ハンズフリーのデバイス操作のための 汎用イヤラブルデバイスのIMUセンサーを用いた 表情認識手法

北森 迪耶[1,a)]　坪内 孝太[2,b)]　西尾 信彦[3,c)]　西山 勇毅[4,d)]　下坂 正倫[1,e)]

概要：本研究では，汎用イヤラブルデバイスに内蔵される IMU センサーを用いて，表情認識が可能となる手法を提案する．表情認識は，ハンズフリーでのデバイス操作や障害者支援システムなど，幅広い応用が期待できる研究分野である．しかし，先行研究では，カメラで撮影した画像を用いた表情認識や外部センサーを取り付けたカスタムデバイスを使用した表情認識が盛んであり，プライバシー問題や日常生活動作の障害となる可能性を孕んでいる．そのため，表情認識技術は商業的な利用が多く，日常生活での活用事例は少ない．そこで，本研究では，表情にコマンドを割り当てることで日常的なハンズフリーデバイス操作を可能とするためのイヤラブルデバイスによる表情認識手法を提案する．提案手法では，汎用イヤラブルデバイスである”AirPods”を使用して，8 人の被験者から 5 種類の表情変化を記録した時系列データを用いて，機械学習モデルによる表情認識の評価実験を行う．その際，訓練データに表情認識対象者のデータを含める場合と含めない場合の 2 種類の評価実験により，ユーザー依存性の有無も考慮した汎用イヤラブルデバイスでの表情認識の可能性を示す．

1. 序論

表情認識技術は，表情にコマンドを当てはめることで，ハンズフリーでのデバイス操作が可能となり，障害者や医療機関でのコミュニケーション支援システムとしても幅広い応用が期待されるため，研究が盛んに行われている．特に，昨今のコロナ禍を経て，非接触が重要となる場面が想定される機会が増え，研究の必要性が高まっている．しかし，既存の手法では，対象者を撮影した画像による表情認識が多く，画像におけるプライバシーや照明条件の問題が未解決である [1][2][3]．そのため，現在の表情認識手法の活用事例は，小売業界による消費者の満足度の定量化や医療機関での患者の健康状態の把握など，商業的利用がほとんどである．

それに対して，画像ベースの表情認識のプライバシー問題や照明条件などの問題点を解決し，表情認識の常用化を目指す手段として，近年，イヤラブルデバイスを利用した表情認識手法が多く研究されている [4][5][6][7]．イヤラブルデバイスによる表情認識では，外部センサーを取り付けたカスタムイヤラブルデバイス [4][5] や学術研究向けに作成されたイヤラブルデバイス [6][7] を使用して，表情変化に伴う顔面筋の動きや輪郭の変化を捉えることで表情認識を可能にしている．これらの手法は，デジタル化が進む社会の中で，急速に，かつ，広く普及したイヤラブルデバイスを用いることで，表情認識の日常的活用やセンサーを利用した表情認識の可能性の提示に加え，プライバシー問題，照明条件の問題点を解決した手法である．

一方で，カスタムイヤラブルデバイスは，デバイスの大きさや重量が外部センサーの取り付けにより増加するという問題が生じる．このことから，日常生活で使用するという観点では，生活動作の障害となり，表情認識の常用化の実現が難しい．また，学術研究向けに作成されたイヤラブルデバイスは，一般には手に入れることができず，表情認識の普及における根本的な問題点の解決には至っていない．

そこで, 本研究では, 汎用イヤラブルデバイスである”AirPods”を使用し，内蔵されている 2 種類の IMU センサーから時系列データを取得することにより，表情認識を可能にする手法を提案する．汎用イヤラブルデバイスを使用することで，表情認識の日常的な利用の根本的な問題であ

[1] 東京工業大学
[2] LINE ヤフー株式会社
[3] 立命館大学
[4] 東京大学
[a)] kitamori@miubiq.cs.titech.ac.jp
[b)] ktsubouc@yahoo-corp.jp
[c)] nishio@is.ritsumei.ac.jp
[d)] yuukin@iis.u-tokyo.ac.jp
[e)] simosaka@miubiq.cs.titech.ac.jp

© 1959 Information Processing Society of Japan

る，生活動作への干渉や一般に入手困難なデバイスを使用しているという点を解消することが可能である．実際に”AirPods”から取得可能なデータを使用して表情認識を行うことで，実現可能性を提示し，その際，イヤラブルデバイスの IMU センサーを使用する上で，重視すべきデータについても言及する．加えて，表情認識の対象者を訓練データに含めた場合と除いた場合の 2 種類の条件で表情認識精度の評価を行うことで，個人による表情の違いについても考慮する．これらの評価により，汎用イヤラブルデバイスに内蔵された IMU センサーを使用した表情認識によるハンズフリーデバイス操作の日常的な利用の可能性について検討する．

本研究の貢献は以下である．

貢献

- 音声を聴くことを主目的として日常的に利用される汎用イヤラブルデバイスに内蔵される IMU センサーを用いた表情認識手法を提案する．
- 複数の評価プロセスにおける実験により，提案手法の有効性とともに，日常生活を阻害することのない表情認識によるハンズフリーデバイス操作の可能性を示す．

関連研究

画像を用いた表情認識

表情変化を通じて，感情表現や非言語によるコミュニケーションの促進のために，ユーザーの前面にカメラを設置して，表情認識を行う研究は，古くから広範囲にわたって研究されている [1][3]．これらの技術は非常に高い精度での認識を行える代わりに，カメラの携帯性の低さや照明条件，遮蔽物問題に加えて，撮影した画像におけるプライバシーの観点から，実生活で利用するには，導入の障害が多い．そのため，これらの技術は，病院での患者の健康管理や小売業における顧客満足度の把握などの商業的利用に留まっている．

カスタムイヤラブルデバイスを用いた表情認識

画像ベースの手法における日常的利用の問題点を解決するための手法として，センサーを用いて表情認識を行う手法が提案されている．特に，近年，SNS や動画配信サービスの浸透により，急速に普及したイヤラブルデバイスにセンサーを取り付けたカスタムデバイスを用いた表情認識が盛んに研究され始めている [4][5][8]．イヤラブルデバイスに付けられるセンサーとしては，スピーカーとマイクによる音響センサー [4] や PPG センサー [5] などがある．しかし，これらの手法は，イヤラブルデバイスの重量の増加やセンサー設置のコストの面を考慮する必要があり，実用化に至っていない．

学術研究用イヤラブルデバイスを用いた表情認識

さらに，カスタムイヤラブルデバイスだけでなく，汎用イヤラブルデバイスとほぼ同型のワイヤレスイヤラブルデバイスである eSense[9] を用いた研究も行われている [6][7]．このデバイスには，加速度計とジャイロスコープの 2 種類の IMU センサーが装備されており，これらのセンサーを用いて表情認識を行っている．しかし，このデバイスを用いて行っている実験は，限りなく日常的な利用を前提としたものであると言えるが，使用している eSense というデバイスは，個人規模の学術研究用のデバイスであり，一般には入手することができないため，実用化には至っていないという現状がある．

2. イヤラブルデバイスを用いた表情認識

2.1 イヤラブルデバイスへの外部センサーの設置

イヤラブルデバイスに外部センサーを装備したカスタムデバイスによる表情認識では，汎用イヤラブルデバイスに音響センサー [4] や PPG センサー [5]，さらには，小型カメラ [2] を取り付けている．これらのセンサーを利用することで，表情変化に伴う輪郭の変化を捉え，表情認識を行っており，外部センサーを表情認識のために設置していることから，センサーの特徴を最大限発揮して．高い精度での表情認識が可能となっている．このことから，表情認識技術においては，表情変化を直接記録するカメラベースの手法でなくとも，認識可能であり，特に，表情変化と輪郭や表情筋の変化は密接に関連しており，イヤラブルデバイスで表情変化を捉えることが有効であることを示している．

しかし，音響センサーを利用する研究 [4] では，センサーによる重量やコストの増加に加え，髪の毛が長い人は音響信号が阻害されてしまったり，歩行中や周囲の音が大きい場合の外部環境によるノイズの影響が問題視されていたり，PPG センサーを用いる研究 [5] においても，デバイスの大きさやコストの問題が言及されているだけでなく，血管内の血流の体積変化を光で読み取る PPG 信号では，照明条件や個人の体調からも影響が出ることが考えられる．したがって，イヤラブルデバイスによる表情認識技術の日常的な利用においては，普及と技術の 2 種類の課題が残っている．

2.2 学術研究用デバイスの使用

カスタムイヤラブルデバイスにおける課題に対して，技術的に外的要因による精度悪化を回避しながら，普及も前提とした研究が，個人規模の学術研究用に複数の研究機関が合同で作成した eSense[9] というデバイスを用いて行われている．このデバイスには，加速度計やジャイロスコープが内蔵されており，追加のセンサーをつけることなく，表情認識を行うことができる．実際に，eSense を利用し

© 1959 Information Processing Society of Japan

た表情認識に焦点を当てた研究 [6] では，被験者の依存性等を考慮しない場合，非常に高い精度の結果が得られており，イヤホンに内蔵されている IMU センサーを用いることでも表情認識が可能であることを示している．しかし，eSense は学術研究において，イヤラブルデバイスを用いた行動認識技術の向上を図ることが目的のデバイスであるため，一部の研究機関で研究活動に使用されているだけであり，日常的に使用されるデバイスのように普及することは難しい．

加えて，センサーを用いた行動認識を前提としているイヤラブルデバイスであるため，センサー自体の性能は非常に高く，音楽を流していない場合には 50Hz と高い周波数でのデータ収集が可能となるが，音楽を流している場合には，流していない場合と同様の周波数でのデータ収集ができることを保証していないとされている [9]．このことから，あくまでカスタムデバイスの重量や大きさの問題を解決した行動認識用のデバイスであり，イヤラブルデバイスの本来の目的である，音声を聞くという役割を十分果たしているとは言えない．日常生活をしている上で，イヤラブルデバイスをつける場面の多くは，音楽を聴いたり，動画を見たり，音声を聴くという行為が主動作となる．したがって，本来のイヤラブルデバイスの役割である音声を聴くという機能を妨げることなく，表情認識をする必要がある．

2.3 先行研究における問題点

イヤラブルデバイスによる表情認識は，現状，画像ベースの表情認識手法の課題の解決へ向かっているが，前述した通り，依然として課題は残っている．特に，カスタムデバイスや学術研究用デバイスであることによる普及面の課題とセンサーの外的要因による技術面の課題の 2 種類がある．それらの課題に対して，イヤラブルデバイスを用いることの最も大きいメリットとして，近年，急速に普及してきたために，非常に身近なウェアラブル端末として利用可能であるという点が挙げられることから，ハンズフリーデバイス操作を目的とした，表情認識技術の普及という面に対する課題解決に臨むことが最優先事項であると考えた．

常用化の手段として，これまでの表情認識に用いられたイヤラブルデバイスは，一般には手にすることのできないものであったことを考慮し，本研究では，既に一般に流通しているイヤラブルデバイスで，かつ，特別なカスタマイズをすることなく，表情認識を行うこととした．採用したイヤラブルデバイスは，ワイヤレスイヤホン市場で世界を席巻していて，かつ，スマートフォンとの親和性の高さがあり，将来的なハンズフリーデバイス操作の需要が期待される”AirPods”である．本デバイスは，イヤラブルデバイスの音声を聴くという本来の機能が前提のデバイスであり，かつ，その機能を損ねることなく，表情認識を行える可能性を持つ．つまり，音声を聴く目的で人々が購入している汎用イヤラブルデバイスを用いた表情認識の可能性を示すことにより，真に，表情認識技術を社会に普及させ，人々から受け入れられることの一助となると考えられる．本研究では，”AirPods”に装備されている eSense[9] と同じような IMU センサーを用いて表情認識が可能であることを提案する．

3. 提案手法

3.1 汎用イヤラブルデバイスを用いた表情認識

汎用イヤラブルデバイスである”AirPods”には，加速度計とジャイロスコープの 2 種類の IMU センサーが内蔵されいる．その 2 種類の IMU センサーから表情変化による顔面筋や顎関節，外耳道の動きの変化を捉えることにより，特徴量を抽出する．本章では，”AirPods”の各センサーから取得可能なデータについて以下で説明する．

3.1.1 加速度センサーによる情報の利用

”AirPods”に搭載された加速度計では，デバイスに対する x，y，z の 3 軸に対する加速度が測定可能である．加速度計から測定された値は，デバイス自身の加速度と重力による加速度との総和となっている．測定値から重力による加速度について，水平方向と垂直方向の 2 成分の重力ベクトルとして取得することが可能である．そして，重力ベクトルを加速度計の測定値から除去することで，デバイス自身の加速度を x，y，z の 3 軸成分として取得することが可能である．

取得した上記 2 種類のデータの単位は，重力加速度 G（$= 9.8\,\mathrm{m/s^2}$）である．このようにして取得した表情変化による 3 軸成分のデバイス自身の加速度と 2 軸成分の重力ベクトルを特徴量として使用する．以下では，デバイス自身の加速度を Acceleration，重力ベクトルを Gravity とする．

3.1.2 ジャイロスコープによる情報の利用

”AirPods”に搭載されたジャイロスコープでは，デバイスに対する x，y，z の 3 軸それぞれを中心にする回転速度が測定可能である．ジャイロスコープから測定された値は，デバイス自身の回転速度を記録したものである．その測定値から地上に対する”AirPods”の向きについて，x，y，z の 3 軸を定義し，それぞれをロール軸，ピッチ軸，ヨー軸として設定する．設定した座標系の各軸に対して回転速度を求めた値をそれぞれロール角，ピッチ角，ヨー角とすることで，床面に対する相対的なデバイスの回転速度を取得することが可能である．

取得した上記 2 種類のデータの単位は，rad/s である．このようにして取得した表情変化による 3 軸成分のデバイス自身の回転速度と床面に対するデバイスの相対的な回転速度を特徴量として使用する．以下では，デバイス自身の

© 1959 Information Processing Society of Japan

回転速度を Rotation，床面に対するデバイスの相対的な回転速度を Roll-Pitch-Yaw とする．

3.2 機械学習モデルの選定

本手法では，1 デバイスの 2 種類のセンサーから取得した複数データを用いて分類するための機械学習モデルを 3 種類作成し，表情認識精度の評価を行った．作成した 3 種類の機械学習モデルは以下である．

3.2.1 1D-Conv-AutoEncoder

一つ目のモデルは，1D-Conv-AutoEncoder モデルである．本モデルでは，特徴量として入力するデータが 1 デバイスの 2 センサーから取得した 4 種類のデータであり，それぞれに 3，あるいは，2 軸の方向成分を持つことから，まずは各データの各軸成分に注目し，複数センサーのマルチモーダル行動認識で用いられたモデル [10] を参考に作成した．図 1a のように，初めに各データの各軸成分に対して 1D Convolutional layer を用いた AutoEncoder を採用することで，それぞれのデータの各軸から特徴量をそれぞれ抽出する．各データの各軸成分から抽出した特徴量を結合し，3 層の全結合層を識別器として適用することによって，より多くの種類の特徴量を考慮した予測結果を取得できるという仮定のもと，学習モデルとして提案した．

3.2.2 2D-Convolution

二つ目のモデルは，2D-Convolution モデルである．本モデルでは, Rotation, Roll-Pitch-Yaw, Acceleration, Gravity の全データをまとめて入力とすることで，同時刻における全てのデータの関連性に注目した．図 1b のように，各データを結合した 2 次元のテンソルを入力とし，2D Convolutional layer を用いることで，データの時間的な繋がりを考慮した特徴量を抽出する．このように抽出した特徴量を用いて，2 層の全結合層を識別器として適用することによって，単一デバイスから取得した時系列データの強みを活かした予測結果を取得できるという仮定のもと，学習モデルとして提案した．

3.2.3 GroupedSensor

三つ目のモデルは，GroupedSensor モデルである．本モデルでは，入力として，Rotation，Roll-Pitch-Yaw，Acceleration，Gravity をそれぞれ分けることで，4 種類それぞれのデータにおける同時刻の各軸成分の関連性に注目した．図 1c のように，入力として，各データごとに 2D Convolutional layer，または，1D Convolutional layer を用いることで，それぞれのデータから特徴量を抽出する．この際，Rotation，Roll-Pitch-Yaw，Acceleration は軸成分による分割をすることなく，それぞれに 2D Convolutional layer を適用したが，Gravity については，x，y 軸の成分ごとに特徴が明確である様子が読み取れたため，各軸成分に分割し，それぞれに対して 1D Convolutional layer を適用した．各データから抽出した特徴量を結合し，3 層の全結合層を識別器として適用することによって，単一デバイスから取得した複数の時系列データを取得している強みを活かした予測結果を取得できるという仮定のもと，学習モデルとして提案した．

4. 実験設定

本研究では，表情認識のデータセットを作成し，そのデータを用いて，表情認識技術の汎用性の検証を行う．そのための実験設定として，データセットの作成方法と評価プロセス，評価指標について以下で説明する．

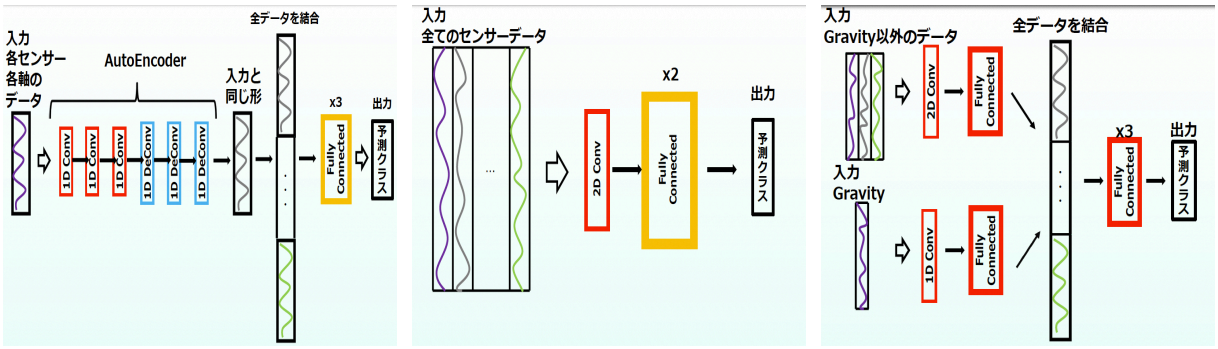
4.1 データセット

本研究では，汎用イヤラブルデバイス”AirPods”での表情認識が実現可能であることを検証するために，まず，被験者が座った状態での表情変化に対して提案手法を評価する．表情の変化としては，我々が日常生活を行なっていく上で，不自然とならないように，以下の図 2 の 5 種類の表情を設定し，それぞれに close，open，left，right，smile のラベルを付与する．

データ収集は次のような手順で行う．表情変化を正確に記録するために，被験者の前のモニターに表情変化の指示を出すスライドを表示し，モニターの下部にアノテーションを行う際に使用する映像記録用カメラを設置する．被験者には，事前に 5 種類の表情をモニターに表示されるスライドの指示に従って行うように手順を説明する．表情変化の指示を出すスライドでは，3 秒ごとに空白のスライドと表情変化の指示のスライドが切り替わるようにし，空白のスライドの時には図 2a のように口を閉じた状態を保ち，表情変化の指示のスライドでは，図 2b から図 2e までの 4 表情を順に指示する．一連のスライドの変化によるデータの記録を 1 セッションとし，図 2b から図 2e の 4 表情の指示が 9 回繰り返されるように設定する．1 セッションの記録が終わるたびに，被験者には”AirPods”を一度外してもらい，データ収集開始時に再び装着してもらうようにする．被験者として，研究室内の学生 8 人に協力してもらい，1 人当たり 5 セッションの表情変化を記録する．

データの収集は，”AirPods”のセンサーに対して，平均 24Hz で表情の変化を記録できるアプリケーションを使用して行う．本研究では，汎用イヤラブルデバイスでの表情認識が可能であることの検討が目的であるため，被験者が音楽を聴いていない状態で表情変化を記録することに集中してもらい，実験を行ったが，データを収集した際の周波数としては，音楽を流している状態でも平均 24Hz で取得可能であることが確かめられたので，イヤラブルデバイスの本来の機能である音声を聞くというものを阻害することなく，表情認識に使用できると考えている．収集したデー

© 1959 Information Processing Society of Japan



(a) 1D-Conv-AutoEncoder **(b)** 2D-Conv **(c)** GroupedSensor

図 1: 選定した 3 モデル

(a) close **(b)** open **(c)** left **(d)** right **(e)** smile

図 2: 予測対象の 5 表情

タに対して，3 秒間のスライディングウィンドウを採用し，スライド幅を 1.5 秒とする．このようにして作成したデータセットでは，各クラスの個数が表 1 のようになる．

表 1: 各クラスごとのセグメント数

close	562
open	157
left	165
right	152
smile	165

4.2 評価プロセス

本提案手法において，汎用イヤラブルデバイスから取得可能なデータごとの重要度の実験と表情の個人依存性についての 2 種類の設定での実験 [11] を行うことで，提案手法の有効性を評価する．実験設定の詳細は以下の 3 つである．

4.2.1 各データの表情認識精度に対する寄与

本研究では，”AirPods”に内蔵されているジャイロスコープと加速度計から取得可能な Rotation，Roll-Pitch-Yaw，Acceleration，Gravity の 4 種類のデータを全て用いて，表情認識することを提案している．それに対して，先行研究 [6] では，Acceleration と Rotation のみを用いて表情認識を行っている．そのため，新たに用いた Roll-Pitch-Yaw と Gravity のデータがどれほど表情認識に寄与しているのかを調査する必要がある．そこで，使用する 4 種類のデータそれぞれを単一の入力として，3.2.2 章で提案した 2D-Convolution モデルを用いて，表情認識の精度の評価を行う．これにより，各データの貢献度がわかるだけでなく，今後の学習モデルの際に重視すべきデータについて検討することも可能となる．

4.2.2 LOSO（Leave-One-Session-Out）

本研究において，複数の被験者の表情データを扱っているため，4.1 章で示した収集対象の 5 表情に対して，被験者ごとの表情筋の発達具合の違いなどから微妙に違いが生じる．そのため，被験者全員に対して，提案手法が適切に表情認識を行えることを確認するため，Leave-One-Session-Out（以下，LOSO）[11] という評価プロセスを採用する．本プロセスでは，収集した各被験者のデータそれぞれ 5 セッションのうち，4 セッションを訓練データにし，残りの 1 セッションをテストデータとして実験を行う．このプロセスにより，表情認識対象者の一部のデータを学習済みのモデルであれば，正しい認識が行えることを示す．

4.2.3 LOPO（Leave-One-Participant-Out）

4.2.2 章の LOSO による評価には欠点があり，特に，表情認識対象者に対して，事前にデータ収集を行う必要があり，時間を要するため，学習させるデータ量を増やすことが難しいという点が挙げられる．したがって，本提案手法の実用化を目指すためには，表情認識対象者のデータを訓練データに含めることなく，実験を行うことが必要である．そのため，被験者個人の表情に依存しない状態でも提案手法が適切に表情認識を行えることを確認するため，Leave-One-Participant-Out（以下，LOPO）[11] という評価プロセスを採用する．本プロセスでは，収集した各被験者のデータのうち，表情認識対象者を一人と定め，その一

© 1959 Information Processing Society of Japan

人の 5 セッションのデータをテストデータとし，残りの被験者のデータを訓練データとして実験を行う．これにより，訓練データを LOSO よりも増やすことを確認するだけでなく，未学習の人物の表情に対しても，表情認識が行えることを示す．

LOSO と LOPO における訓練データとテストデータのサンプル数は表 2 のようになる．

表 2: LOSO と LOPO における訓練データとテストデータのサンプル数

	Train	Test
LOSO	4717	1201
LOPO	5178	740

4.3 評価指標

本研究の評価指標として，Accuracy と Macro-F1 Score の 2 種類の指標を用いる．これらの評価指標は，多クラス分類タスクの性能評価の際に頻繁に用いられる指標である．それぞれの評価指標は以下のような特徴を持つ．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy は，式 1 で表されるようにモデルが正しく認識できた表情の割合を測定できる評価指標である．Accuracy は，分類精度の評価によく用いられる指標であるが，多クラス分類，特に予測対象のセグメント数に偏りがあるデータの分類に対しては，性能を正しく評価できない場合がある．本研究においても，表 1 にあるように，close のセグメント数が他のセグメント数に比べ，3 倍以上になっているため，Accuracy だけでは，正当な評価を行うことができたとは言えない．そのため，Macro-F1 Score を用いた評価も行うことで，提案手法の性能評価の正当性を担保する．

Macro-F1 Score とは，多クラス分類問題に対して，2 クラス分類の精度の評価指標として多く用いられる F1 Score を適応するために定義した評価指標であり，各クラスに対する F1 Score を求め，その値を平均したものである．各クラスに対する F1 Score の求め方は以下のようになる．

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

$$Macro\text{-}F1\ Score = \frac{1}{N} \cdot \sum_{i=1}^{N} F1\ Score_i \quad (N : ClassNum) \quad (4)$$

F1 Score を求めるクラスを Positive，それ以外のクラスを Negative とし，対象クラスに対する 2 値分類と考え，式 2 を用いて Precision と Recall を求める．求めた Precision と Recall を用いて，その調和平均である F1 Score を式 3 を用いて求める．このようにして，各クラスごとに求めた F1 Score を平均したものが Macro-F1 Score であり，式 4 として求めることができる．先に求めた Accuracy に加えて，Macro-F1 Score を用いることで，各クラスが等しい精度で分類されているかを示すことが可能となり，提案手法の評価を正しく行うことができる．

5. 実験結果

5.1 各データの認識精度への寄与

初めに，4.2.1 章の設定の元，各データのうち，どのデータが最も表情認識に寄与しているかを調査した結果が図 3 である．図 3 から，表情認識への寄与が高い順に，Roll-Pitch-Yaw，Rotation，Gravity，Acceleration である．この結果から，”AirPods”の IMU センサーのうち，ジャイロスコープで取得可能なデータの方が加速度計で取得可能なデータよりも表情認識に大きく貢献していると読み取ることができ，表情変化に伴うイヤラブルデバイスの回転変化が重要であることがわかる．したがって，デバイスの回転変化のデータに対するアプローチを詳細に検討することにより精度向上が見込めるのではないかと考える．

加えて，Roll-Pitch-Yaw，Gravity の方が Rotation，Acceleration よりも表情認識に寄与していることから，ジャイロスコープと加速度計のどちらのデータに対しても，イヤラブルデバイス自身の変化よりもイヤラブルデバイスの床面や重力加速度の絶対位置や基準に対する相対的な変化の方が重要であることがわかる．このことは，イヤラブルデバイスの装着位置や角度が被験者によって微妙に異なることに起因するのではないかと考えられ，そのために，固定されているものに対する相対的な変化を重視することで，精度向上の可能性が考えられる．

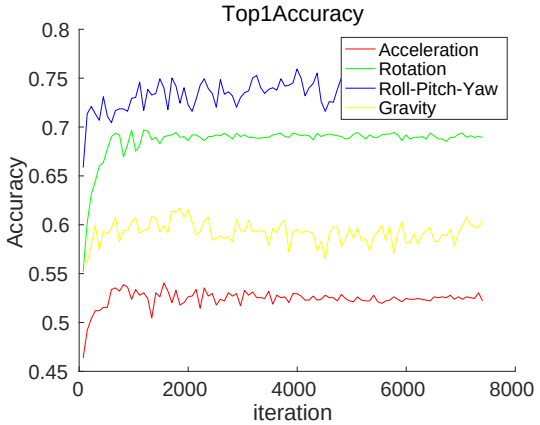


図 3: 取得データ別の表情認識への寄与

5.2 LOSO での各モデルの評価

4.2.2 章の設定の元でデータを分割し，3.2 章で提案した 3 種類の機械学習モデルで表情認識を行った結果が図 4 と

© 1959 Information Processing Society of Japan

表 3a である．表 3a から，Accuracy，Macro-F1 Score のどちらにおいても 2D Convolution モデルが最も精度が良く，79.4%の精度で表情認識を行うことができた．また，Accuracy と Macro-F1 Score を比較したときに，Macro-F1 Score の方が値が小さくなっていることから，表情の種類によって予測精度が不均一であるようにみてとれる．

各表情における精度について，図 4 の混同行列から，図 2 の open から smile までの表情はどのモデルでもほぼ均一な予測結果を出し，close の表情に関連する認識精度が低いことがわかる．close の表情の認識精度が低い原因として，本研究では，前処理として，スライディングウィンドウによる分割を用いているため，時系列データの切り取り箇所によっては，表情変化を行っている途中を一つのセグメントとして扱う可能性があり，close の表情に関連する精度が低くなっていることが考えられる．しかし，実際にハンズフリーデバイス操作をする際には，ウィンドウ幅で切り取ったデータがピッタリと一致している際に，表情認識が正しく行われれば良いと考えられるため，先の考察を仮定すると，実用化の可能性は高いと同時に，セグメントの分け方を工夫する余地があることを示している．

5.3 LOPO での各モデルの評価

5.2 章での LOSO の評価に対して，被験者個人の表情に依存しない状態での評価のために，4.2.3 章の設定の元でデータを分割し，LOSO での評価と同様に 3.2 章で提案した 3 種類の機械学習モデルで表情認識を行った結果が図 5 と表 3b である．表 3b から，LOPO の評価においても，Accuracy，Macro-F1 Score ともに，2D Convolution モデルが最も精度が良く，69.7%の精度で表情認識を行うことができ，Accuracy と macro-F1 Score の比較においても，macro-F1Score の方が値が小さくなっていることから，LOSO と同様に，表情の種類によって予測精度に不均一さが生じることがみてとれる．

また，本評価プロセスにおいて重要となるのは，図 5 の被験者ごとの予測精度の差である．3 つの図の比較から，どの被験者においても，図 5b の 2D Convolution モデルが良いことがわかる．さらに，3 種類すべてのモデルにおいて P4 の精度が他の被験者に比べ著しく精度が落ちているため，表情作りにおける個人依存性についての問題が明白となった．この結果に対して，図 5b の P7 の精度は，79.7%と非常に高く，表情作りが上手い人は，高い精度での表情認識が可能となることが考えられる．以上の個人差を考慮した際に，表情にコマンドを割り当て，ハンズフリーデバイス操作を行う技術の実用化の際には，ユーザーにコマンド操作のための表情を作ることができるように訓練する期間を設けることで，本節で言及した個人差を削減することができると考えている．

5.4 LOSO と LOPO の結果の比較

最後に，汎用イヤラブルデバイスによる表情認識をハンズフリーデバイス操作に実用化する際の大きな障害となり得る表情の個人依存性について改めて，LOSO と LOPO との結果の比較により，言及する．表 3 を参照すると，やはり，LOSO における実験結果の方がどのモデルにおいても精度が高く，訓練データに表情認識対象者のデータを含めて学習させた時のほうが，正確な表情認識を行うことができる．しかし，実際に表情認識を日常のデバイス操作に利用するとなると，ユーザー全員のデータを集めて，学習させることは難しい．したがって，LOSO よりも LOPO での実験の精度について考えることによって，より実用化の可能性が広がる．

ここで，先行研究 [6] の表情認識の結果では，LOSO，LOPO と同様の評価プロセスでの実験の際に，LOPO では LOSO と比べ，45%程度も精度を落としていたことから，結果の比較を行うと，表 3 から，LOPO における評価が LOSO と比較した際に 10%程度の性能低下で留められている．先行研究 [6] とは，表情の種類やデータのサンプル数も異なるため，単純な比較で精度が良いと断言することはできない．しかし，本研究で用いた汎用イヤラブルデバイスは，すでに社会に広く普及しているデバイスであることを考慮すると，ユーザーは本デバイスを装着することに対する抵抗を感じることは少なく，イヤラブルデバイスを用いた表情認識技術によるコマンド操作を実装した際にも，人々から受け入れられる可能性が高いと考えられる．したがって，サンプルデータを十分な数集めることにより，ユーザーの表情を学習することなく，表情認識によるハンズフリーデバイス操作の実用化が可能となるだけでなく，普及の面においても，真に日常生活に浸透する可能性の高い技術となるのではないかと期待できる．

表 3: 各評価プロセスにおける各モデルの予測精度

(a) LOSO

	1D Conv AutoEncoder	2D Conv	Grouped Sensor
Accuracy	0.749	0.794	0.777
Macro-F1 Score	0.707	0.765	0.749

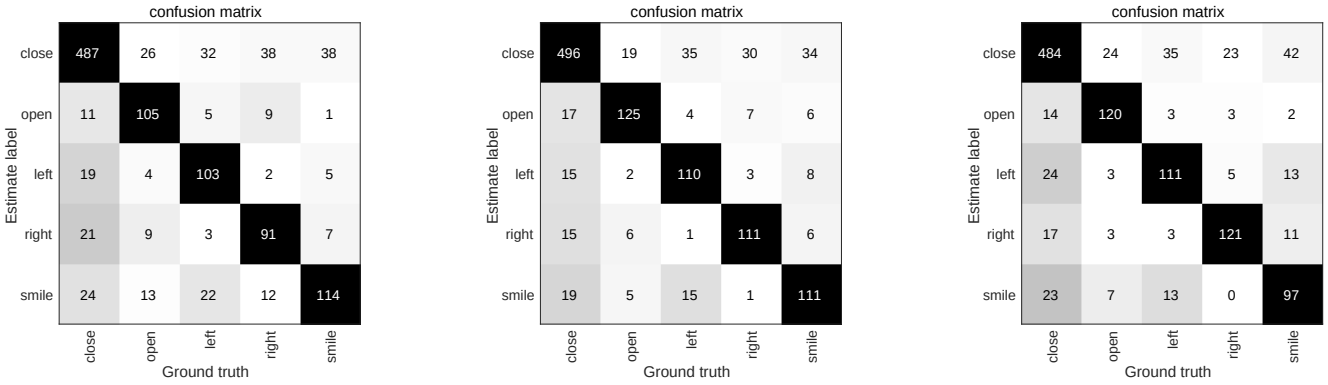
(b) LOPO

	1D Conv AutoEncoder	2D Conv	Grouped Sensor
Accuracy	0.648	0.697	0.686
Macro-F1 Score	0.582	0.652	0.638

6. 結論

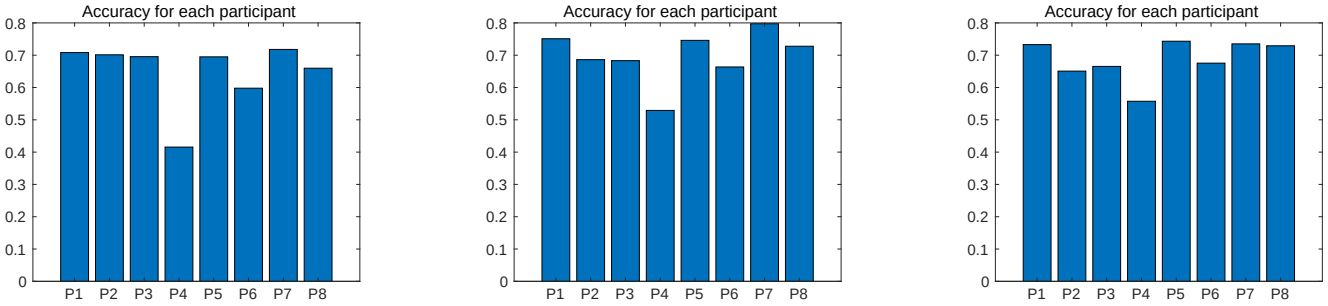
本研究では，汎用イヤラブルデバイスに内蔵された IMU センサーを用いて，表情認識を行う手法の提案，および，

© 1959 Information Processing Society of Japan



(a) 1D AutoEncoder モデル **(b)** 2D Conv モデル **(c)** GroupedSensor モデル

図 4: LOSO における 3 モデルでの各表情の分類精度



(a) 1D AutoEncoder モデル **(b)** 2D Conv モデル **(c)** GroupedSensor モデル

図 5: LOPO における 3 モデルでの各被験者の表情認識精度

表情の被験者依存性を考慮したハンズフリーデバイス操作への応用可能性について検討を行った．実験結果から，5 種類の表情に対して，被験者依存性を考慮しない場合で 79.4%，被験者依存性を考慮した場合で 69.7%の精度で認識を行うことが可能であった．さらに，汎用イヤラブルデバイスから取得可能なデータのうち，デバイスの回転変化の方が加速度変化よりも表情認識への寄与が大きく，また，デバイス自身の回転や加速度の変化よりも床面や重力加速度方向に対する相対的な変化を算出したデータの方が寄与が大きいことを示した．

また，本実験に用いたデータセットでは，8 人の被験者を対象にデータ収集を行なったが，被験者を増やし，様々なサンプルを学習させることで，更なる性能向上が見込めるとともに，表情の種類も増やすことによって，ハンズフリーデバイス操作の際の応用範囲の拡大の可能性を持つ．加えて，本研究で用いたデバイスは，すでに社会に普及していることを考慮すると，表情変化によるハンズフリーデバイス操作を導入した際の社会的受容性が高いことが期待できる．したがって，本研究の今後の課題として，データの拡充とともに，表情の種類を増やし，より実用的なハンズフリーデバイス操作の実現が考えられる．

参考文献

[1] Chen, T. et al.: NeckFace: Continuously Tracking Full Facial Expressions on Neck-Mounted Wearables, *Proc. of IMWUT* (2021).
[2] Chen, T. et al.: C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-Mounted Miniature Cameras, *Proc. of UIST* (2020).
[3] Hsieh, P.-L. et al.: Unconstrained realtime facial performance capture, *IEEE CVPR* (2015).
[4] Li, K. et al.: EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements, *Proc. of IMWUT* (2022).
[5] Choi, S. et al.: Excerpt of PPGface: Like What You Are Watching? Earphones Can “Feel” Your Facial Expressions, *Adjunct Proc. of UbiComp/ISWC* (2023).
[6] Verma, D. et al.: ExpressEar: Sensing Fine-Grained Facial Expressions with Earables, *Proc. of IMWUT* (2021).
[7] Gashi, S. et al.: Hierarchical Classification and Transfer Learning to Recognize Head Gestures and Facial Expressions Using Earbuds, *Proc. of ICMI* (2021).
[8] Amesaka, T. et al.: Facial Expression Recognition Using Ear Canal Transfer Function, *Proc. of ISWC* (2019).
[9] Kawsar, F. and othres: Earables for Personal-Scale Behavior Analytics, *IEEE PerCom* (2018).
[10] Strömbäck, D. et al.: MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices, *Proc. of IMWUT* (2020).
[11] Bhattacharya, S. et al.: Leveraging Sound and Wrist Motion to Detect Activities of Daily Living with Commodity Smartwatches, *Proc. of IMWUT* (2022).

© 1959 Information Processing Society of Japan