフィードバックがある場合のユーザビリティや認識性能を調査したい．

7. 結論

我々は，ヒアラブルデバイスのための空中ジェスチャ認識ステム EarHover を提案した．提案システムは音漏れ信号に着目することで，ヒアラブルデバイスと親和性の高いスピーカと外部マイクでの実装を可能にしている．実験は，27 種類のジェスチャから信号の識別性とユーザの

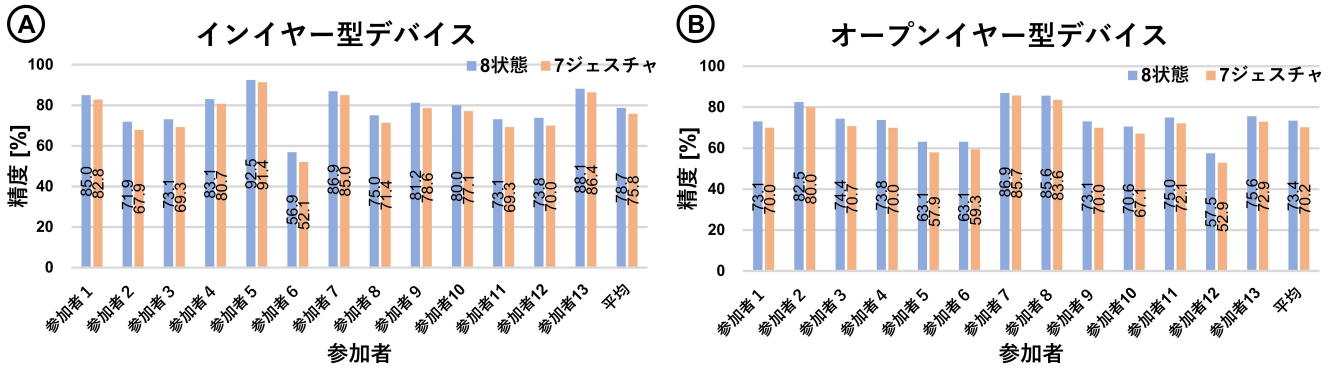

図 14 Stable における各参加者の 8 状態と 7 ジェスチャの認識精度．A: インイヤー型デバイス，B：オープンイヤー型デバイス

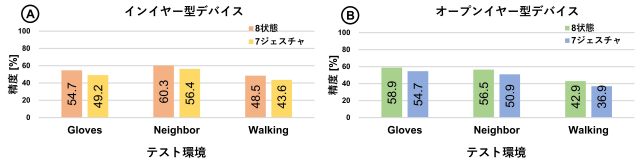

図 15 テスト環境における 8 状態と 7 ジェスチャの認識精度．A: インイヤー型デバイス，B: オープンイヤー型デバイス

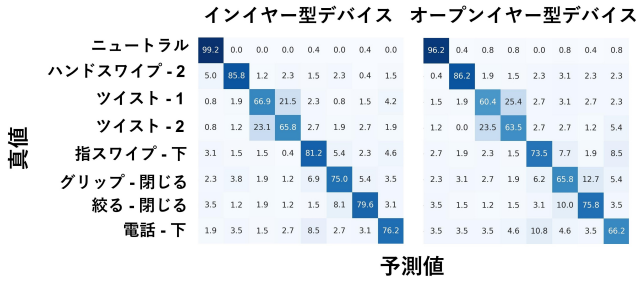

図 16 Stable の各デバイスにおける混同行列

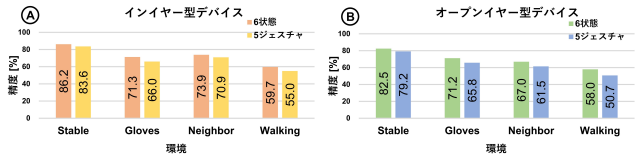

図 17 全環境における 6 状態と 5 ジェスチャの認識精度．A：インイヤー型デバイス，B: オープンイヤー型デバイス

受容性の観点から 7 種類のジェスチャを選定するジェスチャ選定実験を行った．その後，それらのジェスチャセットの認識性能調査を行った．実験の結果，インイヤー型 / オープンイヤー型デバイスにて，8 状態の認識率がそれぞれ 78.7%/73.4% であり，6 状態の認識率がそれぞれ 86.2%/82.5% であることを確認した．

謝辞 本研究の一部は，JST さきがけ（課題番号：JPMJPR2134，JPMJPR2138）および JSPS 科研費（課題番号：JP23KJ1884，JP21H03485）の支援を受けたものである．

参考文献

[1] Apple: AirPods, `https://www.apple.com/airpods-3rd-generation/` (2021). (Accessed on 09/06/2023).

[2] Xu, X., Shi, H., Yi, X., Liu, W., Yan, Y., Shi, Y., Mariakakis, A., Mankoff, J. and Dey, A. K.: EarBuddy: Enabling On-Face Interaction via Wireless Earbuds, *Proc. CHI '20 on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, Association for Computing Machinery, p. 1–14 (online), DOI: 10.1145/3313831.3376836 (2020).

[3] Kikuchi, T., Sugiura, Y., Masai, K., Sugimoto, M. and Thomas, B. H.: EarTouch: Turning the Ear into an Input Surface, *Proc. of the 19th Int'l Conf. on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3098279.3098538 (2017).

[4] Mizumata, T. and Sakamoto, R.: A Pinch up Gesture on Multi-Touch Table with Hover Detection, *ACM SIGGRAPH ASIA 2010 Posters*, SA '10, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/1900354.1900384 (2010).

[5] Sei, Y. and Shizuki, B.: Expanding One-Handed Input Vocabulary for Smartphone Using In-Air Gesture of Index Finger Captured by Rear Camera, *Asian CHI Symposium 2021*, Asian CHI Symposium 2021, New York, NY, USA, Association for Computing Machinery, p. 61–63 (online), DOI: 10.1145/3429360.3468181 (2021).

[6] Apple: Siri (2014). `https://www.apple.com/ios/siri/`.

[7] Google: Google Assistant (2016). `https://assistant.google.com/`.

[8] iprospect: The Future is Voice Activated (2018). `https://www.iprospect.com/en/jp/news-and-insights/insights/the-future-is-voice-activated/`.

[9] Lissermann, R., Huber, J., Hadjakos, A., Nanayakkara, S. and Mühlhäuser, M.: EarPut: Augmenting Ear-Worn Devices for Ear-Based Interaction, *Proc. of the 26th Australian CHI Conf. on Designing Futures: The Future of Design*, OzCHI '14, New York, NY, USA, Association for Computing Machinery, p. 300–307 (online), DOI: 10.1145/2686612.2686655 (2014).

[10] Tamaki, E., Miyaki, T. and Rekimoto, J.: Brainy Hand: An Ear-Worn Hand Gesture Interaction Device, *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, New York, NY, USA, Association for Computing Machinery, p. 4255–4260 (online), DOI:

10.1145/1520340.1520649 (2009).

[11] Manabe, H., Fukumoto, M. and Yagi, T.: Conductive Rubber Electrodes for Earphone-Based Eye Gesture Input Interface, *Proc. of the 2013 Int'l Symposium on Wearable Computers*, ISWC '13, New York, NY, USA, Association for Computing Machinery, p. 33–40 (online), DOI: 10.1145/2493988.2494329 (2013).

[12] Matthies, D. J. C., Strecker, B. A. and Urban, B.: EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input through Facial Expressions, *Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems*, CHI '17, New York, NY, USA, Association for Computing Machinery, p. 1911–1922 (online), DOI: 10.1145/3025453.3025692 (2017).

[13] Röddiger, T., Clarke, C., Wolffram, D., Budde, M. and Beigl, M.: EarRumble: Discreet Hands- and Eyes-Free Input by Voluntary Tensor Tympani Muscle Contraction, *Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3411764.3445205 (2021).

[14] Sun, W., Li, F. M., Steeper, B., Xu, S., Tian, F. and Zhang, C.: TeethTap: Recognizing Discrete Teeth Gestures Using Motion and Acoustic Sensing on an Earpiece, *26th Int'l Conf. on Intelligent User Interfaces*, IUI '21, New York, NY, USA, Association for Computing Machinery, p. 161–169 (online), DOI: 10.1145/3397481.3450645 (2021).

[15] Amesaka, T., Watanabe, H. and Sugimoto, M.: Facial expression recognition using ear canal transfer function, *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, ISWC '19, New York, NY, USA, Association for Computing Machinery, p. 1–9 (online), DOI: 10.1145/3341163.3347747 (2019).

[16] Zou, Y., Yang, Q., Han, Y., Wang, D., Cao, J. and Wu, K.: AcouDigits: Enabling Users to Input Digits in the Air, *2019 IEEE Int'l Conf. on Pervasive Computing and Communications (PerCom)*, pp. 1–9 (online), DOI: 10.1109/PERCOM.2019.8767415 (2019).

[17] Chen, M., Yang, P., Cao, S., Zhang, M. and Li, P.: WritePad: Consecutive Number Writing on Your Hand With Smart Acoustic Sensing, *IEEE Access*, Vol. 6, pp. 77240–77249 (online), DOI: 10.1109/ACCESS.2018.2880980 (2018).

[18] Gao, Y., Jin, Y., Li, J., Choi, S. and Jin, Z.: EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 4, No. 3 (online), DOI: 10.1145/3411830 (2020).

[19] Jin, Y., Gao, Y., Zhu, Y., Wang, W., Li, J., Choi, S., Li, Z., Chauhan, J., Dey, A. K. and Jin, Z.: SonicASL: An Acoustic-Based Sign Language Gesture Recognizer Using Earphones, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 5, No. 2 (online), DOI: 10.1145/3463519 (2021).

[20] Sony: wf-1000xm4, `https://www.sony.jp/headphone/special/WF-1000XM4/` (2021). (Accessed on 09/06/2023).

[21] Apple: AirPods Pro, `https://www.apple.com/airpods-pro/` (2019). (Accessed on 09/06/2023).

[22] Apple: EarPods, `https://www.apple.com/jp/shop/product/MNHF2FE/A/earpods-with-35-mm-headphone-plug` (2012). (Accessed on 09/06/2023).

[23] Victor: HA-NP35T, `https://www.victor.jp/headphones/lineup/ha-np35t/` (2022). (Accessed on 09/06/2023).

[24] SHOKZ: OpenFit, `https://shokz.com/products/openfit` (2023). (Accessed on 10/11/2023).

[25] Baken, R. and Orlikoff, R.: *Clinical Measurement of Speech and Voice*, Speech Science, Singular Thomson Learning (2000).

[26] Audio-Technica: AT9904, `https://www.audio-technica.com.hk/index.php?op=productdetails&pid=461&lang=eng` (2008). (Accessed on 09/06/2023).

[27] Audio-Technica: AT-MA2, `https://www.audio-technica.co.jp/product/AT-MA2` (2002). (Accessed on 09/06/2023).

[28] ASUS: Zenbook 14 UM425QA, `https://www.asus.com/laptops/for-home/zenbook/zenbook-14-um425-qa/` (2022). (Accessed on 09/06/2023).

[29] steinberg: UR22C, `https://www.steinberg.net/audio-interfaces/ur22c/` (2019). (Accessed on 09/06/2023).

[30] WHO: Safe listening devices and systems: a WHO-ITU standard, `https://www.who.int/publications/i/item/9789241515276/` (2019). (Accessed on 09/07/2023).

[31] Chen, Y.-C., Liao, C.-Y., Hsu, S.-w., Huang, D.-Y. and Chen, B.-Y.: Exploring User Defined Gestures for Ear-Based Interactions, *Proc. ACM Hum.-Comput. Interact.*, Vol. 4, No. ISS (online), DOI: 10.1145/3427314 (2020).