話者照合のための話者性を考慮した音素情報に基づくポップノイズ検出法による声の生体検知*

◎望月紫穂野, 塩田さやか, △貴家仁志 (首都大東京)

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．一方で，登録話者の声を録音し，再生するなりすまし攻撃や少量の学習データから目標話者の声を作る技術である音声合成などを用いて登録話者を模倣するなりすまし攻撃によって話者照合の精度が大幅に低下してしまうことも報告されている [1]．そのため，話者照合システムの課題として精度向上だけでなく，なりすまし攻撃に対する頑健性向上も重要となり，国内外で活発に研究が行われている [2]．これまで提案されてきたなりすまし攻撃への主な対処法である話者照合システムのモデル表現や特徴量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人間が実際に発声したものか否かを判定する枠組みである声の生体検知が提案されている [3]．声の生体検知を実現する手法として，入力音声にポップノイズ [4] が発生し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．ポップノイズの発生原理と人の発声器官の仕組みから，人間が発声する際にポップノイズを発生させやすい音と発生させにくい音があると考えられる．そこでポップノイズ検出時にポップノイズ区間内の音素を考慮して声の生体検知を行う音素情報に基づくポップノイズ検出法を提案し，なりすまし攻撃に対する話者照合の頑健性が向上することを報告してきた [5]．

音素情報に基づくポップノイズ検出法では，音素情報を用いる際の音素リストを全話者共通で用いてきた．しかし，発音の傾向は話者毎に異なるため適切な音素リストも話者毎に異なると考えられる．そこで本研究では，話者毎に音素情報を設定することで，提案法がなりすまし攻撃に対してより頑健になることを報告する．

2 話者照合のための声の生体検知 [3]

2.1 ポップノイズ情報を用いた声の生体検知

近年，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．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発声したものなのかを識別する枠組みが提案された．声の生体検知は図 1 に示すように，話者照合と組み合わせて使用することを想定している．図 1 の例では声の生体検知部で入力された音声信号が実際に人間から発せられたものか否かを識別し，生体であると判定された場合のみ後段の話者照合に入力信号を渡すというフローになっている．これまでに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発生し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マイク内部に息や風が入りこむことにより変則的に振動板が揺れることで発生してしまうノイズのことを指す．

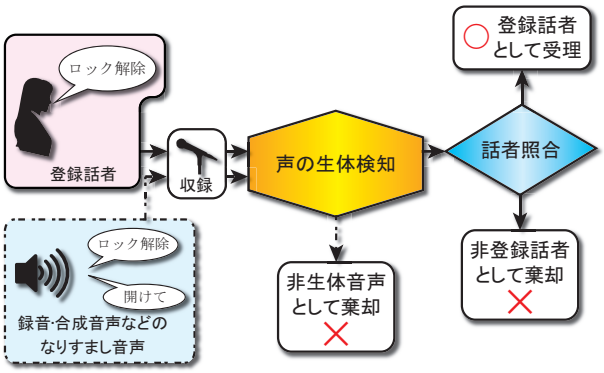

図 1 声の生体検知と話者照合システムのフロー

2.2 ポップノイズ検出法

本稿では，入力音声のポップノイズを検出するためにシングルポップノイズ検出法 [3] を用いた．ポップノイズは発話内で突発的におこるノイズのため，局所的に強いエネルギー変動を持つ性質がある．そのため，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．手順としてはまず，短時間フーリエ変換を行い，入力音声の周波数分解を行う．次にフレーム毎のパワースペクトルの低周波領域のみの平均を求める．この平均が低周波成分のエネルギーの推移を表し，フレーム間でのエネルギー変動が閾値より大きくなる区間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 1 本のマイクで実現可能であり，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

*Speaker adapted phoneme-based pop-noise detector for voice liveness detection and speaker verification. by MOCHIZUKI, Shihono, SHIOTA, Sayaka and KIYA, Hitoshi (Tokyo Metropolitan University)

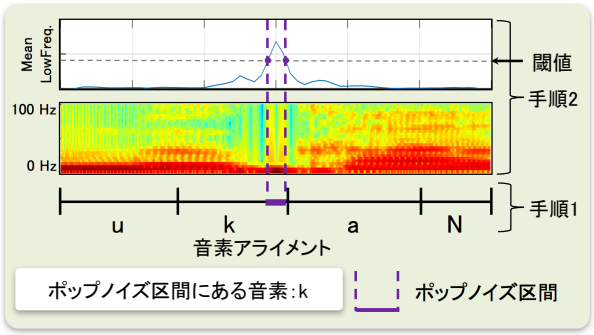

図 2　ポップノイズ区間にある音素の抽出

あげられる．

3　音素情報に基づくポップノイズ検出法による声の生体検知 [5]

3.1　ポップノイズと音素の依存性

2.2 節で述べたポップノイズ検出法を用いた声の生体検知では，なりすまし音声の中にはポップノイズが発生しておらず，人間の発話にはポップノイズが必ず発生していることを前提としていた．しかしながら，実際には再生音声でもポップノイズが検知される場合があった．

ポップノイズの発生原理と人の発声器官の仕組みから，言語毎にポップノイズを発生させやすい音と発生させにくい音があると考えられる．そこでポップノイズ検出後にポップノイズ区間内の音素の出現傾向を考慮した上で，生体音声か再生音声かを判定することでポップノイズ検出がより頑健になると期待できる．

3.2　ポップノイズの発生頻度と音素の傾向分析

ポップノイズとして検出された区間にある音素の傾向の調査には声の生体検知のために収録されたデータベースである VLD データベース [3] を用いる．VLD データベースには，風防カバーを装着しないで収録した音声データが収録されており，風防カバーなしのマイクで収録した音声データにはポップノイズが比較的多く発生している状態を想定している．そこで，傾向調査には風防カバーなしで収録したデータを用いた．ポップノイズ区間にある音素を抽出するための手順は以下に示す通りである．

1：音声データに対して音声認識を行い，音素アライメントを取得．
2：音声データに対し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比較して，ポップノイズ区間にある音素を抽出（図 2）．

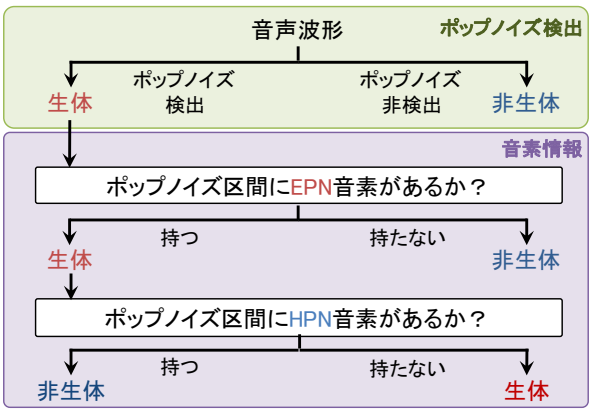

図 3　音素情報に基づくポップノイズ検出法による声の生体検知のフロー

ここで，ポップノイズを発生させやすい音素を EPN（Easily pop-noise phenomenon caused; EPN）音素，ポップノイズを発生させにくい音素を HPN（Hardly pop-noise phenomenon caused; HPN）音素とする．

3.3　音素情報に基づくポップノイズ検出法

音素情報に基づくポップノイズ検出法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発生しているならばその音声を生体による音声として受理する．発生していないならば非生体による音声として棄却する．次にポップノイズの検出精度を向上させるために，ポップノイズ区間にある音素情報を用いてさらに生体検知を行う．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，ポップノイズ区間に EPN 音素がある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 音素があるならば，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．逆にないならば，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．ここで，EPN 音素部分に背景雑音等でポップノイズが発生してしまった場合，なりすまし攻撃を誤受理してしまうことが想定される．このような誤受理を減らすために，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に，HPN 音素があるかどうかでさらに生体検知を行う．人間が発声した場合，HPN 音素の場所ではポップノイズが非常に発生しづらい．つまりポップノイズ区間に HPN 音素があることは，人間の発声としては不自然である．そこでポップノイズ区間に HPN 音素がある場合は再生音声として棄却し，HPN 音素がない場合は実発話として受理する．

表 1　VLD2 データベースの収録環境

マイク	AKG P170
再生用スピーカー	ELECOM LBT-SPP300
話者数	女性 15 名
サンプリング周波数	48 kHz

表 2　ポップノイズ検出および話者照合システムの設計条件

シングルポップノイズ検出	
サンプリング周波数	48 kHz
周波数帯域	[0,10] Hz
分析窓幅	200 msec
窓シフト幅	25 msec
話者照合（UBM-GMM）	
サンプリング周波数	16 kHz
分析窓長	25 msec
窓シフト幅	10 msec
特徴量	MFCC19 次+Δ+ΔΔ
UBM データベース	JNAS（女性のみ）
UBM 学習データ	165599 文
特定話者モデルデータベース	VLD2 [5]
特定話者モデル学習データ	60 文章 × 15 名

表 3　従来の音素リストおよび話者毎の音素リスト（F02，F05 のみ）

	話者	音素	従来音素と同じ音素リスト	従来音素と違う音素リスト
従来	共通	EPN	**b**, **e:**, **hy**, **k**, **ky**, **o**, **o:**, **s**, **sh**, **t**, **u:**	—
		HPN	**i:**, **m**, **ry**	—
話者毎	F02	EPN	**k**, **o**, **o:**, **s**, **sh**, **u:**	ch, d, ny, ts, y
		HPN	—	b, by, f, gy, my, p, q, sp, z
	F05	EPN	**hy**, **k**, **o**, **o:**, **s**, **sh**, **u:**	ch, e, g, w
		HPN	**i:**	ky, my, ny, sp

3.4　ポップノイズ区間内の音素の出現傾向と話者性

これまで EPN 音素および HPN 音素は，複数の話者から調査した一般的なポップノイズ区間内の音素の出現傾向から選択していた．しかしながら，話し方の違いなどにより話者毎に EPN 音素および HPN 音素が異なると考えられる．そこで，話者毎に EPN 音素および HPN 音素のリストを設定することで提案法の検出性能が向上することが期待できる．

4　評価実験

4.1　実験条件

話者性を考慮して EPN 音素および HPN 音素を選択することの有効性を確認するため，人間の実発話と収録した実発話をスピーカーで再生し収録した再生音声を用いて生体検知実験および話者照合実験を行った．実験のために，表 1 の条件で各話者それぞれ実発話 100 文/再生音声 100 文を防音室で収録し，データベースを構築した（以降 VLD2 データベース）．テストデータには VLD2 データベースから各話者それぞれ実発話 40 文/再生音声 40 文を用いた．発話内容はポップノイズの発生頻度を考慮したものを用いている．ポップノイズ検出および話者照合システムの設計条件を表 2 に示す．ポップノイズ検出で用いる音素アライメントの抽出には汎用大語彙連続音声認識エンジン Julius（Ver.4.3.1）のディクテーションキット（Ver.4.4，DNN-HMM 版）を使用した．話者照合に用いた UBM は，JNAS の音声および，JNAS の音声に電子協騒音データベースの展示会場の雑音を SN 比が 0，5，10，15，20，30dB となるよう重畳した音声を用いて学習した．

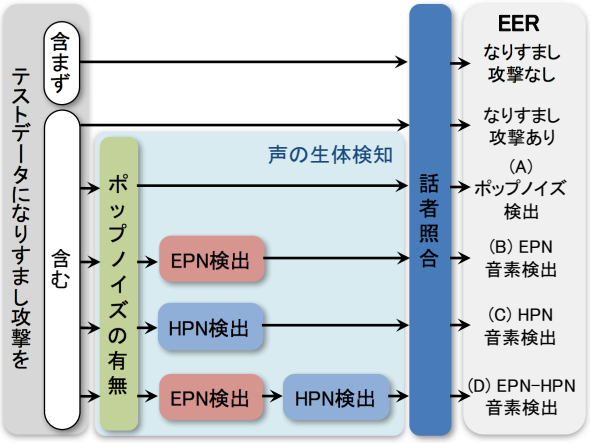

図 4　実験フロー

EPN 音素および HPN 音素のリストには，話者性を考慮していない従来の音素リストと話者毎の音素リストを用意した．各リストの作成方法は以下の通りである．

従来の音素リスト： VLD データベースに対してポップノイズ検出を行い，ポップノイズ区間にある音素とその音素の総数からポップノイズ区間にある割合を求め，ランキングを作成した．EPN 音素にはランキング上位 11 個の音素を選択し，HPN 音素にはランキング下位の音素 3 個を選択した．

話者毎の音素リスト： VLD2 データベースに対してポップノイズ検出を行い，ポップノイズ区間にある音素とその音素の総数からポップノイズ区間にある割合を求め，話者毎にランキングを作成した．EPN 音素にはランキング上位 11 個の音素を選択した．HPN 音素にはポップノイズ区間にあった割合が 0%の音素のみを選択した．

実際に用いた従来の音素リストおよび話者毎の音素リストを表 3 に示す．生体検知実験の評価尺度には以下に示す生体受理率を用いた．

$$\text{生体受理率} = \frac{\text{生体として受理されたサンプル数}}{\text{全サンプル数}}. \quad (1)$$

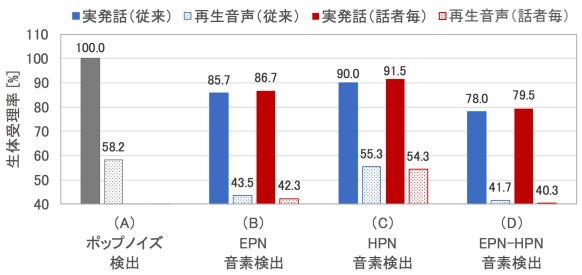

図 5　各手法の生体受理率

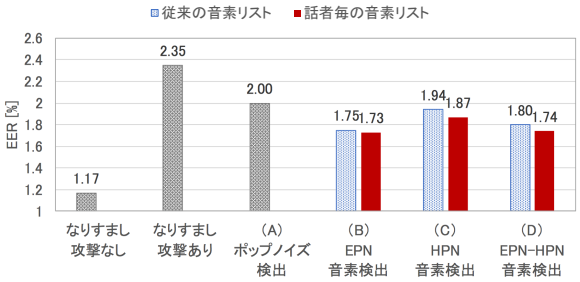

図 6　各手法における EER

話者照合の評価尺度には本人棄却率と他人受理率が等しくなる点である等価エラー率（EER）を用いた．図 4 に実験フローを示す．生体検知実験での各手法の詳細は以下の通りである：

（A）ポップノイズ検出：ポップノイズの有無のみで判定（従来法）．

（B）EPN 音素検出：ポップノイズ検出後に EPN 音素情報を用いて判定．

（C）HPN 音素検出：ポップノイズ検出後に HPN 音素情報を用いて判定．

（D）EPN-HPN 音素検出：ポップノイズ検出後に EPN 音素情報を用いて生体検知し，その後 HPN 音素情報を用いて判定．

話者照合実験での比較手法の詳細は以下の通りである：

なりすまし攻撃なし：なりすまし攻撃を含まないテストデータに対し，声の生体検知を行わずに話者照合を行い，EER を算出．

なりすまし攻撃あり：なりすまし攻撃を含むテストデータに対し，声の生体検知を行わずに話者照合を行い，EER を算出．

声の生体検知との組み合わせる手法は上述の各手法（A）～（D）を行った後，話者照合を行い EER を算出した．ただしポップノイズ検出に用いる閾値は，(A) において実発話が全て受理される最大値とした．

4.2　実験結果

図 5 に各手法の生体受理率を示す．従来の音素リスト使用時（青）に比べ，話者毎の音素リスト（赤）を用いたときの方が全ての手法で生体受理率が実発話で増加し，再生音声で減少した．このことから，話者性を考慮した音素リストを用いることで生体検知の精度が向上するといえる．また，音素情報を用いた手法を比較すると，EPN 音素もしくは HPN 音素どちらかだけを用いた手法（B，C）よりも両方用いた（D）の方がなりすまし攻撃に対して高い頑健性を持っていることがわかる．一方で，実発話の棄却率が高いことが課題だと言える．

図 6 に各手法の EER を示す．なりすまし攻撃なしの EER に比べ，なりすまし攻撃ありの EER が非常に高くなっている．これはなりすまし攻撃に用いられた再生音声によって話者照合システムの頑健性が低下することを示している．次に（A）の EER に着目すると，攻撃なしよりも EER が低下している．これはポップノイズ検出により話者照合に入力される再生音声の数を減らすことができたためである．(A) の EER と比較したとき，音素リスト使用時の EER は全て低くなっている．さらに話者毎の音素リストを使用することで，従来法の音素リストを使用したときよりも EER が低下することを確認できる．これは図 5 の結果からも妥当な結果といえる．

5　おわりに

本稿では，話者性を考慮した音素情報に基づくポップノイズ検出法による声の生体検知ついて提案し，実験により，その有効性を示した．今後の課題として，ポップノイズが生じた再生音声のテストデータでの実験等が挙げられる．

謝辞　本研究の一部は科学研究比基盤（B）2628006 による．

参考文献

[1] Z. Wu, et. al., “Spoofing and countermeasures for speaker verification: a survey,” Speech Communication, Vol. 66, pp.130-153, 2015.

[2] Z. Wu, et. al., “Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,” Training, Vol.10, No.15, p.3750, 2015.

[3] S. Shiota, et. al., “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” In INTERSPEECH, pp.239-243, 2015.

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

[5] 望月ら, “話者照合のためのポップノイズの発生頻度を考慮したプロンプト文を用いた声の生体検知,” 情報処理学会音楽情報科学研究会 音学シンポジウム, vol.2017-MUS-115, no.57, pp.1-6, 2017.