話者照合のための低周波および高周波成分情報を用いた声の生体検知の提案*

☆矢口凌也, 塩田さやか, △貴家仁志 (首都大東京)

1　はじめに

近年，銀行やスマートフォンなどで様々な生体認証技術が使われるようになってきている．また，声を用いた生体認証システムである話者照合も実用化が進んでいる．しかしながら，登録話者の声を録音・再生することで登録話者を詐称するなりすまし攻撃や，音声合成・声質変換といった声を生成し再生するなりすまし攻撃により話者照合の精度が低下してしまうことが重大な問題となってきている [1]．特に録音・再生によるなりすまし攻撃は，専門的知識をもたない詐称者でも簡単に詐称が行えてしまうため対策が急務と言える．この問題に対して，これまでに様々な音響的特徴量を用いた統計モデルによる対策手法が提案されてきた [2–4]．しかし，音響的特徴量を模倣する手法も考えられているため，より根本的な解決策が必要となる．そこで，入力音声が実際に人間が発声したものか否かを判別する声の生体検知が提案されている [5]．同論文で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否かを検出するポップノイズ検出法が提案されており，人間による実発話と再生機器を用いた再生音声の判別において高い識別性能が得られることが報告されている．一方，なりすまし音声は再生時および録音時にエイリアシング防止のためのフィルタリング処理によって，ナイキスト周波数付近の高周波数成分が減衰する傾向があることが報告されている [6–8]．そこで本研究ではこれまでのポップノイズ検出法による判別に加え，入力音声におけるナイキスト周波数付近の高周波数成分の有無による判別を行うことで，なりすまし攻撃に対し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 [5]

2.1　ポップノイズを用いた声の生体検知

声の生体検知とは，入力音声が実際に人間が発声したものなのか，スピーカー再生によるものなのかを判定するシステムである．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で，なりすまし攻撃による話者照合システムの精度低下を防ぐことを主な目的としている．センサーや機器を増やしシステムを煩雑化することなく，声の生体検知で用いる入力音声をそのまま話者照合システムにも使用できる状況を実現するために，本研究では入力音声に含まれるポップノイズに注目する．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呼気などが入り込むことで起こるノイズを指す．

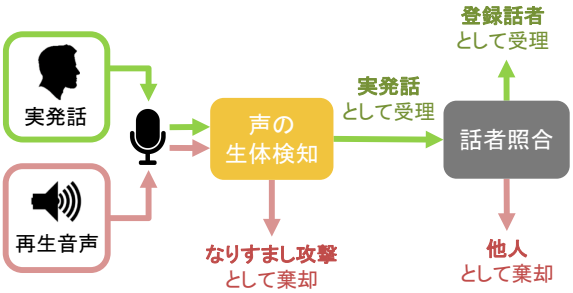

図 1　話者照合と声の生体検知の全体図

2.2　シングルポップノイズ検出法

ポップノイズには低周波数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の急激な変動を捉えることで検出を行う．具体的な手順は以下の通りである．

1. 入力音声にフレーム処理を施し，フレームごとに短時間フーリエ変換を行う．
2. 得られた振幅スペクトルの低周波領域のみの値の平均を求め低周波エネルギーとする．
3. 閾値を設定し，閾値以上の低周波エネルギーを持つフレームをポップノイズとして検出する．
4. 入力音声にポップノイズが検出された場合は実発話として受理，検出されない場合はなりすまし攻撃として棄却する．

シングルポップノイズ検出法はマイク一つで実現可能であり，導入コストが低く，また話者照合

*Voice liveness detection using low and high frequency components for speaker verification.
by YAGUCHI Ryoya, SHIOTA Sayaka, and KIYA Hitoshi (Tokyo Metropolitan University)

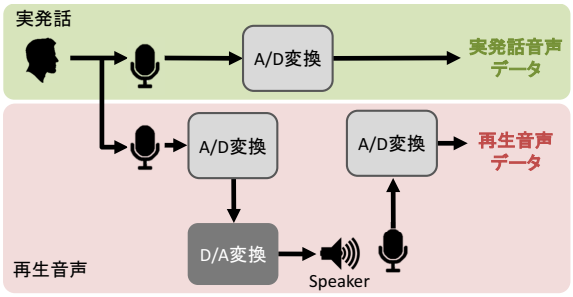

図 2　入力音声の詐称過程

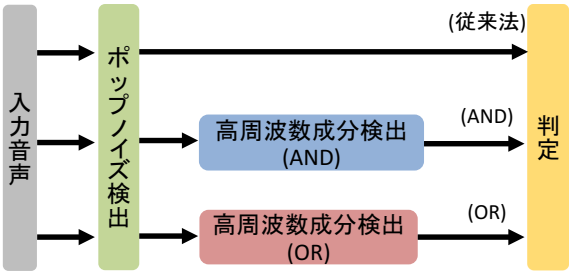

図 3　実験フロー

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　高周波数成分を用いた声の生体検知

3.1　なりすまし攻撃における高周波数成分

話者照合の精度低下につながる再生音声によるなりすまし攻撃をより高い精度で検出することが急務となってきている．実発話およびなりすまし攻撃音声の収録過程について図示すると図 2 のようになる．人間の発話においては A/D 変換が 1 回行われた後システムへの入力音声となるが，なりすまし攻撃の場合にはさらに，D/A 変換からスピーカーによる再生を経て再び A/D 変換が行われる．この過程において，エイリアシングを避けるために低域通過フィルタ処理が施されることにより，入力音声の高周波数成分が実発話と比べて減衰することが報告されている [6–8]．本稿では，収録音声のナイキスト周波数付近の高周波数領域に着目し，実発話と再生音声の高周波成分の違いを生体検知に用いることを考える．

3.2　高周波数成分検出法

高周波数成分を用いて生体検知を行うために，2.2 節で述べたポップノイズ検出法と同様の手法でナイキスト周波数に近い周波数帯のパワーの変動を検出することを考える．手順としては，ポップノイズ検出法と同様で，音声をフレームに分割し，各フレーム毎に短時間フーリエ変換を行い周波数分解を行い，得られた振幅スペクトルの高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高周波成分のエネルギーを表し，フレーム間のエネルギー変動が極大となる点を検出する．閾値を設定し，閾値以上のエネルギーを持つ場合生体として受理する．

3.3　ポップノイズ検出法との統合

なりすまし攻撃に対する声の生体検知の頑健性向上のため，ポップノイズ検出法と高周波数成分検出法を統合することを考える．図 3 に示すように，入力音声に対してポップノイズ検出を行い，入力音声にポップノイズが含まれるか否かを判定する．また，高周波数成分情報を用いる手法 (AND) および手法 (OR) では，入力音声に対して高周波数成分検出を行い，ナイキスト周波数付近の高周波数成分情報が存在するか否かを検出する．高周波数成分検出 (AND) では，ポップノイズ検出において生体として受理されかつ高周波数成分検出において生体として受理された入力音声のみを生体として受理する．高周波数成分検出 (OR) では，ポップノイズ検出，または高周波数成分検出どちらかで生体として受理された入力音声を生体として受理する．上記の 2 段階で構成される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の頑健性が向上すると考えられる．なお，ポップノイズ検出と高周波数成分検出はそれぞれ独立したシステムであるため実行順序は問わない．

4　評価実験

4.1　実験条件

ポップノイズ検出法および高周波数成分検出法を用いた声の生体検知の性能を評価するために VLD [5]，AVspoof [9]，ASVspoof2017 [10] という 3 つデータベースを用いて評価実験を行った．各データベースの詳細は表 1 にまとめてある．ポップノイズ検出法および高周波数成分検出法の評価に使用する尺度は，以下に示す実発話誤棄却率 (False Rejection Rate; FRR) と再生音声誤受理率 (False Acceptance Rate; FAR) が等しくなる点である等価エラー率 (Equal Error Rate; EER) を用いた．また，ポップノイズ検出法と高周波成分検出法の統合手法の評価尺度に

表 1　各データベースの詳細

VLD データベース	
サンプリング周波数	48 kHz
量子化ビット数	24 bit
話者	女性 15 名
サンプル数	実発話：1500 文 再生音声：1500 文
マイク	AKG
再生機器	BOSE 111AD
AVspoof データベース	
サンプリング周波数	16 kHz
量子化ビット数	16 bit
話者	男性 31 名
サンプル数	実発話：155 文 再生音声：155 文
マイク	話者照合用マイク
再生機器	高性能スピーカー
ASVspoof2017 データベース	
サンプリング周波数	16 kHz
量子化ビット数	16 bit
話者	男性 19 名
サンプル数	実発話：2268 文 再生音声：285 文
マイク	BQ Aquaris M5 Samsung Galaxy 7s
再生機器	Dell laptop speaker Dynaudio BM5A VIFA M10MD-39-08

は FRR と FAR を用いた．

$$\mathrm{FRR} = \frac{\text{誤棄却された実発話サンプル数}}{\text{全実発話サンプル数}} \quad (1)$$

$$\mathrm{FAR} = \frac{\text{誤受理された再生音声サンプル数}}{\text{全再生音声サンプル数}} \quad (2)$$

本実験での比較手法の詳細は以下の通りである (図 3)．

(従来法)：入力音声にポップノイズが検出された場合生体として受理．

(AND)：入力音声がポップノイズ検出および高周波数成分検出の両方で生体として判定された場合のみ受理．

(OR)：入力音声がポップノイズ検出または高周波数成分検出の少なくとも一方で生体と判定された場合に受理．

表 2　各データベースにおける EER

DB	EER [%]	
	ポップノイズ検出	高周波数成分検出
VLD	1.33	6.27
AVspoof	3.23	41.38
ASVspoof2017	23.42	11.78

ポップノイズ検出および高周波数成分検出は周波数分解能 5 Hz，分解窓幅 200 msec，窓シフト幅 25 msec の設定のもと行った．ポップノイズ検出に使用する周波数範囲は全データベース共通して 0 から 40 Hz とした．高周波数成分検出に使用する周波数範囲を VLD には 11 から 24 kHz，AVspoof，ASVspoof2017 には 7 から 8 kHz を使用した．上述の 3 つの比較手法においてポップノイズ検出に用いる閾値については VLD，AVspoof は実発話を全て受理する最小値とした．ASVspoof2017 では，実発話を全て受理する閾値が存在しないため，ポップノイズ検出法を用いた際の EER となる閾値を用いた．高周波数成分検出に用いる閾値は，全データベースで実発話を全て受理する閾値の設定が困難であるため，実発話受理率と再生音声受理率の差が最大となる閾値を用いた．

4.2　実験結果

表 2 にデータベースごとのポップノイズ検出法および高周波数成分検出法それぞれを用いた声の生体検知実験における EER を示す．まずポップノイズ検出法についてみると，ASVspoof2017 における EER は他のデータベースと比べてかなり悪くなっている．これは，ASVspoof2017 のデータがポップノイズを含むように収録されておらず，また背景ノイズなどが強く乗っている収録環境があまり良くないデータベースであることが原因であった．しかし，VLD および AVspoof においては低い EER となっており，ポップノイズ検出法の有効性が確認できる．次に，高周波数成分検出を用いた声の生体検知実験では，AVspoof における EER は他のデータベースに比べてかなり悪くなっている．これは，AVspoof のなりすまし攻撃音声の再生に高周波数領域の再現力が高い高性能スピーカーを使用しているため，実発話となりすまし攻撃音声の高周波数成分の差が小さいためであると考えられる．しかし，VLD および複数種の再生機器を使用している ASVspoof2017 において

表 3　各比較手法における FRR と FAR

DB	手法	FRR [%]	FAR [%]
VLD	(従来法)	0.00	8.93
	(AND)	1.53	0.47
	(OR)	0.00	8.93
AV spoof	(従来法)	0.00	10.97
	(AND)	59.35	0.65
	(OR)	0.00	23.87
ASV spoof 2017	(従来法)	24.03	20.35
	(AND)	26.68	8.77
	(OR)	0.35	48.77

は低い EER となっている．特に ASVspoof2017 に関してはポップノイズ検出法よりも低い EER となっており，収録条件に依存するものの高周波数成分検出法の有効性が確認できる．

表 3 にデータベースごとの従来法と統合手法の比較結果を示す．EER の時と異なり，ポップノイズ検出法における閾値が実発話をなるべく受理するように設定されている．まず VLD について見てみる．ポップノイズ検出のみの従来法と高周波数成分検出と統合する AND の FAR を比較すると数値が大幅に下がっていることがわかる．VLD については高周波数成分と統合した AND の方が FRR をあまり上昇させずに FAR を低下させることができているためシステムが頑健になると期待できる．次に AVspoof について見ると．従来法と統合手法である AND を比較すると FAR が大幅に改善されているものの，FRR が非常に悪化している．表 2 で示した EER においても AVspoof の高周波数成分検出法は非常に悪く高周波数成分検出法が機能していないためだと考えられる．一方，ASVspoof2017 においては，ポップノイズ検出自体の精度が低いため従来法の時点で FRR・FAR どちらも悪くなっている．しかし，統合手法 AND においては FRR が多少悪化しているものの，FAR が大幅に改善していることがわかる．なりすまし攻撃検出のコンペティションである ASVspoof2017 の結果においても，高周波数成分を利用した音響的特徴量を用いたグループは高い性能を得られていたことからも，ASVspoof2017 は高周波数成分を判定に用いるのに適したデータベースであったと言える．

以上の結果より，収録環境に依存するものの高周波数成分を用いた場合になりすまし攻撃に対して頑健になることを確認できた．特にマイクや背景雑音などの収録条件を調整できる ATM や建物の入出管理などに有効だと期待できる．

5　おわりに

本稿では，ポップノイズ検出法と高周波数成分検出法を統合した声の生体検知法を提案した．3 つの言語や収録環境の異なるデータベースを用いた性能評価を行い，提案法の有効性を確認した．今後の課題として，高周波数領域における分類手法の検討や統計モデルに手法などが挙げられる．

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. Jelil *et. al.*, “Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features,” in Proc. Interspeech, pp.22–26, 2017.

[3] R. Font *et. al.*, “Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge,” in Proc. Interspeech, pp.27–31, 2017.

[4] X. Wang *et. al.*, “Feature selection based on CQCCs for automatic speaker verification spoofing,” in Proc. Interspeech, pp.32–36, 2017.

[5] 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.

[6] M. Witkowski *et. al.*, “Audio Replay Attack Detection Using High-Frequency Features,” in Proc. Interspeech, pp.27–31, 2017.

[7] G. Lavrentyeva *et. al.*, “Audio replay attack detection with deep learning frameworks,” in Proc. Interspeech, pp.82–86, 2017.

[8] P. Nagarsheth *et. al.*, “Replay Attack Detection using DNN for Channel Discrimination,” in Proc. Interspeech, pp.97–101, 2017.

[9] AVspoof, http://www.idiap.ch/dataset/avspoof

[10] ASVspoof 2017, http://www.asvspoof.org