Vol.2019-MUS-123 No.59
Vol.2019-SLP-127 No.59
2019/6/23

ASVspoof 2019データを用いた環境ごとにおけるなりすまし検出の性能評価に関する調査

奥野 桜子[1,a)]　塩田 さやか[1,b)]　貴家 仁志[1,c)]

概要：近年，人間の身体的特徴を利用して本人認証を行う生体認証システムが普及してきている．このうち人間の声を生体情報として用いる生体認証を話者照合と呼ぶ．話者照合はマイクがあれば実現可能であり導入コストが低いという利点があるが，他の生体認証技術と同様にスマートフォンやICレコーダーなどを使うことで簡単になりすましを行えることが問題視されている．近年話者照合のためのなりすまし検知手法に関するコンペティションASVspoofが2015年，2017年および2019年に開催された．特にASVspoof 2019ではなりすまし音声の収録環境や攻撃時の再生環境などの27通りの音響環境が用意された．そこで本研究では，ASVspoof 2019データセットおよびベースラインシステムを用いて録音再生音声の収録環境やなりすまし検出実験時の周囲の環境などによるなりすまし検出性能の違いについて調査した．実験結果より，再生機器の性能が非常に高い場合にのみなりすましが成功すること，また，収録環境では，部屋の大きさは広く，残響の少ない方が実発話データおよび録音再生音声データの検出性能が良くなることがわかった一方，部屋が小さい場合LFCC，CQCCでの検出は非常に困難であることを確認した．

Investigation of spoofing detection performance in different acoustic configurations with ASVspoof 2019 database

Abstract: In recent years, biometric authentication systems that perform personal authentication using human physical features have become widespread. Automatic speaker verification (ASV) is one of the biometric authentication by using human voice. Since ASV systems can be installed with a microphone only, it is easy to introduce to many applications. However, ASV systems suffer from spoofing attacks, e.g., speech synthesis and replay, as same as other biometric authentication systems. Therefore, ASVspoof challenges were held in 2015, 2017 and 2019, in order to establish the research topic of spoofing countermeasures. The published database in 2019 is created according to a total of 27 different acoustic configurations. Therefore, this paper investigates how the spoofing detection performance in the combinations of the acoustic configurations differ by using the baseline systems of ASVspoof 2019. From the experimental results, spoofing is successful only when a playback device is perfect quality. The spoofing detection performs well when the room size is large and the reverberation time is short. It also shows that it is difficult to detect spoofing attacks in small rooms each case of using LFCC or CQCC as acoustic feature.

1. はじめに

近年，人間の身体的特徴を利用して本人認証を行う生体認証システムが普及してきている．実用化の例として，スマートフォンの指紋認証や，入出国管理に用いられる顔認証などがある．生体認証システムのうち，人間の声を生体情報として用いる技術を話者照合と呼ぶ．話者照合はマイクがあれば実現可能であり導入コストが低いという利点がある．一方で，録音再生音声や合成音声を用いたなりすまし音声攻撃に対して脆弱であることが報告されている [1]．そこで，図1に示すようになりすまし検出を話者照合の前段で行うことを考える．近年，なりすまし検出手法について比較評価するコンペティションASVspoofが隔年で開催されているが，その際に想定されているなりすまし攻撃のフローが2系統ある．その1つが登録者の録音音声や合成音声などのなりすまし音声を認証時にスピーカーで再生

[1] 現在，首都大学東京 システムデザイン研究科 情報科学域
Presently with Tokyo Metropolitan University, FacultySchool of Systems Design, Department of Computer Science
a) okuno-sakurako@ed.tmu.ac.jp
b) sayaka@tmu.ac.jp
c) kiya@tmu.ac.jp

Vol.2019-MUS-123 No.59
Vol.2019-SLP-127 No.59
2019/6/23

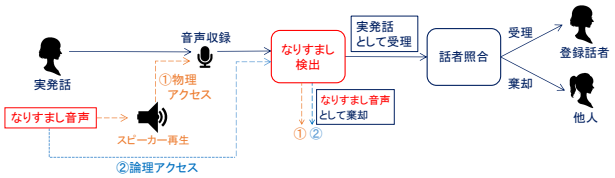

図 1　話者照合システムとなりすまし検出システムおよびなりすまし攻撃のアクセスフロー

する物理アクセス（図 1①) であり，もう 1 つがシステムの入力系統に直接合成音声等を割り込ませる論理アクセス (図 1②) である．物理アクセスと論理アクセスの根本的な違いはシステムに入力される音声がスピーカーで再生されてマイクで再収録されるというプロセスが存在するか否かということである．録音再生音声による物理アクセスは，専門的な知識を必要とせず容易に行えてしまうため，特に認証のフレーズが固定の場合現実的かつ検出が難しいなりすまし攻撃となっている．ASVspoof では，これまでに論理アクセス (2015 年)，物理アクセス (2017 年) の検出手法のコンペティションを開催してきた [2,3]．その結果それぞれのなりすまし攻撃に対して，先行研究として様々な音声特徴量およびモデルを用いた対策手法が提案された [4–7]．それらを踏まえて開催された ASVspoof 2019 では論理アクセスと物理アクセスそれぞれを想定したデータを公開して，コンペティションが開催された．ASVspoof 2019 の物理アクセスの特徴は，テスト音声の収録環境や録音音声の収録環境の条件がラベル付けされて公開されていることにある．そこで，本研究では ASVspoof 2019 データセットおよびベースラインシステムを用いたなりすまし検出実験を行い，様々な音響条件がなりすまし検出の精度に与える影響について調査した．実験結果より，再生機器の性能が非常に高い場合にのみなりすましが成功すること，また，収録環境では，部屋の大きさは広く，残響の少ない方が実発話データおよび録音再生音声データの検出性能が良くなることがわかった一方，部屋が小さい場合 LFCC，CQCC での検出は非常に困難であることを確認した．さらに，残響時間が長いほど入力音声はなりすまし音声と認識されやすくなることも報告する．

2. ASVspoof 2019

2.1 開催概要

2019 年に開催された ASVspoof 2019 [8] では，なりすまし攻撃として合成音声 (Text-To-Speech; TTS)，声質変換 (Voice Conversion; VC)，録音再生の 3 つの主要な攻撃に対する対策に焦点を当てている．ASVspoof 2019 で公開されたデータセットは物理アクセス，論理アクセス両方であった．本研究では再生音声のみの物理アクセスについて着目する．

表 1　ASVspoof 2019 なりすまし音声収録環境 ID

	A	B	C
攻撃者と話者の距離 [cm] (Da)	10-50	50-100	>100
再生機器の品質 (Q)	perfect	high	low

表 2　ASVspoof 2019 照合時の収録環境 ID

	a	b	c
部屋の大きさ [m](S)	2-5	5-10	10-20
残響時間 [ms](R)	50-200	200-600	600-1000
話者と ASV の距離 [cm] (Ds)	10-50	50-100	100-150

2.2 ベースラインシステム

本研究では ASVspoof 2019 で公開されたベースラインシステムを用いて調査を行う．使用するモデルは混合ガウスモデル (Gaussian Mixture Model; GMM) であり，入力特徴量に Linear Frequency Cepstrum Coefficients (LFCC) [9] と Constant Q Cepstral Coefficients (CQCC) [10] それぞれを用いた 2 つのシステムがベースラインシステムと設定されている．これらのシステムは過去のコンペティションにおいて比較的高い検出性能を得ていることからベースラインとして採用されている．

2.3 環境ごとの評価

ASVspoof 2019 で公開された物理アクセスのデータベースには，収録・攻撃環境ごとの ID が付けられている．なりすまし音声の録音環境および照合時の収録環境を表 1,2 に示す．収録音声の各環境 ID はそれぞれ 3 つのレベルに割り当てられている．そこで本研究では，各環境によるなりすまし検出性能への影響を調査することを目的として，ASVspoof 2019 データセットを用いた検出性能の評価を行い，なりすまし検出が困難な条件について調査した．

3. 実験

3.1 実験条件

本実験では ASVspoof 2019 のデータセットおよび 2 種類のベースラインシステムを用い，収録・攻撃環境ごとのなりすまし検出性能の比較を行った．なお，ASVspoof 2019 の条件設定が，ASVspoof 2019 データセット以外のデータを用いることは不可としているため他のデータは使用していない．ベースラインシステムの学習条件を表 3 にまとめる．本実験では環境の組み合わせごとにエラー率を算出するために，閾値を設定する必要がある．そこで LFCC-GMM，CQCC-GMM それぞれの等価エラー率 (Equal Error Rate; EER) を計算し，EER の時点での他人受入率 (False Acceptance Rate; FAR)，本人棄却率 (False Rejection Rate; FRR) を条件ごとに調査した．FAR，FRR の数値は全てパーセンテージを示している．実発話のデー

Vol.2019-MUS-123 No.59
Vol.2019-SLP-127 No.59
2019/6/23

表 3　ベースラインシステム学習条件

特徴量	LFCC，CQCC
次元数	20 次元 (LFCC), 29 次元 (CQCC)
モデル	GMM
サンプリング周波数	16kHz
GMM 混合数	64
攻撃音声	録音再生音声 (物理アクセス)
dev データ数	実発話：5400 なりすまし音声：24300
train データ数	実発話：5400 なりすまし音声：48600

LFCC

R	Ds	S a	S b	S c
a	a	10.00	0.50	0
	b	11.50	8.00	2.50
	c	4.50	6.00	2.00
b	a	9.00	3.50	2.00
	b	9.50	5.00	3.00
	c	8.00	9.00	6.00
c	a	34.00	19.50	12.00
	b	34.50	23.00	14.50
	c	30.00	23.00	22.00

CQCC

R	Ds	S a	S b	S c
a	a	17.00	3.50	3.50
	b	17.50	11.50	4.50
	c	18.00	22.50	5.50
b	a	3.50	3.50	6.00
	b	5.50	5.50	1.00
	c	4.00	6.50	4.00
c	a	17.00	12.50	12.00
	b	22.00	16.50	13.00
	c	19.50	14.50	18.00

図 2　実発話の FRR(%)

タ数は各環境ごとに 200 発話ずつである．なりすまし音声のデータ数は各環境ごとに異なっている．一番データ数の多い環境で 220 発話，一番少ない環境で 6 発話となっている．

3.2　実験結果

照合時環境ごとの実発話データの FRR を図 2 に示す．色が薄いほどエラー率が低いことを示している．この際の LFCC，CQCC の EER はそれぞれ 11.60%，10.66%であった．図 2 左の LFCC の結果より，実発話データでは部屋の大きさ (S) は広く，残響時間 (R) は短いほうが本人検出性能が高くなっている．一方，CQCC では，残響時間の条件が b の時に最も低いエラー率となっており，部屋の広さと検出性能に比例関係はない．また，話者と ASV の距離 (Ds) は，LFCC では関係性が見られないが，CQCC では残響時間がある程度短い場合に距離の近さと検出率に若干相関があることがわかる．このような特徴量による傾向の違いがあることが確認できた．

図 3 になりすまし収録・照合時環境ごとのなりすまし音声データの FAR を示す．なりすまし音声の FAR は，再生機器の品質が B または C のときには全て 0%となるため掲載していない．つまり，再生機器の品質が非常に高い場合のみ検出が難しいといえる．実発話データと同様に，部屋の大きさ (S) は右に行くほど色が薄くなることから広いほうが他人検出性能が高くなっている．一方，残響時間が長いほうが他人検出性能が高くなっている．これは実発話データとは反対の傾向である．これより，残響時間が長いと録音再生音声と判断される傾向であることがわかる．残響が重畳されることで音声がなりすまし攻撃とみなされやすくなると考えられる．実発話となりすまし音声の残響時間が短いときと長いときのスペクトログラムを図 4 に示す．発話内容は 4 つとも同じである．それぞれ残響時間の短い音声と長い音声を比較している．残響時間の長いほうのスペクトログラムは，時間方向に信号が流れているため低周波数領域にパワーが集まっているように見える．この影響により，残響時間による検出性能の偏りが生まれていると考えられる．また，なりすまし収録・照合時環境のうち，話者と ASV の距離 (Ds)，攻撃者と話者の距離 (Da) は

Da=A, LFCC

R	Ds	S a	S b	S c
a	a	68.32	12.66	8.22
	b	16.57	6.23	4.30
	c	6.47	3.41	1.68
b	a	5.12	3.24	3.62
	b	2.21	2.09	2.04
	c	1.75	0.71	0.59
c	a	0.10	0.77	0.44
	b	0.62	0.30	0.14
	c	0.51	0.08	0.03

Da=A, CQCC

R	Ds	S a	S b	S c
a	a	59.90	12.81	7.65
	b	14.79	6.03	4.30
	c	4.98	2.53	1.71
b	a	4.79	2.18	2.62
	b	3.31	2.14	1.98
	c	2.50	0.25	0.69
c	a	0.70	0.14	0.06
	b	0	0.21	0.26
	c	0.09	0.04	0.03

Da=B, LFCC

R	Ds	S a	S b	S c
a	a	40.00	3.09	0.76
	b	25.36	6.85	4.56
	c	18.18	6.12	5.86
b	a	0.36	0.40	0
	b	5.33	2.53	2.85
	c	2.82	2.25	1.86
c	a	0	0.17	0
	b	1.80	0.32	0.39
	c	0.64	0.15	0.10

Da=B, CQCC

R	Ds	S a	S b	S c
a	a	32.50	2.97	0.76
	b	23.19	6.29	4.49
	c	13.22	2.49	5.20
b	a	1.43	0.10	0.07
	b	8.38	3.38	2.97
	c	3.83	0.78	2.21
c	a	0	0	0
	b	0	0.24	0.28
	c	0.26	0.08	0.26

Da=C, LFCC

R	Ds	S a	S b	S c
a	a	74.14	4.49	3.06
	b	29.03	8.31	3.73
	c	21.09	9.96	5.25
b	a	1.59	1.52	1.21
	b	1.83	1.56	1.72
	c	4.96	1.83	1.01
c	a	0.18	0.49	0.17
	b	0.17	0.20	0.08
	c	0.81	0.18	0.08

Da=C, CQCC

R	Ds	S a	S b	S c
a	a	67.24	5.24	2.67
	b	16.94	7.21	3.29
	c	22.27	4.84	4.09
b	a	4.14	1.78	1.60
	b	4.45	3.11	2.10
	c	6.35	2.86	2.45
c	a	0.37	0.07	0.08
	b	0	0.20	0.37
	c	0.67	0.06	0.15

図 3　なりすまし音声の FAR(%)
(再生機器の品質 Q=A)

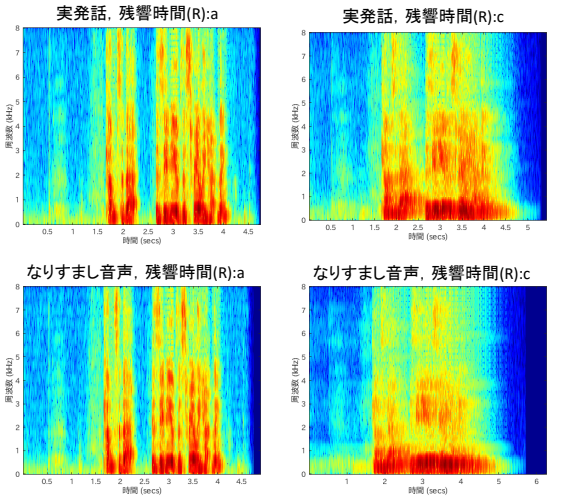

図 4　実発話となりすまし音声の残響時間ごとのスペクトログラム

結果にはあまり影響を与えていないことがわかる．

これらの結果からなりすまし検出には部屋が大きく，残響の少ない環境が適していること，残響時間が長い場所ではLFCC，CQCCを用いてもなりすまし検出が難しいことが確認できた．これらの環境においても頑健に動く検出手法を今後考えていく必要がある．

4. おわりに

本稿では，ASVspoof 2019データとベースラインシステムを用いてなりすまし音声検出，攻撃環境ごとの性能評価を行った．実験結果より，再生機器の性能が非常に高い場合にのみなりすましが成功すること，また，収録環境では，部屋の大きさは広く，残響の少ない方が実発話データおよび録音再生音声データの検出性能が良くなることがわかった一方，部屋が小さい場合LFCC，CQCCでの検出は非常に困難であることを確認した．

今後の課題として，ほかの特徴量やモデルを用いた場合との性能の比較，より適切な特徴量の提案などが挙げられる．

謝辞 本研究の一部はJSPS科研費若手研究JP19K20271の助成を受けたものです．

参考文献

[1] Z. Wu, *et al.*,“Spoofing and countermeasures for speaker verification: A survey” in Proc. Speech Communication, pp.130–153, 2015.
[2] ASVspoof 2015, http://www.asvspoof.org/index2015.html
[3] ASVspoof 2017, http://www.asvspoof.org/index2017.html
[4] M.S. Saranya, *et al.*,“Decision-level feature switching as a paradigm for replay attack detection” in Proc. Interspeech, pp.686–690, 2018.
[5] G. Suthokumar, *et al.*,“Modulation Dynamic Features for the Detection of Replay Attacks” in Proc. Interspeech, pp.691–695, 2018.
[6] F. Tom, *et al.*,“End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention” in Proc. Interspeech, pp.681–685, 2018.
[7] K. Sriskandaraja, *et al.*,“Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric” in Proc. Interspeech, pp.671–675, 2018.
[8] ASVspoof 2019, http://www.asvspoof.org/index2019.html
[9] T. Kinnunen, *et al.*,“Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research” in Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5395–5399, 2017.
[10] M. Todisco, *et al.*,“Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification” in Proc. Computer Speech & Language, pp.516–535, 2017.