The Seventh Symposium on Biometrics, Recognition and Authentication
2017/11/13-14 at National Institute of Advanced Industrial Science and Technology

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

ASVspoof2017-based evaluation of pop-noise detection method for voice liveness detection

矢口　凌也†　塩田　さやか†　貴家　仁志†
†首都大学東京　システムデザイン学部

RYOYA YAGUCHI†　SAYAKA SHIOTA†　HITOSHI KIYA†
†Tokyo Metropolitan University, Faculty of System Design

アブストラクト　本研究ではポップノイズ検出法による声の生体検知の ASVspoof2017 データによる性能評価とその分析結果について報告する．近年，声を用いた生体認証技術である話者照合において，スピーカーを用いた再生音声によるなりすまし攻撃が大きな問題となってきている．そこで人間の実発話とスピーカー再生音声を識別するポップノイズ検出法が提案され高い識別性能が得られることが報告されている．一方，話者照合に対するなりすまし検出手法のコンペティションである ASVspoof2017 において，実環境を想定した様々な録音・再生機器を用いたデータベースが公開された．そこで，本研究では ASVspoof2017 データベースを用いてポップノイズ検出を行い，特に収録機器ごとに結果を分析した．

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実用化に伴い，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への対策が急務となっている [1]．なりすまし検出手法のコンペティションである ASVspoof2017 [2] では，実際のなりすまし方法を考慮した様々な録音・再生機器を用いたデータをなりすまし攻撃として用いている．これまでに様々な音響的特徴量を用いたモデル化による対策手法が提案されてきた．一方，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ものか否かを判別するポップノイズ検出法による声の生体検知が提案されている [3]．これまでに高性能スピーカーによる再生音声を小型のコンデンサマイクや指向性マイクで収録した場合に高い検出性能が得られることが報告されている．しかしながら，ポップノイズ検出法では収録された音声における息の回り込みによって発生するポップノイズを検出するため，収録に用いるマイクの種類に対して依存度が高いことが想定される．そこで本研究では ASVspoof2017 データを用い，ポップノイズ検出法の精度を評価し，また収録機器ごとの結果についても分析し考察した．

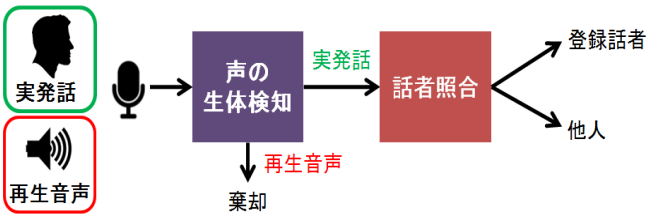

図 1: 声の生体検知と話者照合の全体図

2　ポップノイズ検出法を用いた声の生体検知 [2]

声の生体検知とは入力音声が人間の実発話かスピーカーによる再生なのかを判定する枠組みのことで，図 1 に示すように話者照合への前段として使用し，なりすまし攻撃による話者照合システムの性能低下を防ぐことを主な目的としている．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含まれているかを検出する手法が提案された．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．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．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　実験条件

シングルポップノイズ検出法を用いて ASVspoof2017 データの評価実験を行った．ASVspoof2017 データではサンプリング周波数 16kHz，量子化ビット数 16bit で収録された音声データが用意されている．学習・開発・評価データの文章数はそれぞれ実発話 760 文，1508 文，150 文，再

表 1: 収録機器 ID と再生機器 ID の関係

ID	収録機器	ID	再生機器
R02	BQ Aquaris M5 smartphone	P07	Dell laptop speaker
R03	Headset (desktop)	P04	Beyerdynamic DT 770 PRO
R04	H6 Handy Recorder	P02	All-in-one PC speaker
		P05	Creative A60
		P10	High Quality GENELEC Studio Monitors speakers
R06	Nokia Lumia	P09	HP Laptop speakers
R07	Rode NT2 (laptop)	P08	Dynaudio BM5A
		P15	VIFA M10MD-39-08
R08	Rode smartlav+ (laptop)	P08	Dynaudio BM5A
		P15	VIFA M10MD-39-08
R11	Samsung Galaxy 7s	P08	Dynaudio BM5A
		P15	VIFA M10MD-39-08

生音声が 950 文，1508 文，12008 文となっている．また，話者数は学習データが男性 8 名，開発データが男性 11 名である．ASVspoof2017 学習・開発データに使用されている収録機器は 7 種，再生機器は 8 種で，再生機器と収録機器の名称と組み合わせは表 1 の通りである．評価データに関しては機器情報が公開されていないため分析については学習・開発データを用いて行うことにする．次にシングルポップノイズ検出法の実験条件として，40Hz 以下の周波数範囲を使用し，周波数分解能は 20 Hz，分析窓幅は 50 msec，窓シフト幅は 6.25 msec となっている．評価尺度には生体棄却率と再生音声誤受理率が等しくなる点である等価エラー率（Equal error rate; EER）を用いた．

4 実験結果

表 2 に ASVspoof2017 における生体検知実験の EER を示す．ASVspoof2017 はポップノイズを含むように収録されていないため EER は悪くなっている．そこでさらに，収録条件が公開されている学習・開発データベースにおいて，使用された再生機器及び収録機器で分けた場合の EER の値を図 2，3 に示す．まず，再生機器別の結果を比較すると，P07 が一番低い EER となっている．一方で収録機器別の結果を比較すると R02 と R11 が低い EER となっている．しかしながら，P07 で再生した音声は R02 でのみ収録をしているためどちらの機器の影響か確認できない．一方，再生機器別で次に EER が低い P05 は R04 で収録されている．R04 で収録した他の P02，P10 はどちらも EER が高くなっていることから P02，P10 は再生音声にも関わらず低周波数帯に何らかのノイズが発生しやすく誤受理されやすく，逆に P05 はポップノイズ検出されにくい機器であると考えられる． 次に，収録機器のうち R07，R08，R11 の結果を比較すると，3 つとも再生に使用した機器が同じにもかかわらず R11 のみ低い

表 2: 声の生体検知実験結果

データの種類	EER[%]
学習 (train)	50.00
開発 (dev)	42.15
評価 (eval)	29.16

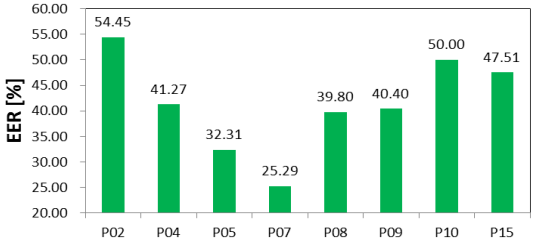

図 2: 再生機器別の EER[%]

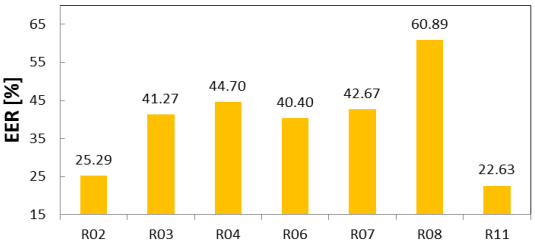

図 3: 収録機器別の EER[%]

EER となっている．これらの結果から，使用する再生機器よりも収録機器の方が，ポップノイズ検出法の精度に影響を与えると考えられる．そこでさらに追加実験として，iPhone6s，ELECOM LBT-SPP300，laptop 内蔵スピーカーを再生機器に，これまでのポップノイズ検出実験でも高い性能を示していた AKG P170 を収録機器とした収録を行った．収録条件は ASVspoof2017 と同様である．文章数は実発話が 9 文，再生音声が 150 文である．追加収録したデータを用いてポップノイズ検出を行った結果，どの再生機器を用いた場合にも EER が 6.67%と高い性能を得ることを確認した．

5 おわりに

本稿では，ポップノイズ検出法による声の生体検知法を用いて ASVspoof2017 データの性能評価を行い，録音・再生機器条件がポップノイズ検出法の精度に与える影響を調査した．今後の課題として，録音・再生機器，収録環境の多様性にも適応できるシステムへの改善等が挙げられる．

参考文献

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

[2] “ASVspoof 2017,” http://www.asvspoof.org .

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