テキスト依存型話者照合のための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，話者照合が普及しつつある一方で，登録話者の声を録音した音声などを用いた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が報告されている [1]．そのため，話者照合システムの課題としてシステム自体の精度向上だけでなく，スピーカー再生によるなりすまし攻撃に対する頑健性向上も重要課題となり，国内外で活発に研究が行われている [2]．

これまでにスピーカー再生によるなりすまし攻撃に対する根本的な解決策の 1 つ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声の生体検知が提案されている [3]．また，同論文にて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．また，著者らはポップノイズ検出後にポップノイズ区間にかかる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，話者依存の音素リストやプロンプト文の使用を前提としてきたが，近年普及しつつあるスマートスピーカーやスマートフォン等の認証システムは発話内容を固定することを前提にしている場合が多い．そこで本研究ではテキスト依存の音素情報を用いた生体検知法を行い，日本語および英語の特定フレーズを用いた生体検知実験による性能評価と考察について報告する．

2 声の生体検知 [3]

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かを識別する枠組みであり，図 1 に示すように話者照合の前段としてなりすまし攻撃を棄却する役割を担っている．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ために発生してしまう

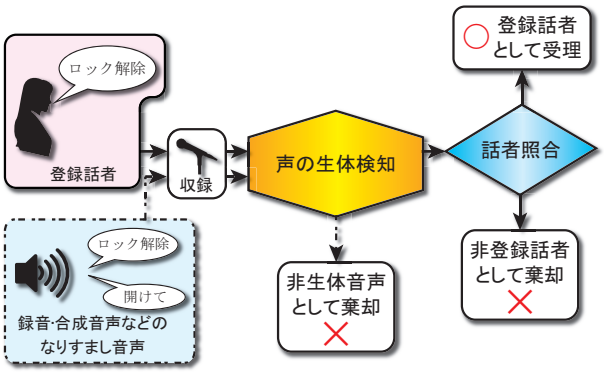

図 1 声の生体検知と話者照合システムのフロー

ノイズのことを指す．

2.2 シングルチャネルポップノイズ検出法 [3]

本稿では，入力音声のポップノイズを検出する手法としてシングルチャネルポップノイズ検出法を用いた．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．そこで，シングルチャネ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（図 2（b））．次にフレーム毎に低周波領域 $[0, F]$ Hz のパワースペクトルの平均を求める（図 2（c））．この平均が低周波成分のエネルギーの変動を表し，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が発生している区間として検出する（図 2）．シングルチャネ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あげられる．

3 テキスト依存型の音素情報に基づくポップノイズ検出法

3.1 音素情報を用いたポップノイズ検出法 [5]

ポップノイズの発生原理と人の発声器官の仕組みから，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にかかる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

*Voice liveness detection based on pop-noise detector considering phoneme information for text-dependent speaker verification. by MOCHIZUKI, Shihono, SHIOTA, Sayaka and KIYA, Hitoshi (Tokyo Metropolitan University)

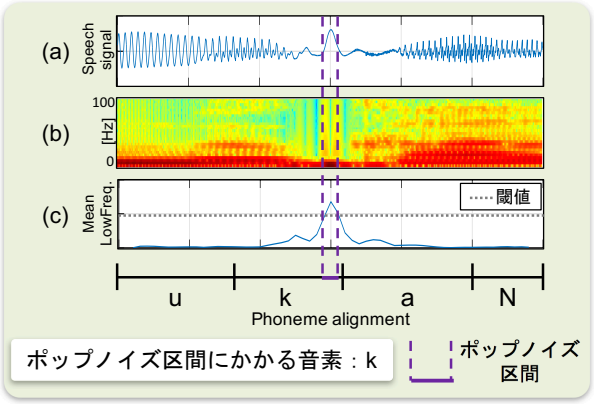

図 2　ポップノイズ区間にかかる音素の抽出

ポップノイズ検出がより高精度になると考えられる．

ポップノイズとして検出された区間にかかる音素の傾向の調査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チャネ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かかる音素を抽出（図 2）．

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

上述の手順により得られた EPN，HPN 音素を用いた音素情報に基づくポップノイズ検出法について説明する．全体の流れは図 3 に示すとおりとなっており，まず 2.2 節のシングルチャネ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区間を検出する．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し，発生していないならばなりすまし攻撃として棄却する．次に検出されたポップノイズ区間に EPN 音素がかかるかを判定し，かかる場合には人による発話によって発生したポップノイズとして入力音声を受理する．逆にポップノイズ区間に EPN 音素がかからない場合は，なりすまし攻撃と想定されるため棄却する．最後に，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 音素がかかるかどうかで生体検知を行う．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ため，ポップノイズ区間に HPN 音素がかか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 音素がかかる場合はなりすまし攻撃として棄却し，HPN 音素がかからない場合は生体による音声として受理する．HPN 音素まで確認するのは，詐称者が音声を再生中に故意にポップノイズを発生させた場合にも適切な区間でポップノイズが発生していないと受理できないようにするためである．

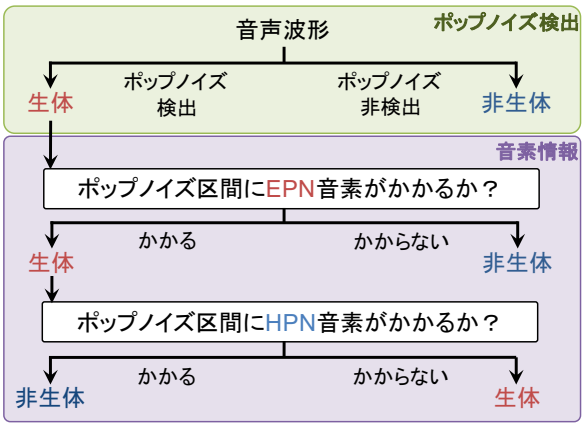

図 3　音素情報を用いたポップノイズ検出法による声の生体検知のフロー

3.2　テキストを考慮した音素情報の選択

スマートフォンやスマートスピーカー，ネットバンキングなどに代表される話者照合システムは，発話内容が固定される場合が多い．本稿では，テキスト依存型話者照合と親和性の高いポップノイズ検出法について提案する．これまでの音素情報に基づく手法では，EPN 音素および HPN 音素のリストを複数の文章から調査した一般的な音素の出現傾向から作成してきた．しかしテキスト依存型話者照合を想定する場合，事前に入力される文章が固定されているため，ポップノイズの発生頻度などを考慮できない一方で，フレーズ毎に音素リストを設定することが可能である．文章固有の音素リストを用いることでテキストにより則したものになると考えられるため，なりすまし攻撃に対する頑健性が向上することが期待できる．

4　評価実験

4.1　実験条件

文章毎に EPN 音素リストおよび HPN 音素リストを設定することの有効性を確認するため，VLD データベース [3]，VLD2 データベース [5] および AVspoof データベース [6] を用いて生体検知実験を行った．ここで，VLD および VLD2 データベースはポップノイズ検出法のために収録された日本語データベースで，マイク（AKG P170）に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いる．収録されている音声は実発話および実発話をスピーカー（ELECOM LBT-SPP300）で再生し収録した再生音声となっている．VLD および VLD2 データベースの発話内容は同じで

表 1　実験条件

ポップノイズ検出条件		
データベース	VLD,VLD2	AVspoof
周波数帯域	[0,50] Hz	[0,65] Hz
分析フレーム長	20 msec	15.4 msec
フレームシフト	5 msec	7.7 msec
音素リスト作成データ		
データベース	VLD	AVspoof(学習)
サンプリング周波数	48 kHz	16kHz
話者数	17 名	14 名
文章数	10 文	5 文
サンプル数	実発話 170 文	実発話 280 文
テストデータ		
データベース	VLD2	AVspoof(テスト)
サンプリング周波数	48 kHz	16kHz
話者数	15 名	18 名
文章数	10 文	5 文
サンプル数	実発話 150 文 再生音声 150 文	実発話 360 文 再生音声 360 文

ある．一方，AVspoof データベースは話者照合に対するなりすまし攻撃検出のための英語データベースであり，スマートフォンや高性能マイク（AT2020USB+）などを用いた収録を行っているものである．収録されている音声は実発話および実発話の録音再生や合成音声・声質変換による合成音声のスピーカー再生音声となっている．発話内容は 3 つのパートに分かれているが，本研究ではパスフレーズの読み上げのみを用いた．また，ポップノイズ検出法では低周波成分に着目するため，検出精度が収録マイクの性能にも依存することが報告されていることから [7]，本実験では高性能マイクを用いたデータのみを使用している．表 1 にデータベース毎の実験条件を示す．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）を使用した．VLD および VLD2 データベースに対してはディクテーションキット（Ver.4.4，DNN-HMM 版）の音響モデルと言語モデルを使用し，AVspoof データベースに対しては，ニュースの読み上げデータや会話データなど，複数のデータベースから学習した英語音響モデル [8] を，言語モデルにはパスフレーズの発話内容に適したものを用いた．また，EPN 音素および HPN 音素のリストには，全文章共通の音素リスト（従来法）と文章毎の音素リスト（提案法）を用意した．各リストの作成方法としては，まず表 1 の音素リスト作成データに対してポップノイズ検出を行い，ポップノイズ区間にかかる音素とその音素の総数からポップノイズ区間にかかる割合を求め，ランキングを作成した．ここで，全文章共通の音素リストに用いたランキングは全ての音素リスト作成データから作成し，文章毎の音素リストに用いたランキングは文章毎に音素リスト作成データから作成した．EPN 音素にはランキング上位 11 個の音素を選択し，

表 2　VLD2 の全文章共通の音素リストおよび文章毎の音素リスト（文章 1，文章 2 のみ）

	文章	音素	全文章共通音素と同じ音素リスト	全文章共通音素と違う音素リスト
従来法	共通	EPN	**ky**,**e:**,**my**,**py**,**sh**,**by**,**d**,**u:**,**N**,**s**,**m**	—
		HPN	**r**	—
提案法	文章 1	EPN	**py**,**sh**,**u:**,**N**	ts,o,z,b,y,k,a
		HPN	—	q
	文章 2	EPN	**sh**,**by**,**d**,**m**	b,e,n,o,t,a,o:
		HPN	—	sp

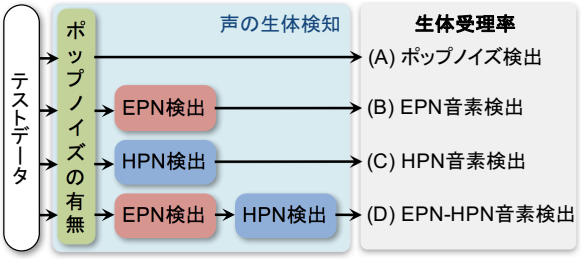

図 4　手法毎の実験フロー

HPN 音素にはランキング最下位の音素のみを選択した．実際に用いた全文章共通の音素リストおよび文章毎の音素リストを表 2 に示す．また，本実験で用いたテストデータを表 1 に示す．ただし発話内容は音素リスト作成データと同じであり，データベース毎にそれぞれ全話者共通となっている． 生体体検知実験の評価尺度には以下に示す生体受理率を実発話，なりすまし攻撃それぞれのデータに対して用いた．

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

実験に用いた各手法の詳細は以下の通りである：

（A）ポップノイズ検出：発話内におけるポップノイズの有無のみで判定．

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

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

（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．

図 4 に各手法の実験フローを示す．また，ポップノイズ検出に用いる閾値については実発話を 90%受理する値で固定とした．

4.2　実験結果

図 5 に VLD2 データベースを用いたときの手法毎の生体受理率を示す．実発話（塗りつぶし）の割合が高く，再生音声（ドット）の割合が低い方が理想的な

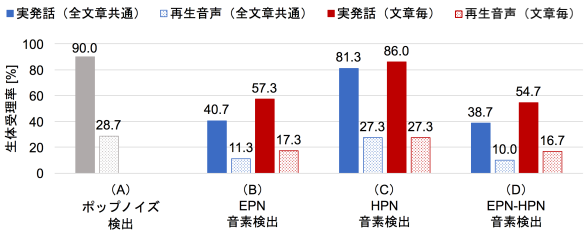

図 5　生体受理率（VLD2 データベース）

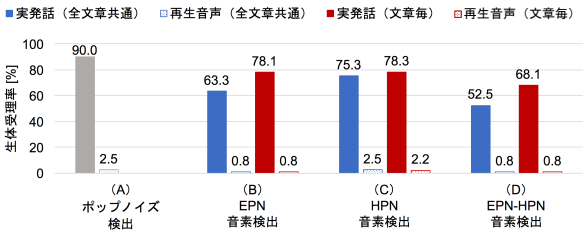

図 6　生体受理率（AVspoof データベース）

状態を表す．まず全文章共通の音素リストを用いたときの生体受理率（青）に着目すると，音素情報を用いることで（A）ポップノイズ検出に比べ，より多くの再生音声が棄却できる一方で，実発話も多数誤棄却してしまっていることがわかる．次に全文章共通の音素リスト使用時と文章毎の音素リスト使用時（赤）の生体受理率を比較すると，文章毎の音素リストを用いることで音素情報を用いた全手法（B，C，D）で実発話の受理率が改善していることがわかる．全文章共通の音素リストは複数文章を用いて作成したポップノイズ区間にかかる音素のランキングから選択されているため，文章によっては音素リストにかかる音素が文中に 2，3 個しか含まれていない等，音素情報の判定に用いることができる音素数が少ない場合があった．しかし文章毎の音素リストはその選択方法から，音素リストに含まれる全ての音素が判定に利用可能なため受理率が改善したと考えられる．また，音素毎のポップノイズの発生しやすさは文章内容によっても変動するので，文章毎に音素リストを設定することでそのような変動もカバーできたと考えられる．一方，(B) EPN 音素検出および（D）EPN-HPN 音素検出で再生音声の生体受理率も増加している．これは全文章共通の音素リストに比べて，文章毎の音素リストは判定に利用できる音素数が増えるため，検出されたポップノイズ区間に EPN 音素が被りやすくなったためと考えられる．

図 6 に AVspoof データベースを用いたときの生体受理率を示す．実発話は VLD2 データベースの結果同様，文章毎の音素リストを用いることで全手法（B，C，D）で生体受理率が改善した．さらに再生音声の誤受理率も減少していることがわかる．

以上より文章毎に音素リストを設定することで音素情報に基づくポップノイズ検出法の検出精度が向上するといえる．また，言語の違いやデータベースの変化に対しても頑健であるといえる．

5　おわりに

本稿では，テキスト依存型話者照合と親和性の高いテキスト固有の音素リストを用いたポップノイズ検出法について提案した．日本語および英語の発話内容固定データベースを用いた実験から，文章毎に音素リストを設定することで再生音声の誤受理率は若干増加してしまうものの，実発話の誤棄却率を大幅に減少させられることを示した．今後の課題として，話者性およびテキストの両情報を考慮した音素情報の選択などが挙げられる．

謝辞　本研究の一部は科学研究費基盤（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] K. Tomi, et al., “The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection,” In INTERSPEECH, pp.2-6, 2017.

[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，“話者照合のための音素情報を考慮したポップノイズ検出法による声の生体検知，” 電子情報通信学会 論文誌, vol.J101-D, no.3, 2018 年.

[6] S. K. Ergünay, et al., “On the vulnerability of speaker verification to realistic voice spoofing,” In Proc. Int. Conf. on Biometrics: Theory, Applications and Systems, 2015.

[7] 矢口ら，“声の生体検知のためのポップノイズ検出法の ASVspoof2017 に基づく評価，” 電子情報通信学会 バイオメトリクスと認識・認証シンポジウム, no.S2-16, 2017 年.

[8] VoxForge `http://www.voxforge.org`