なりすまし音声検出に対する録音再生攻撃の収録条件に関する影響分析 *

☆堀江涼花 (都立大), 高道慎之介 (慶応義塾大/東大), 塩田さやか (都立大)

1 はじめに

近年，情報セキュリティの重要性が高まる中，生体認証技術が注目されている．生体特徴として声を用いた生体認証技術を話者照合と呼ぶが，他の生体特徴を用いる生体認証技術と同様になりすまし攻撃への対策が急務となっている．そのため，なりすまし音声攻撃の検出技術の重要性も高まっており，その一環としてさまざまなデータベースが公開されてきている [1–3]．特に J-SpAW [3] は，話者照合となりすまし音声検出の両方の評価が可能な日本語音声データベースとして公開されており，実発話および録音再生によるなりすまし音声が多様な収録環境下で収録されている．しかし，J-SpAW におけるなりすまし音声は検出が比較的容易であり，困難ななりすまし音声検出タスクではないことが報告されている．そこで，本研究ではより検出が困難ななりすまし音声を作成するために，なりすまし音声収録時の条件がタスクに与える影響について分析調査した．具体的には，J-SpAW の不正録音音声を複数の攻撃収録環境で再収録して作成したなりすまし音声が，なりすまし音声検出及び話者照合の性能に与える影響を調査した．特に再生機器と攻撃収録機器の距離および信号対雑音比（Signal-to-Noise Ratio; SNR）の関係性に着目して攻撃収録を行い，分析を行っている．実験結果より，再生機器と攻撃収録機器の距離が離れている場合かつ平均 SNR が 5dB 程度の音量で再生した際になりすまし音声検出が困難になる傾向になることがわかった．また，話者照合においては実発話のみの場合と比べて今回作成したなりすまし音声を加えると，話者照合の評価が悪化することから，話者照合を突破できるより検出困難ななりすまし音声が作成できたことを報告する．

2 関連研究

2.1 録音再生によるなりすまし音声攻撃と検出

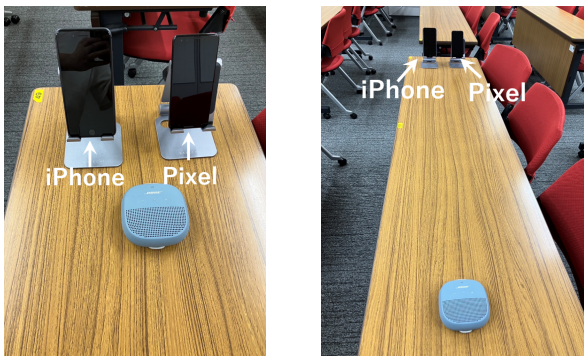
録音再生によるなりすまし音声攻撃とは，攻撃者がターゲット話者の音声を不正に録音し，スピーカーなどの再生機器で再生する攻撃手法である．なりすまし音声検出は入力された音声が実際に人間が発声した音声である実発話かなりすまし音声のどちらであるかを識別するタスクである．録音再生攻撃によるなりすまし音声攻撃の場合は，入力された音声が実発話であると受理，攻撃者が不正録音した音声を再生したなりすまし音声であると棄却と判断される．これまでの研究において，録音再生攻撃によるなりすまし音声検出の性能は，攻撃者が不正録音を行う際の収録環境や，なりすまし音声攻撃時の攻撃収録環境の影響を受けることが明らかになっている [4,5]．つまり，使用された再生機器や不正録音が収録された環境などの条件が，検出の性能に関わる重要な要素となっており，頑健性に影響を与えているといえる．このような背景から，なりすまし音声検出の頑健性向上を目指すためには，さまざまな収録条件を含むデータベースの整備が重要である．これまでになりすまし音声検出のためのデータベースとして様々なものが公開されているが，多くは声質変換や音声合成による攻撃を対象としており，録音再生による物理的な攻撃を扱ったものは限られている．上述のとおり，収録環境などの影響が大きいことからなりすまし音声検出の頑健性を向上させるためには様々な環境で評価できる必要があるため，現状の評価では不十分であるといえる．さらに，なりすまし音声の収録が検出性能に与える影響について調査された事例は少なく，収録環境の影響を緩和する方法についての知見は十分ではない．より頑健ななりすまし音声検出を実現するためにはこれらの影響を分析することが重要である．

2.2 J-SpAW

J-SpAW は話者照合評価セットとなりすまし音声検出評価セットの 2 つで構成された音声コーパスである．話者照合評価セットでは，様々な環境で実際に人が発声した音声の収録を行い，同時に攻撃者がやや離れた位置から不正に録音するという不正収録状況を想定した収録も行われている．なりすまし音声検出評価セットは，不正録音された音声を再生機器から再生し，再収録することにより作成されたものをなりすまし音声として扱っている．J-SpAW では攻撃者が用いる不正録音音声を収録する際に，マイクの位置が正規ユーザの口元近くではなく，正規ユーザに気づかれにくいようある程度距離をおいて収録されており，より現実的な録音再生によるなりすまし音声攻撃を行っているといえる．一方で，公開されている

---

* Impact of recording conditions in replay attacks on spoofed speech detection,
by HORIE, Suzuka (Tokyo Metropolitan University), TAKAMICHI, Shinnosuke (Keio University/The University of Tokyo),SHIOTA, Sayaka (Tokyo Metropolitan University)



(a) 近距離　(b) 遠距離

Fig. 1　近距離・遠距離での収録状況 (Bose)

なりすまし音声の検出率は高く，なりすまし音声検出のデータとして難易度はあまり高くないという問題がある．

3　なりすまし音声収録

録音再生攻撃の収録条件がなりすまし音声検出の性能に与える影響を調査するために，再生機器と攻撃収録機器の距離やなりすまし音声の SNR に着目し，なりすまし音声収録を行った．不正録音音声を再生する機器は iPad mini (第 5 世代) (iPad)，MacBook Pro (M2, 2022 年製) (Mac)，Bose Soundlink Micro Bluetooth Speaker Bundle (Bose)，Sony SRS-ZR7 (Sony) の 4 つである．iPad および Mac は内蔵スピーカーで音質より手軽さを重視しており，Bose はポータブルであるが音質が高く，Sony は据え置き型で音質が高い，という特徴を持っている．なりすまし音声の収録機器には，一般的なスマートフォンとして Pixel3 (Pixel)，iPhone8 Plus (iPhone) の 2 つを用いた．収録の際は，再生機器と攻撃収録機器の距離として遠距離と近距離の 2 種類，再生機器の音量として音量大と音量小の 2 種類の組み合わせで 4 種類の攻撃収録条件を用意した．再生機器と攻撃収録機器の距離は近距離では 10cm 程度，遠距離では 1m 程度の距離とした．Figure 1(a) に近距離での収録状況，Fig. 1(b) に遠距離での Bose を使ったなりすまし音声攻撃収録状況を示す．4 種類の収録条件は，音声信号の強度やノイズの影響に違いが生じるように設定した．近距離・音量大の条件では，収録機器への音圧レベルが高く，話者の音響的特徴が明瞭に記録されやすい．一方，遠距離・音量小の条件では，収録における音に比べて空間的反響や環境ノイズの影響が相対的に大きくなるため，話者固有の特徴が劣化する傾向がある．他の 2 条件（近距離・音量小，遠距離・音量大）は，この 2 つの条件の中間的な特性を示すと考えられる．Table 1 にそれぞれの収録条件と再生機器ごとの発話

Table 1　収録条件と再生機器ごとの平均 SNR [dB] (括弧内は標準偏差)

	iPad	Mac	Bose	Sony	全体
近距離	17.8	25.21	37.9	31.9	28.2
音量大	(2.29)	(2.07)	(2.22)	(4.01)	(2.65)
近距離	7.66	8.88	19.9	14.4	12.7
音量小	(1.85)	(2.85)	(2.46)	(3.42)	(2.65)
遠距離	8.93	10.11	16.6	22.1	14.4
音量大	(1.81)	(2.15)	(2.90)	(2.82)	(2.42)
遠距離	1.75	1.82	9.11	7.81	5.12
音量小	(0.75)	(1.02)	(2.68)	(2.73)	(1.80)

平均の SNR と標準偏差を示す．Table 1 より，特に Sony は収録条件によって SNR が大きく変化しており，近距離・音量大では 31.9dB，遠距離・音量小では 7.81 dB と，24dB 以上の差がある．一方，iPad や Mac のような内蔵スピーカーでは，音量や距離を変えても比較的一貫して低い SNR を示しており，iPad では 1.75〜17.8dB，Mac では 1.82〜25.21dB の範囲に収まっている．また，標準偏差の観点から見ると，近距離条件（特に Sony の 4.01dB や Bose の 2.29dB など）では，音量条件の影響も加わって SNR のばらつきが大きくなる傾向がある．このことから，近距離では再生機器の特性や発話内容による影響を強く受け，SNR にばらつきが生じやすいことがわかる．

4　実験条件

4.1　なりすまし音声収録

J-SpAW に含まれる不正収録音声 2,100 発話 (21 話者× 25 発話× 4 不正収録機器) を用いて，3 章で示した機器や攻撃収録条件のもとでなりすまし音声攻撃の収録を行った．本実験で収集されたなりすまし音声は 67,200 発話 (不正収録音声 2,100 発話× 4 攻撃収録条件× 4 再生機器× 2 攻撃収録機器) であった．攻撃収録時のサンプリング周波数は 48kHz となっており，評価時にはダウンサンプリングにより 16kHz にしたものを用いた．

4.2　なりすまし音声検出

本実験で収録した 67,200 発話のなりすまし音声と J-SpAW に収録されている話者照合用の実発話 800 発話 (40 話者× 5 発話× 4 収録環境) を用いてなりすまし音声検出実験を行った．なりすまし音声検出モデルにはなりすまし音声検出の国際コンペティション ASVspoof2021 [1] のベースラインとして公開されている事前学習済み Linear frequency cepstral coefficients gaussian mixture model (LFCC-GMM) [6] と ASVspoof2021 の評価セットで非常に高い性能が

Table 2　LFCC-GMM における再生機器と収録条件ごとの EER_{cm}(%)

	音量大		音量小	
	近距離	遠距離	近距離	遠距離
iPad	11.74	11.74	39.50	**53.11**
Mac	23.50	21.50	56.12	**78.74**
Bose	38.00	24.76	38.73	**40.00**
Sony	15.24	15.24	36.12	**48.48**

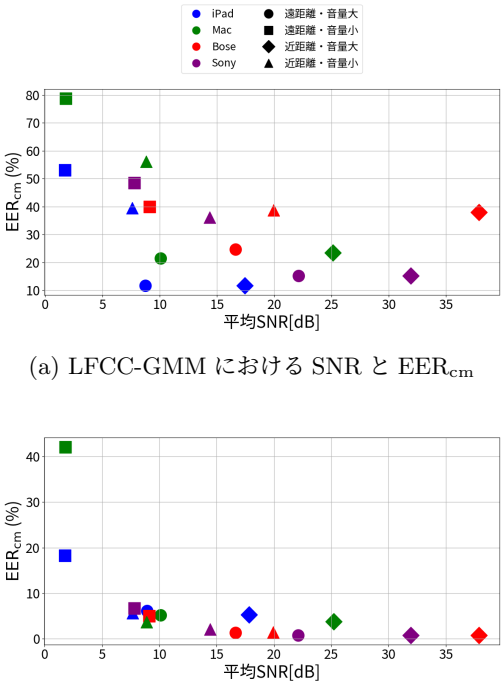
Table 3　w2v2+AASIST における再生機器と収録条件ごとの EER_{cm}(%)

	音量大		音量小	
	近距離	遠距離	近距離	遠距離
iPad	5.23	6.13	5.62	**18.24**
Mac	3.72	5.12	3.62	**42.12**
Bose	0.74	1.24	1.37	**5.00**
Sony	0.73	0.73	2.00	**6.62**

得られていることが報告されている wav2vec 2.0 と AASIST を組み合わせたモデル (w2v2+AASIST) [7] を使用した．性能評価の指標にはなりすまし音声誤受理率と実発話誤棄却率が等しくなる点であるなりすまし音声検出の等価エラー率 (Equal error rate for countermeasure; EER_{cm}) を使用した．EER_{cm} が高ければなりすまし音声検出が失敗した，つまりなりすまし音声攻撃が成功したことを表す．本研究では，なりすまし音声検出が検出困難ななりすまし音声を作ることを目標としているため，EER_{cm} が高くなることを目指す．

4.3　話者照合

なりすまし音声が話者照合システムをどの程度突破できるかを評価するため，話者照合実験も行った．まず，通常の話者照合として J-SpAW で用意されている照合ペアのトライアルを用いて話者照合評価を行い，次に，そのトライアルに登録者の実発話となりすまし音声の照合ペアを追加した話者照合実験を行った．つまり，実発話同士で話者照合を行うなりすまし音声攻撃なしの条件と，実発話となりすまし音声の照合ペアを含むなりすまし音声攻撃ありの条件両方を評価し，比較した．実発話音声の本人同士の照合ペア数が 7,600，実発話音声の他人同士の照合ペア数が 30,000，同一話者の実発話音声となりすまし音声の照合ペア数が 336,000 の合計 373,600 ペアを使用した．話者照合の評価に使用したモデルは，Voxceleb2 [8] で学習された ECAPA-TDNN [9] である．評価には本人誤棄却率と他人誤受理率が等しくなる点である，話者照合の等価エラー率 (EER for automatic speaker vertification; EER_{asv}) を用いた．EER_{asv} は通常，低いほど照合性能が高いことを示すが，本研究では，なりすまし音声と実発話のペアを「他人同士」としてトライアルに含めており，話者照合システムがこれらを本人同士と誤って判定した場合，誤受理として扱われる．そのため，EER_{asv} が高いほど，実発話音声となりすまし音声が本人と判別されたケースが多く，照合システムを突破できたことを意味しているため，本研究では高い EER_{asv} を目指す．



(a) LFCC-GMM における SNR と EER_{cm}

(b) w2v2+AASIST における SNR と EER_{cm}

Fig. 2　モデル別に見た SNR と EER_{cm} の関係

5　実験結果

5.1　なりすまし音声検出：距離と音量の影響

Table 2, 3 に LFCC-GMM および w2v2+AASIST を用いて本実験で収録したなりすまし音声の検出実験を行った結果を示す．Table 2 より，LFCC-GMM ではどの条件でも EER_{cm} が非常に高くなっており，なりすまし音声攻撃が成功していることがわかる．LFCC-GMM が ASVspoof のデータで学習されていたとしても，なりすまし音声検出の頑健性が十分に高くないことが原因の一つだと考えられる．特に，4 つの再生機器全てにおいて遠距離・音量小の条件で EER_{cm} が高い傾向にある．Table 3 の w2v2+AASIST を用いた際の結果では，LFCC-GMM と比較すると全体的に EER_{cm} が低くなっている．しかし，LFCC-GMM と同様に 4 つの再生機器全てにおいて，遠距離・音量小の条件下では特に EER_{cm} が高い傾向にある．w2v2+AASIST

Table 4　収録条件ごとの EER_{asv}(%)

なりすまし	音量大		音量小	
音声なし	近距離	遠距離	近距離	遠距離
1.69	22.59	14.08	20.41	8.33

は文献 [3] の実験結果において，どの攻撃収録環境・再生機器でも EER_{cm} が 6%以下であったため，今回作成したなりすまし音声の中では遠距離・音量小の条件の検出難度が高いといえる．また，その条件に次いで遠距離・音量大の条件の値が高い．再生機器においては，Bose と Sony より iPad と Mac が EER_{cm} の高い傾向にあることがわかる．この結果より，音質の高い Bose，Sony に比べて内臓スピーカーである iPad, Mac はなりすまし音声攻撃が成功しやすいといえる．

5.2　なりすまし音声検出：SNR の影響

次に，再生機器・攻撃収録条件ごとの平均 SNR と EER_{cm} の傾向を調べる．Figures 2(a)，2(b) はそれぞれ LFCC-GMM と w2v2+AASIST における平均 SNR と EER_{cm} の散布図である．Figures 2(a)，2(b) から，どちらも SNR が低いほど EER_{cm} が比較的高くなることがわかる．また，w2v2+AASIST は特に相関が強いことがわかる．また，w2v2+AASIST においては，Fig. 2(b) からも，SNR が同じ程度の条件において Bose や Sony よりも iPad と Mac が EER_{cm} の高い傾向にあることが読み取れる．つまり，内臓スピーカーのような音質があまり高くないスピーカーの方がなりすまし音声攻撃が成功しやすいことがわかる．

5.3　話者照合

本収録で行った 4 攻撃収録条件のそれぞれにおいて話者照合実験を行った結果の EER_{asv} を Table 4 に示す．Table 4 より，なりすまし音声なしの条件に比べて，なりすまし音声ありの 4 つの条件全てにおいて EER_{asv} が高くなった．これは，今回作成したなりすまし音声が本人と誤認識されており，話者照合システムの識別性能を低下させていることを示している．特に，話者の音響的特徴が明瞭に記録されやすい近距離・音量大の条件では EER_{asv} が 22.59%と最も高くなっており，話者照合が突破しやすい環境であることが示される．一方，遠距離・音量小の条件でも EER_{asv} が 8.33%と高くなっている．これらの結果から，作成したなりすまし音声は攻撃収録条件に影響されるが全体的になりすまし音声検出も話者照合も突破できる難しい攻撃音声になったことが確認できた．ただし，なりすまし音声検出においては，平均 SNR が下がれば EER_{cm} が上がる傾向にあったが，話者照合では逆の傾向であった．つまり，なりすまし音声検出実験で失敗する条件の収録方法では，話者照合は成功しやすくなるトレードオフの関係があるといえる．

6　おわりに

本研究では，なりすまし音声検出を困難にするデータベースの作成を行い，モデルによる変化や録音再生攻撃の収録条件が与える影響について調査した．その結果，なりすまし音声検出および話者照合を困難にするなりすまし音声攻撃の作成に成功した．さらに，再生機器，攻撃収録条件や使用したモデルによって検出性能に大きな影響があることが示された．また，距離と音量によってもかなり性能が異なることが示された．今後の課題として，なりすまし音声収録環境を増やすこと，w2v2+AASIST などの高性能で難しいなりすまし音声検出モデルを使用して高い EER_{cm} を得ること，より EER_{cm} が高い攻撃収録条件の検討など，難しい音声データセットの作成を行うことが挙げられる．さらに，話者照合においても EER_{asv} を上げ，同一話者の実発話音声となりすまし音声のペアが同一話者と判定されるような音声を収録することが課題である．

謝辞　本研究の一部は JSPS 科研費 JP24K14993, SCAT および ROIS データサイエンス共同利用共同研究拠点 (DS-JOINT) の助成 (課題番号：026RP2025) の助成を受けたものである．

参考文献

[1] Héctor Delgado et. al., “ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” 2021.

[2] Yuan Gong et. al., “ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems,” *Proc. Interspeech*, pp. 2355–2359, 2019.

[3] Sayaka Shiota et. al., “J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset,” Proc. *Interspeech*, 2025 (accepted).

[4] Kinnunen, Tomi et. al., “The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” Proc. *Interspeech*, pp. 2–6, 2017.

[5] Xin Wang et. al., “ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech,” Trans. on *Computer Speech and Language*, vol. 64, 2020.

[6] Md. Sahidullah et. al., “A comparison of features for synthetic speech detection,” Proc. *Interspeech, pp. 2087–2091*, 2019.

[7] Hemlata Tak et. al., “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” Proc. *Odyssey*, 2022.

[8] Joon Son Chung et. al., “VoxCeleb2: Deep Speaker Recognition,” *Proc. Interspeech*, 2018.

[9] Desplanques et. al., “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” *Proc. Interspeech 2020, pp. 3830–3834, 2020.*