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 *

☆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法である話者照合システムに登録話者の声を録音した音声や合成音声をスピーカーで再生したものを使用するなりすまし攻撃が問題となってきている．そこで話者照合のための声の生体検知という枠組みが提案された．本稿ではポップノイズを発生させやすい音素，発生させにくい音素の情報を用いることで声の生体検知の頑健性を向上させることについて検討する．

2　ポップノイズ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たものを入力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処するために様々な手法が提案されているが，主に音響的特徴量を用いるものが主流であり，なりすまし攻撃に話者照合システムが使用する音響的特徴量を用いることで簡単に話者照合システムを破ることができてしまうという問題があった．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 [1]．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．従来のポップノイズ検出法では，なりすまし攻撃には恣意的にも偶発的にもポップノイズは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恣意的にまたは，風などで偶発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くなる．一方，人が発話したときにポップノイズを発生させやすい音素と発生させにくい音素には傾向があることが報告されている．この傾向を用いてポップノイズ区間内に実発話のポップノイズを発生させやすい音素を持つかどうかで生体検知をする方法を提案し，高いなりすまし検出精度が得られることを報告した [2]．

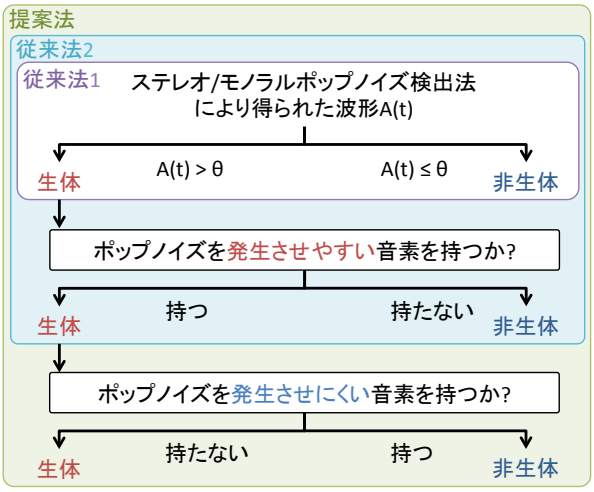

Fig. 1　提案法フロー

3　提案法

従来のポップノイズを発生させやすい音素情報を用いることに加え，本稿ではポップノイズを発生させにくい音素情報を用いた声の生体検知を提案する．図 1 に示すように，提案法では従来法で生体として受理されたサンプルに対して，さらにポップノイズを発生させにくい音素を持つかどうかで生体検知を行う．具体的な手順としては，ポップノイズを発生させにくい音素リストを先行音素，中心音素，後続音素毎に設定し，サンプルのポップノイズ区間にその中の 1 つ以上該当する音素を含む場合そのサンプルを非生体として棄却し，そうでない場合は生体として受理する．これによりポップノイズを恣意的に発生させたなりすまし攻撃に対しても頑健性が維持できると考えられる．

4　実験

4.1　実験条件

提案法の性能を評価するために生体検知実験を行った．評価のためポップノイズを恣意的に生じさせた音声とそうでない音声を用意した．これは収録時，故意に風を発生させた音声と発生させていない音声を屋外で収録したデータである．ポップノイズ検出にはステレオ/モノラルポップノイズ検出法の縦列接続を用いた [1]．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．また再

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

生音声に今回収録した音声をスピーカー（ELECOM LBT-SPP300）で再生したものを収録し用いた．発話内容はポップノイズの発生頻度を考慮していない．使用したデータ数は男性 5 名，文章数は各話者につき実発話 3 文/再生音声 3 文である．サンプリング周波数は 16kHz とした．ポップノイズ区間に含まれる音素の抽出は [2] と同様の手順である．評価尺度には生体による音声を非生体として拒絶してしまう生体拒否率（FRR）と非生体による音声を生体として受理してしまう非生体受入率（FAR）を用いた．比較手法として従来法 1，2 および提案法の 3 つを用いてそれぞれ FRR，FAR を算出した．従来法 1 は閾値のみを用いて，従来法 2 は従来法 1 の後ポップノイズを発生させやすい音素情報を用いて生体検知を行った．図 1 に各手法のフローを示す。

4.2　実験結果

図 2 は風無しで収録したデータ（Out）と風有りで収録したデータ（Out-Wind）それぞれでの閾値と生体として受理されたサンプル数の変化を示している．図 2 の Out と Out-Wind を比較すると，Out-Wind のデータベースの方が生体として受理されるサンプル数の変化が実発話と再生音声で近いことがわかる．これは故意に発生させた風がポップノイズを発生させたためである．次に閾値を Out-Wind で収録したデータベースのすべての実発話が生体として受理される値に定め（図 2 中の黒線の値），従来法および提案法を実行した．表 1 に実験結果を示す．はじめに従来法の結果について考察する．従来法 1 のとき，Out-Wind の FAR は Out の FAR に比べ 13%高い．また従来法 2 のとき，Out は従来法 1 に比べ FRR は変化せず FAR は 7%減少している．一方，Out-Wind のときは従来法 1 に比べ FRR が 7%増加している．これは今回使用したポップノイズを発生させやすい音素リストが合っていないため実発話を正しく検知できなかったと考えられる．また FAR は従来法 1 と変化がないことから，ポップノイズを発生させやすい音素情報によって再生音声を正しく棄却できなかったことを示す．これはポップノイズを発生させやすい音素部分に，再生音声も風によってポップノイズ区間を持ったためであると考えられる．以上より音声に意図的にポップノイズを発生させることで従来法による生体検知が破られてしまうことがわかる．次に提案法の結果について考察する．Out のとき，提案法では従来法 2 よりも FRR が 27%増加していることから，生体として正しく受理されていた実発話をポップノイズを発生させにくい音素情報によって誤棄却したことがわかる．これは再生音声に比べ実発話の方が多くポップノイズ区間を持つことから，ポップノイズ区間に含まれる音素の種類も多くなり，ポップノイズ区間に発生させにくい音素を含んだためと考えられる．Out-Wind のときも同様の理由で FRR が 20%増加したと考えられる．一方，FAR は Out のとき 46%，Out-Wind のとき 66%減少したことがわかる．このことから生体として誤受理されていた再生音声をポップノイズを発生させにくい音素情報によって誤棄却したことがわかる．また，Out と Out-Wind で FAR および FRR が同じになることから，提案法によって恣意的にポップノイズを発生させた音声の誤認識をそうでない音声と同じ水準まで減らすことができることが確認できた．

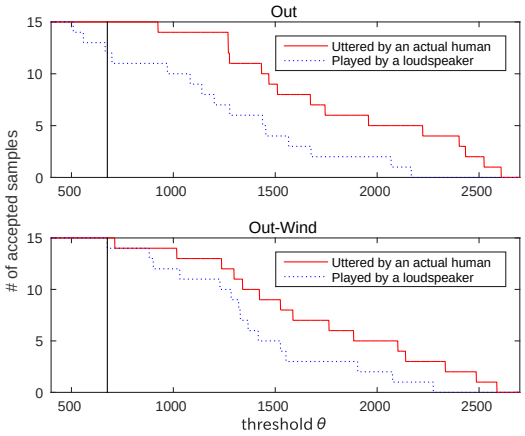

Fig. 2　風無しで収録したデータ（Out）と風有りで収録したデータ（Out-Wind）それぞれでの閾値と受理数の変化

Table 1　従来法および提案法の FRR，FAR

	従来法 1		従来法 2		提案法	
	FRR	FAR	FRR	FAR	FRR	FAR
Out	0%	80%	0%	73%	27%	27%
Out-Wind	0%	93%	7%	93%	27%	27%

5　おわりに

本稿では，ポップノイズを発生させやすいまたは発生させにくい音素情報を利用した声の生体検知について提案し，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ポップノイズの音素バランスを考慮したプロンプト文の使用，またテキスト依存型の話者照合システムとの連結などが挙げられる．

参考文献

[1] Sayaka Shiota, et al., “VOICE LIVENESS DETECTION FOR SPEAKER VERIFICATION BASED ON A TANDEM SINGLE/DOUBLE-CHANNEL POP NOISE DETECTOR,” Proc. Odyssey 2016, pp. 259-263, 2016.

[2] 望月紫穂野 他, “声の生体検知のためのポップノイズの音素バランスを考慮したプロンプト文についての考察,”研究報告音楽情報科学, 2016-MUS-111, No. 26, pp. 1-4, 2016.