ポップノイズに含まれる音素情報を用いた声の生体検知と話者照合システムの統合*

◎望月紫穂野, 塩田さやか, △貴家仁志 (首都大)

1 はじめに

近年，声を用いた生体認証システムである話者照合の精度向上に伴い実用性が高まってきている．同時に，登録話者の声を録音再生するなりすまし攻撃や音声合成・声質変換といった声を作る技術を用いて登録話者を模倣するなりすまし攻撃によって精度が大幅に低下してしまうことが報告されている [1]．これらのなりすまし攻撃に対処するために提案されてきた手法は，音響的特徴量として様々な特徴量を用いるものが主であった．一方，話者照合システム内でのモデル学習や特徴抽出による対策ではなく，なりすまし攻撃に対する根本的な解決策として入力音声が実際に人間が発声したのか否かを判定する声の生体検知という枠組みが提案された [2]．これまでに声の生体検知を実現するためにポップノイズ検出法が提案されたが，ポップノイズ検出による声の生体検知では，なりすまし攻撃にはポップノイズが生じていないことを前提としていた．そのため再生音声に攻撃者が息を吹きかける等により恣意的にポップノイズを生じさせた場合，再生音声を生体として誤受理してしまう可能性が高い．そこで声の生体検知の頑健性向上のため，ポップノイズ区間内に含まれる音素情報を用いた声の生体検知を提案し，高いなりすまし検出精度が得られることを報告した [3]．本研究では実際にポップノイズに含まれる音素情報を用いた声の生体検知と話者照合システムを統合し，話者照合実験を行うことでなりすまし攻撃に対する頑健性が向上することを報告する．

2 話者照合のための声の生体検知

2.1 ポップノイズ情報を用いた声の生体検知

話者照合に登録話者の声を録音した音声や合成音声などをスピーカーで再生して入力音声とするなりすまし攻撃が問題となってきている．そこでなりすまし攻撃に対する根本的な解決策として，声の生体検知という入力音声がスピーカーで再生されたものなのか人間が実際に話したものなのかを識別する枠組みが提案された．図 1 に示すように，話者照合の前段として使用することで，なりすまし攻撃に対する話者照合システムの頑健性を向上させることを目的としている．声の生体検知の実現手法として入力音声にポップノイズが含まれているかを検出する方法が有用であることが報告されている．ここでポップノイズとは人間がマイクに向かって発声する際にマイク内部に息や風が入りこむことにより発生してしまうノイズのことを指す [4]．

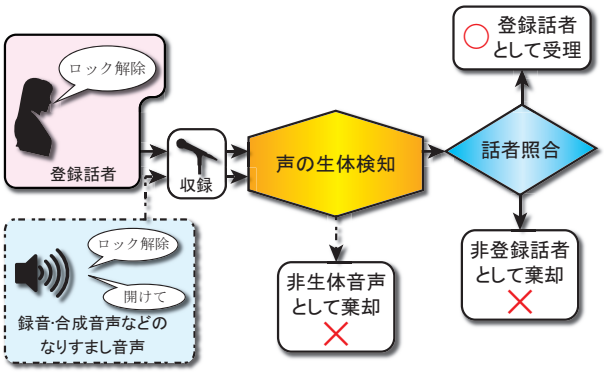

Fig. 1 話者照合システムと声の生体検知の概要

2.2 ポップノイズ検出法

ポップノイズの有無を検出するために，シングルポップノイズ検出法 [2] を用いた．ポップノイズが低周波成分に突発的な強いエネルギーを持つ性質があるため，シングルポップノイズ検出法では，そのエネルギー変動を捉えることで検出を行う．具体的な手順としてはまず，各フレームごとに短時間フーリエ変換を行い，周波数分解を行う．次に振幅スペクトルの低周波領域のみの値の平均を求める．この値が各フレームにおける低周波成分のエネルギーを表すため，フレーム間でのエネルギー変動が極大になる点をポップノイズとして検出する．シングルポップノイズ検出法はマイク 1 つで実現可能であり，導入コストが低く，また話者照合システムとの親和性も高いことが利点としてあげられる．

3 ポップノイズに含まれる音素情報を用いた声の生体検知

3.1 ポップノイズ検出と風に対する脆弱性

ポップノイズの有無による声の生体検知では，なりすまし音声の中には恣意的にも偶発的にもポップノイ

*System integration between voice liveness detection based on pop-noise detector considering phoneme information and automatic speaker verification. by MOCHIZUKI, Shihono, SHIOTA, Sayaka and KIYA, Hitoshi(Tokyo Metropolitan University)

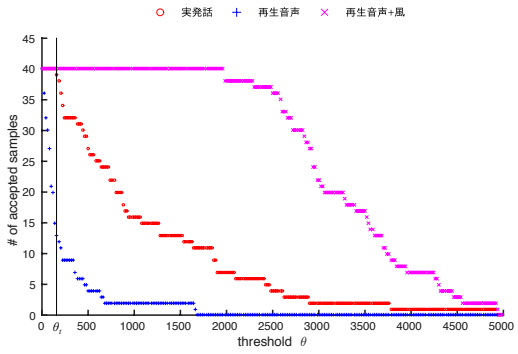

Fig. 2　テストデータ毎（実発話，再生音声，再生音声+風）のポップノイズ検出の閾値 θ と生体受理数の変化

ズは生じないことを前提としており，攻撃者がなりすまし音声を再生しながら自分の呼気や風を使ってポップノイズを発生させた場合に，なりすまし音声でも生体音声として誤受理されてしまうという可能性がある．実際に，再生音声に風を恣意的に発生させた場合の変化を図 2 に示す．図 2 は各テストデータ（実発話，再生音声，再生音声+風）に対してポップノイズ検出の閾値 θ を変化させた時に生体として受理されたサンプル数を示している．ここで，実発話とは人間が実際に発話した音声を収録した音声，再生音声とは収録した実発話を再生したものをなりすまし音声として収録した音声，また，再生音声+風とはなりすまし音声をスピーカーで再生しながら同時に団扇で風を起こして収録した音声のことを表す．各データの変化の推移を見ると，再生音声の生体受理数の減少が実発話に比べて早いことから，再生音声は実発話よりもポップノイズが含まれていないことがわかる．このことからポップノイズ検出の閾値を調整することで，ポップノイズの有無による再生音声の棄却は容易であると考えられる．一方で，再生音声+風の生体受理数の減少は実発話よりも遅い．このことから故意に発生させた風がポップノイズとして検知されていることがわかる．以上よりポップノイズの有無のみによる声の生体検知では，再生音声を棄却することは可能であるものの，再生音声+風を棄却することは難しいことがわかる．ポップノイズが含まれる再生音声を棄却するために，生体として受理された音声のポップノイズ区間内の音素の出現傾向を考慮する必要がある．ここで，人が発話したときにポップノイズを発生させやすい（Easily caused Pop-noise ; EPN）音素と発生させにくい（Hardly caused Pop-noise ; HPN）音素には傾向があることがわかっている．この傾向を用いて，ポップノイズ区間内に EPN 音素および HPN 音素を含むかどうかで生体検知をする方法を提案する．

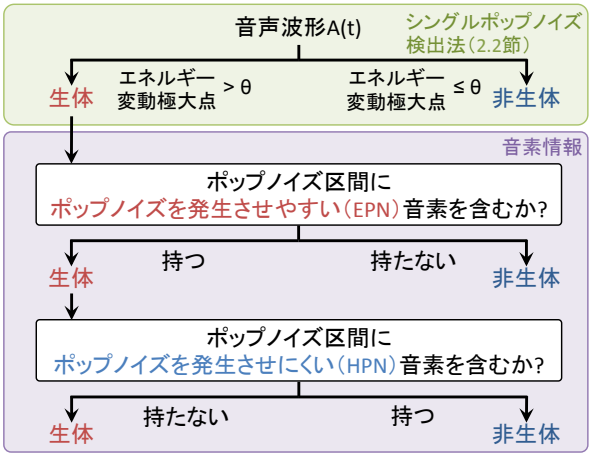

Fig. 3　ポップノイズに含まれる音素情報を用いた声の生体検知のフロー

3.2　ポップノイズに含まれる音素の抽出

EPN 音素および HPN 音素を用いた声の生体検知を行うために，VLD データベース [2] を用いてポップノイズ区間に含まれる音素の傾向を調査した．VLD データベースでは，入力音声は 2 本の同じ種類マイクを用意し，片方には風防カバーを装着，もう片方には風防カバーを装着しないで収録したステレオ収録の音声データを使用することを想定している風防カバーありの音声データに関してはポップノイズがほぼ発生しておらず，風防カバーなしの音声データに関してはポップノイズが発生している状態を想定している．また，2 チャンネルのデータは同時に収録を行うため時間のずれは生じない．ポップノイズが発生した区間に含まれる音素情報を以下の手順で抽出した．

1：汎用大語彙連続音声認識エンジン Julius を用いて風防カバーありのマイクにより収録した音声データに対して音声認識を行い，モノフォンの音素アライメントを取得．

2：風防カバーなしの音声データを用いてシングルポップノイズ検出法を用い，ポップノイズ区間のアライメントを取得．

3：手順 1，2 で得られたアライメント情報を用いて，ポップノイズ区間に含まれる音素を抽出．

3.3　ポップノイズに含まれる音素情報による判定

ポップノイズに含まれる音素情報を用いた声の生体検知について説明する．フローを図 3 に示す．はじめにシングルポップノイズ検出法を用いて入力音声のポップノイズを検出する．入力音声にポップノイズが含まれるならばその音声を生体による音声として受理する．含まないならば非生体による音声として棄却する．次にシングルポップノイズ検出法にて生体として受理された音声に対し，ポップノイズが生じた再

ポップノイズ検出	
周波数帯域	10 Hz
周波数分解能	5 Hz
分析窓幅	200 msec
窓シフト幅	25 msec
閾値 θ	実発話を 100%受理する値（図 2 の黒線の値 θ_t）
話者照合	
サンプリング周波数	16 kHz
量子化ビット数	16 bit
特定話者モデル学習データ	平均 45 文章 × 4 名
UBM データベース	JNAS（男性のみ）
UBM 学習データ	14951 文章
分析窓幅	25 msec
窓シフト幅	10 msec
特徴量	MFCC19 次+Δ+ΔΔ

Table 1　ポップノイズ検出および話者照合における実験条件

生音声を棄却するためにポップノイズ区間内に EPN 音素を含むかどうかで生体検知を行う．3.2 節で述べた手順により，もしポップノイズ区間に EPN 音素を含むならば，それは人による発話によって発生したポップノイズと想定されるため生体として受理する．逆に含まないならば，それはなりすまし攻撃と想定されるため非生体として棄却する．しかしながら，図 2 に示した通り，再生音声+風のサンプルにも実発話同様ポップノイズを含んでおり，EPN 音素部分にポップノイズ区間が生じた再生音声が入力された場合，誤受理してしまう．そういった再生音声を棄却するために，EPN 音素情報で生体として受理された音声のポップノイズ区間内に，HPN 音素を含むかどうかで更に生体検知する．もし HPN 音素を含むならば，そのポップノイズは人による発話と考えにくいため非生体として棄却する．逆に含まないならば生体による音声として受理する．

3.4　声の生体検知と話者照合システムの統合

なりすまし攻撃に対する話者照合の頑健性を向上させるために，声の生体検知と話者照合システムを統合する．様々な統合手法が考えられるが本研究では最もシンプルな統合を行う．図 1 に示すように，声の生体検知部で入力された音声信号が生体から発せられたものか否かを識別し，生体であると検知できれば後段の話者照合に信号を渡すようになっている．なりすまし攻撃を声の生体検知部分で棄却することで，なりすまし攻撃に対する話者照合の頑健性を維持できると考えられる．また，声の生体検知のを破るための工作によって，本来の目的である話者照合に対しての精度が下がることも期待できる．

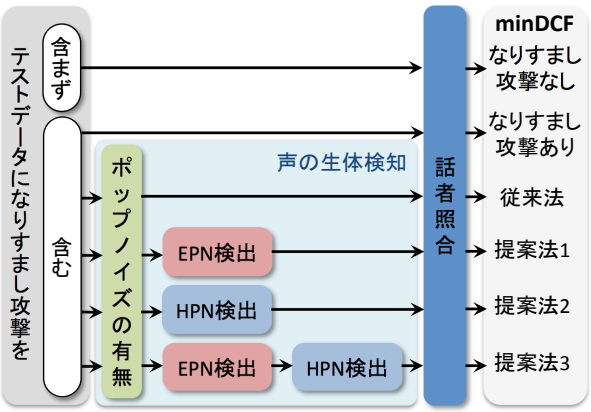

Fig. 4　実験フロー

4　評価実験

4.1　実験条件

ポップノイズに含まれる音素情報を用いた声の生体検知の性能を評価するために話者照合実験を行った．評価のために，人が実際に発話した音声を収録した実発話と，なりすまし攻撃用に再生音声および風をのせた再生音声+風を収録した．音声収録は静かな部屋で行った．それぞれの収録方法は以下の通りである．収録には 2 本のマイク（AKG P170）を用い，1 本は風防カバーを装着し，1 本は風防カバーを装着しない 2 チャンネル同時収録を行った．マイクの音量は各話者毎に調節し，再生音声は今回収録した実発話をスピーカー（ELECOM LBT-SPP300）で再生したものを収録した．また再生音声+風は，マイクに向かって団扇で無作為に風を起こしながら今回収録した実発話を再生したものを収録した．このときマイクと口およびスピーカーの距離は約 5cm とした．収録したテストデータは男性 4 名，文章数は各話者につき実発話 10 文/再生音声 10 文/再生音声+風 10 文の計 120 文である．サンプリング周波数 48kHz，量子化ビット数 24bit とした．ポップノイズ検出および話者照合における実験条件は表 1 に示す．ポップノイズ検出には風防カバーを装着していないマイクで収録した音声を使用した．話者照合には入力音声には風防カバーを装着しているマイクで収録した音声を入力した．また声の生体検知に用いる EPN 音素および HPN 音素は，3.1 節の予備実験により得られた傾向からポップノイズ区間に含まれやすかった上位 10 個の音素を EPN 音素 (hy，f，o:，ch，p，q，t，s，n，h) とし，ほとんどポップノイズ区間に含まれなかった音素を HPN 音素 (u:，ry，ny，i:，m) とした．評価尺度には NIST の最小決定コスト関数（minDCF）を使用した [5]．minDCF は式（1）の通りである．

$$\begin{aligned}\text{minDCF} = {} & P_{target} \times P_{Miss|Target}(\theta) \\ & + (1 - P_{target}) \times P_{FalseAlarm|NonTarget}(\theta)\end{aligned} \quad (1)$$

ただし，$P_{target} = 0.01$ とし，$P_{Miss|Target}(\theta)$ および $P_{FalseAlarm|NonTarget}(\theta)$ はそれぞれ閾値 θ に対する本人棄却率と他人受入率とする．

図 4 に実験フローを示す。まずなりすまし攻撃を含まないテストデータおよび含むテストデータに対し，声の生体検知をせずに話者照合しそれぞれ minDCF を算出する．なりすまし攻撃を含んだときの minDCF をベースラインとし，話者照合前に声の生体検知を行うことでなりすまし攻撃に対する話者照合システムの頑健性を改善できるか確認する．次の 4 通りの声の生体検知と話者照合を統合し，それぞれ minDCF を算出した．声の生体検知手法は次の 4 手法になる．：ポップノイズの有無のみ（従来法），EPN 音素情報のみ（提案法 1），HPN 音素情報のみ（提案法 2），EPN 音素と HPN 音素の両情報（提案法 3）．

4.2 実験結果

図 5 に各手法での minDCF の値を示す．はじめに，なりすまし攻撃なしの minDCF となりすまし攻撃ありの minDCF を比較すると，後者の方が 0.0625 高くなる．minDCF の値は話者照合システムの誤受理に対して特に大きく変動する評価尺度であるため，なりすまし攻撃に用いられた再生音声の多くが話者照合システムのみでは誤受理されてしまったことがわかる．次に声の生体検知と話者照合を統合したときの minDCF について考察する．なりすまし攻撃なしの minDCF とポップノイズの有無だけで判定する従来法の minDCF を比較すると，従来法の方が低い値を取るものの大きな変化はない．これは，主に再生音声+風のサンプルが風由来のポップノイズを持つためにポップノイズの有無だけでは棄却することができなかったからである．次に提案法 1 の minDCF に着目すると，従来法に比べ大幅に値が低下する．これは風によるポップノイズの発生区間が EPN 音素でない音素にかかっており，なりすまし音声だと正しく判定されたためである．このことから EPN 音素情報によって，話者照合システムに入力するなりすまし攻撃の数を減らせたことを意味する．しかしながら，なりすまし音声のポップノイズ区間が EPN 音素にかぶってしまい，棄却できないケースもあった．提案法 2 の minDCF は提案法 1 に比べると minDCF の改善は少ないものの，従来法よりも低い．これは，3.3 節でも述べたように HPN 音素は再生音声+風のようなケースを想定し，過剰にポップノイズが発生する場合に声の生体検知の頑健性向上を目指して取り入れたものであるため，EPN 音素と組み合わせないと性能が出ないためである．最後に提案法 3 の minDCF に着目すると，他手法の中で最も低い minDCF が得られた．EPN 音素情報だけでは棄却できなかった，EPN

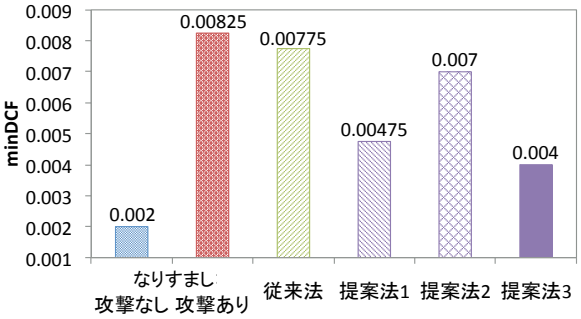

Fig. 5　各手法毎の minDCF

音素部分にポップノイズ区間がかぶった再生音声+風のサンプルを，HPN 音素情報によって正しく棄却したためである．提案法 2 から HPN 音素のみの検出は効果が低いものの，EPN 音素情報と組み合わせることでより頑健な声の生体検知となることが示された．以上より，音素情報を用いた声の生体検知を行うことで話者照合システムがなりすまし攻撃に頑健になるといえる．しかしながら，今回用いたテストデータに EPN 音素あるいは HPN 音素が元々含まれない文章もあった．そのため事前に音素情報を考慮したプロンプト文を用いることで更なる改善が見込まれる．

5 おわりに

本稿では，ポップノイズに含まれる音素情報を用いた声の生体検知と話者照合システムの統合について提案し，実験によりその有効性を示した．今後の課題としてサンプル数の増加や音素バランスを考慮したプロンプト文の使用などが挙げられる．

参考文献

[1] N. Evans, *et.al.*, “Spoofing and countermeasures for automatic speaker verification,” in Proc. Interspeech, pp. 925-929, 2013.

[2] S. Shiota, *et.al.*, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” in Proc. Interspeech, pp. 239-243, 2015.

[3] 望月ら，“話者照合のためのポップノイズに含まれる音素情報を用いた声の生体検知の検討，”日本音響学会秋季大会, No. 3-Q-9, pp.107-108, 2016 年 9 月 16 日.

[4] G. Elko, *et.al.*, “Electronic pop protection for microphones,” in Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on, Oct 2007, pp. 46-49.

[5] NIST 2016 Speaker Recognition Evaluation Plan, `https://www.nist.gov/sites/default/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf`