Vol.2022-DPS-190 No.48
Vol.2022-CSEC-96 No.48
2022/3/11

身体的特徴に依存しない音声を用いた個人識別手法

田中 勇護[1,a)]　清水 一樹[1,b)]　渡邊 健之[1,c)]　川尻 秀憲[1,d)]　高松 晴仁[1,e)]　高橋 知暉[1,f)]
宇田 隆哉[1,g)]

概要：暗証番号による個人認証や指紋認証，顔認証などの個人認証方式が存在するが，障がい者の利用面を考えると入力時の負担が大きい．そこで本研究では, それらに代わる音声の音波における波形を用いた個人識別手法を提案する．本手法では, 録音した音声の波形を画像化し，CNN による機械学習で分類するものである．本研究の評価実験では，CNN で訓練したモデルを用いてテストデータを分類した. 実験結果から，音声の音波における波形を用いた個人識別方式は非常に高い精度であることが明らかになった．しかし，課題として雑音環境下における可用性の向上と少数のデータによる機械学習精度の向上が挙げられる．さらに，本人確認時に大きな声量を出すため，一般環境下における音声入力手法の改善が必要である.

Personal Classification Method by Voice Independent of Biological Characteristics

1. はじめに

現代では個人認証に暗証番号だけではなく，バイオメトリクス認証を利用している．

その中で，指紋認証や顔認証が世の中で多く利用されている．ただ，指紋認証や顔認証はそれぞれ専用の機材を使用しないと認証機能を十分に働かせることができないと考えられる．加えて暗証番号での個人認証では，ブルートフォースアタックによる不正アクセスの可能性，暗証番号を覚えなければならないことや，障害者からすると入力が難しく疲れる [1] という意見もある．しかし，音声を用いた認証方式はあまり実装されておらず，論文や研究報告書には多くは語られていない．

そこで我々は，声を使った個人識別方式を提案する．声であれば障害者を含めて，手や目を使わずに個人識別することができると考えている．また，複雑なパスワードのような入力の難しい方式から脱却できる可能性があると考えられる．

なお，身体的特徴の声紋を用いると，取得した情報が漏えいした場合に声紋を変更できず問題となるため，本論文では行動的特徴となる声の出し方を用いて個人を識別する．これにより，認証に用いられるような絶対的な精度を失う代わりに，情報漏えいの際にも身体的特徴となる情報が守られる．

[1] 東京工科大学
東京都八王子市片倉町 1404-1, 192–0982
[a)] c011919697@edu.teu.ac.jp
[b)] c0119157e8@edu.teu.ac.jp
[c)] c011934396@edu.teu.ac.jp
[d)] c0119088f7@edu.teu.ac.jp
[e)] c011918842@edu.teu.ac.jp
[f)] c01181674c@edu.teu.ac.jp
[g)] uda@stf.teu.ac.jp

2. 関連研究

2.1 多重バイオメトリクスによる個人認証

坂野らは，顔と声紋を用いた個人認証実験で，識別的手法を導入し，その有効性を検証した．手法自体は簡便であるものの，セキュリティレベルの設定が困難であることや，最適な識別器械・技術の選択が困難であることを明らかにした [2]．技術選択の必要があり，応用場面の状況に応じて手法を変更していく必要があると述べている．

2.2 視覚障害者の大学生における情報セキュリティ疲れの考察

垣野内らは，情報セキュリティ疲れに陥ることに加え，それに対する施策が視覚障害者には対応できない先行研究が実施されていることに注目し，視覚障害者に限定した情報セキュリティ疲れの調査を実施した．結果，晴眼者と視覚障害者ではコンディションマトリクス上の分布が大きく

異なり，セキュリティ対策実施度が高い学生ほどセキュリティ疲労度も高いことが判明している [1]．この論文にもあるように，視覚障害者への個人認証を取り上げたものは少なく，その対応もできていないことが分かっている．

2.3 覗き見耐性を持つマウス操作を用いた個人認証方式の提案

長友らは，マウス操作を用いた認証方式について検討した．マウスを動かし，クリックをする動作を個人認証に用いている [3]．この手法では，マウスを隠して個人特定を行うことが前提であるので，マウスを隠せない環境では覗き見耐性が激減する問題がある．またクリック音で個人を推測することが可能である．

大声での認証な声自体が公なものであることから，覗き見や再現が効かない．

2.4 唇動作と音声を用いたカーネル判別分析による個人認証方式

市野らは，唇動作と音声の認証器から出力されるスコアを統合して認識するアルゴリズムにカーネル判別分析を用いた個人認証方式を提案し，その有効性を確認している [4]．この手法では，唇動作と音声の認証器を必要とする．また，録音録画の環境が整っているオープンデータベースを用いた研究内容であり，汎用性に乏しい．

2.5 話者認識技術の実用化に向けて

松井らは，任意の言葉を発生するテキスト独立型の話者認識方法を用いて，声による個人認証システムの構築を試みた。

男性 19 名が約 1 か月にわたる 2 時期に，4 種類の有線電話機を通じて，オフィス環境で発生した音声を使用して実験を行った．話者 19 名のうち 10 名は学習と認識で同じ電話機を使用し，その他の 9 名は異なる電話機を使用した．発声内容は名前，生年月日，電話番号，住所，出身地などの個人情報や 4 種類の一般的な単語（銀行名など）で各話者によって異なる．

この結果から学習と認識で発声練習や電話機が同じ/異なる条件は，認識性能に大きな影響を与えること，劣化要因としては，発生内容，電話機の順に大きいことがわかった．課題としては学習と認識の磁気さや話者に判定のためのしきい値の設定などがある．特に電話網を利用する場合には，少ない情報量や回線やハンドセットの違いによる歪みの問題があった [5].

このことから，発声内容は同一であった方が認証における本人確認の精度は有利であるといえる．

2.6 スマートウォッチの竜頭型コントローラを用いた暗証番号入力方法

稲村らはスマートウォッチを利用した，個人認証方法を提案した [6]．この手法では，スマートウォッチのリュウズを使用し，個人認証を行う．20 歳から 23 歳の男女 20 人を被験者とし，テンキーとリュウズを用いた本人確認方法で覗き見耐性の実験を行った．結果，テンキーでは覗き見攻撃が 100 パーセントであったが、リュウズだと 10 パーセントに減った．

しかし，個人識別を行う際に，専用の機械が必要のため利便性が落ちる．

3. 提案手法

この問題の解決法は，録音した声を波形で表したものを画像化し，CNN を用いた機械学習で分類するものである．

録音に際しては，声を 5 秒間録音する．最初の声量は 0 から，録音時間の中間である 2.5 秒で最大の音量になるように徐々に声を大きくし，2.5 秒を超えたら徐々に声を小さくしていく．最終的に 5 秒で再び声量が 0 になるようにする．この時発声する言葉は「あー」である．

録音，機械学習ともに使用したノートパソコンは dynabook GX83/JLE，プログラミング言語は Python3，実行環境は Visual Studio Code である．

Python で録音する環境の構築に, PySimpleGUI モジュール，pyaudio モジュールをインストールする必要がある．PySimpleGUI モジュールは，Python のコマンドプロンプトを管理者権限で実行して pip を使ってインストールすればよい．pyaudio モジュールをインストールするには，ダウンロードページ `https://www.lfd.uci.edu/~gohlke/pythonlibs/#pyaudio` から自分の Python の環境のバージョンとアーキテクチャーにあったものをダウンロード後，「python -m pip install –upgrade pip setuptools」をコマンドで実行する．次にダウンロードしたファイルがあるディレクトリーから「pip install (ダウンロードしたファイル名)」をコマンドで実行すればインストールが完了する．

またライブラリ numpy，matplotlib もインストールする必要がある．

プログラムを実行することで，図 1 に示すポップアップ (GUI) が表示される．「録音」ボタンを押すと，白い四角い枠の中に「rec」次の行に「3 秒後に始めます」と表示され，3 秒後に録音が開始する（この時「Now Recording...」と白枠に表示される．）．

先程提示した録音方法で録音が終了し，録音から 5 秒経過すると，白い四角白い枠に「Finished Recording.」と表示される．録音が終了するとプログラムファイルと同一のディレクトリに「voicewav」と「voiceflg」ディレクトリが生成される．音声データは「voicewav」に wav 形式で保存

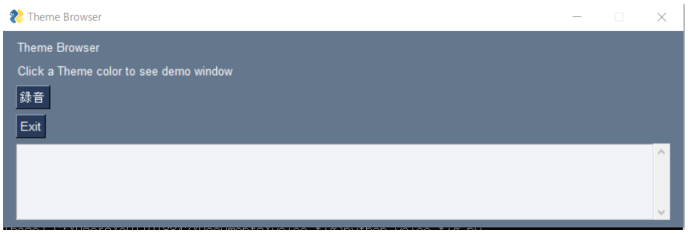

図 1: 録音用ブラウザ
Fig. 1 Recording browser

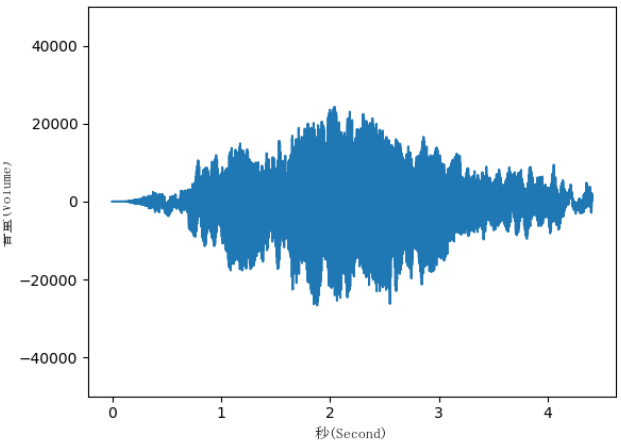

図 2: 波形グラフ
Fig. 2 Waveform graph

される．matplotlib で生成された音声データの波形グラフは，「voiceflg」に png 形式で保存される．この波形データを図 2 に示す．グラフのプロットに当たっては，グラフの目盛りは可変できないように設定してある．CNN を用いて機械学習を行うのは，このうち波形グラフのみである．

被験者 10 名にそれぞれ 100 回ずつ録音を実施した．回数を重ねることによる疲労や声の変化に対応するため，20 回録音で必ず休憩を挟み，水分補給や飴による喉のケアは随時実施できる状態とした．

録音後，CNN を用いた機械学習を行うため，被験者ごとにディレクトリを用意した．

具体的な波形グラフについては 5 章にて示すが，この波形グラフは CNN に入力する際の効率を考慮し，小さいものにしている．これにより，高周波の波形はつぶれ，声紋としての特徴は残っていない．サーバに保存されるのはこの波形グラフの画像である．

4. 評価

4.1 評価方法

本研究の評価としては，CNN を用いてトレーニングを行った後テストデータを分類させたのち，これらの結果より，評価を算出し，精度を比較することで行った．10 名から集めたデータにおいてはそれぞれ，トレーニング用，テスト用の割合は 9:1，エポック数は 100 回である．これらは同時期に取得した音声であるが，テスト用データの情報がトレーニング時のパラメータに影響しないこと，テスト用データはトレーニング用データとは独立して録音されていることから，検証用データではなくテスト用データとして扱った．

入力層の出力ユニット数は 16*16px の 16 枚である．1 つ目のプーリング層は 2*2px，1 つ目の二次元畳み込み層の出力ユニット数は 3*3px の 128 枚，2 つ目の二次元畳み込み層の出力ユニット数は 3*3px の 256 枚，2 つ目のプーリング層は 2*2px，全結合層の出力ユニット数は 128 枚とした．そして出力ユニットの 25 パーセントを無効化した上で softmax 関数を出力時に用いた．softmax とは総和が 1 となるように確率を出していく関数である．また，EarlyStopping の条件として，予測データの正解率が 5 回以上変わらなければトレーニングを終えるものとした．

表 1: 評価結果
Table 1 Evaluation results.

CNN 精度	96.7%
CNN 標準偏差	0.0167508

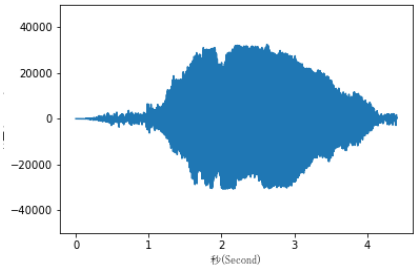

図 3: 1 人目の被験者の 10 回目の波形
Fig. 3 The 10th time waveform for the first person

4.2 結果

CNN を使用した際の分類結果を表 1 に示す．結果として精度は 96.7%であるということがわかる．また標準偏差は 0.0167508 であることがわかった．

5. 考察

本研究で評価した結果から考えられる原因を，精度が高かった原因と精度が低かった原因に分けて考察する．精度が高かった原因として考えられるのは，音量の減少にあたって特徴が出ていたためだと考えられる．

精度が高かった原因として考えられるのは，音量の減少にあたって特徴が出ていたためだと考えられる．1 人目の被験者のデータのうち，10 回目に録音した波形を図 3，70 回目に録音した波形の方を図 4 に示す．また，2 人目の被

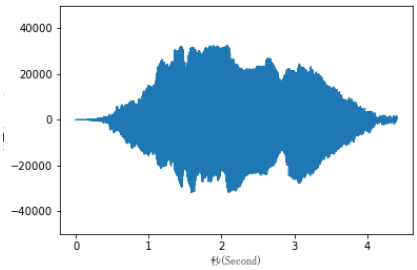

図 4: 1 人目の被験者の 70 回目の波形
Fig. 4 The 70th time waveform for the first person

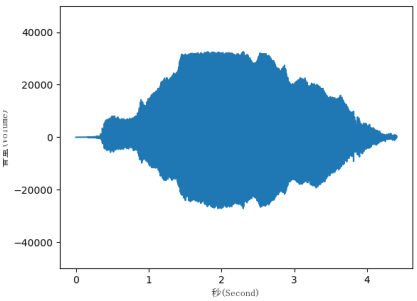

図 5: 2 人目の被験者の 10 回目の波形
Fig. 5 The 10th time waveform for the second person

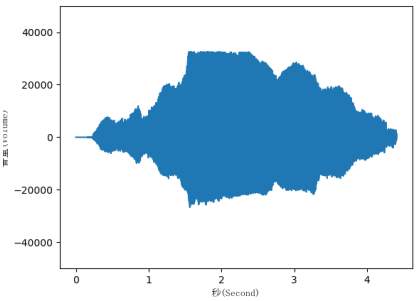

図 6: 2 人目の被験者の 70 回目の波形
Fig. 6 The 70th time waveform for the second person

験者のデータのうち，10 回目に録音した波形を図 5，70 回目に録音した波形の方を図 6 に示す．1 人目の被験者では，後半の波形の減少傾向が似ており，2 人目の被験者も同様に後半の波形の減少傾向が似ている．しかし 1 人目の被験者と 2 人目の被験者の後半の波形を比較するとそれぞれで波形の減少傾向が異なっている．被験者の声量がデータを取得していくうえで徐々に減少しなかったのは，こまめに休憩をはさみながらデータを取得していたためであると考えられる．

精度が低かった原因として考えられるのは，本実験で使用したマイクの精度の低さによって起こりうる音割れである．音割れによって本来取得できる音量が取れない可能性がある．被験者によっては録音のタイミングにずれが生じているため，精度が低くなったと考えられる．また，録音する際に，周囲の雑音などが入ってしまうと，状況によって取得できる音量は異なるため，精度が低くなったと考えられる．

6. 結論

本研究では，特徴のある大声を機械学習でトレーニングとテストを行い，その精度を評価した．評価の結果，平均で 96.7%という非常に高い精度が得られた．この結果は，後半，声を下げるときの特徴が強く出ていたからと考えられる．また，精度が 100%ではなく，約 3%適切に振りけることができなかったものがあるが，これについては，ズレの生じや雑音による影響であると考えられる．

問題点として，この個人特定方法は，論文執筆時点の社会情勢に合っていないところがある．また，実用化を目指す場合，テストは 1 回でよいが，トレーニング時に 100 回録音するのは現実的とはいえない．どのようにウイルス感染対策を行えばいいのか，少ないデータ数でも同じ精度が出るようなデータの水増し，調整や機械学習の手法を模索していかなければならないと考える．今後は，このような問題点の解消方法，録音攻撃や推測攻撃などの脅威への耐性についても検討を行っていきたい．

参考文献

[1] 垣野内将貴，面和成：視覚障害者の大学生における情報セキュリティ疲れの考察，情報処理学会研究報告，Vol.2021-SPT-45, No.15, pp.1–8, (2021).
[2] 坂野鋭，劉偉傑：多重バイオメトリックスによる個人認証，情報処理学会研究報告コンピュータセキュリティ（CSEC），Vol.1999, No.45, pp.37–42, (1999).
[3] 長友誠，朴美娘，岡崎直宣：覗き見耐性を持つマウス操作を用いた個人認証方式の提案，情報処理学会研究報告，Vol.2017-CSEC-78, No.29, pp.1–8, (2017).
[4] 市野将嗣，坂野鋭，小松尚久：唇動作と音声を用いたカーネル判別分析による個人認証方式，電子情報通信学会論文誌，Vol.92, No.8, pp.1363–1372, (2009).
[5] 松井知子，吉岡理，南泰浩：話者認識技術の実用化に向けて，映像情報メディア学会技術報告，Vol.22.45, pp.43–48, (1998).
[6] 稲村勝樹，市村泰佑：スマートウォッチの竜頭型コントローラを用いた暗証番号入力方法，情報処理学会研究報告，Vol.2019-IOT-44, No.38, pp.1–6, (2019).