一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

J-SPAW2: 録音再生攻撃によるなりすまし音声の収録環境を分析可能な日本語音声コーパス

堀江　涼花[†]　高道慎之介[††]　塩田さやか[†]

† 東京都立大学システムデザイン研究科
†† 慶應義塾大学理工学部

あらまし　近年，話者照合に対するなりすまし音声攻撃への対策が重要視されている．そこで本研究では，話者照合となりすまし音声検出に対する録音再生攻撃の脅威を解明するため，検出が困難な攻撃条件を分析可能な日本語音声コーパス J-SPAW2 を構築した．具体的には，既存の音声コーパスである J-SPAW の音声を再生機器，収録条件などの物理的な環境条件を変化させて再収録し，より検出が難しいなりすまし音声の作成を行った．実験の結果，収録条件が音量小・遠距離の SNR が低い条件では最先端の検出モデルでも性能が著しく低下し，逆に音量大・近距離では話者照合が容易に突破されることを確認した．さらに統合評価指標 t-DCF を用いた分析により，特定の条件下ではシステム全体が深刻な脆弱性を抱えることを明らかにし，実環境を想定した評価における本コーパスの有用性を示した．

キーワード　話者照合, なりすまし音声検出, 録音再生攻撃, 音声コーパス, 収録環境

J-SPAW2: A Japanese Speech Corpus for Analyzing Recording Conditions in Replay Attacks

Suzuka HORIE[†], Shinnosuke TAKAMICHI[††], and Sayaka SHIOTA[†]

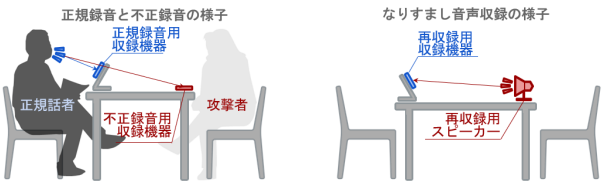
† graduate school of system design, Tokyo Metropolitan University
†† Faculty of Science and Technology, Keio University

1. ま え が き

近年，情報セキュリティの重要性が高まる中，生体認証技術が広く利用されている．生体特徴として声を用いる生体認証技術は話者照合と呼ばれ，利便性の高さから注目されている [1]．一方で，他の生体認証技術と同様に，なりすまし攻撃に対する脆弱性が指摘されており，なりすまし音声検出技術の重要性が高まっている [2],[3] ．なりすまし音声は，音声合成や声質変換などによる論理的アクセス攻撃 Logical Access（LA）と，録音再生による物理的アクセス攻撃 Physical Access（PA）に大別される．このような攻撃に対して，なりすまし音声検出および話者照合の性能評価を目的とした音声データベースが公開されてきた [4]–[6]．特に，J-SPAW [6] は，話者照合となりすまし音声検出の両方を評価可能な日本語音声データベースとして構築され，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，既存の J-SPAW に含まれるなりすまし音声は，なりすまし音声検出において比較的容易に検出可能であり，検出が困難な攻撃条件を十分に含んでいないことが報告されている [7]．

そこで本研究では，より検出が困難ななりすまし音声を体系的に収録・分析することを目的として，日本語音声コーパス J-SPAW2 を構築した(注1)．J-SPAW2 では，J-SPAW に含まれる不正録音音声を基に，複数の攻撃収録環境において再収録を行い，攻撃収録条件がなりすまし音声検出および話者照合性能に与える影響を詳細に分析可能とした．J-SPAW2 は，録音再生攻撃を対象とした PA タスクに加え，音声合成・声質変換攻撃を想定した LA タスクを含む評価用音声コーパスであるが，本論文では作成した J-SPAW2 における PA タスクを対象として分析および評価を行っている．具体的には，再生機器と攻撃収録機器の距離条件や信号対雑音比 Signal-to-Noise Ratio（SNR）に着目し，これらを系統的に変化させた録音再生攻撃音声を収録している．さらに，構築した J-SPAW2 を用いた実験を通して，検出がより困難ななりすまし音声条件を明らかにするとともに，話者照合においても実発話のみの場合と比較して評価性能が悪化することを確認した．これにより，話者照合を突破し得る，より現実的かつ高難度ななりすまし音声を含むデータベースを構

(注1)：https://github.com/takamichi-lab/j-spaw2



(a) 実発話収録　　(b) なりすまし音声収録

図 1: (a) 正規録音と不正録音，(b) 録音再生攻撃のイメージ図（不正録音で得られた音声を録音再生攻撃に用いる）

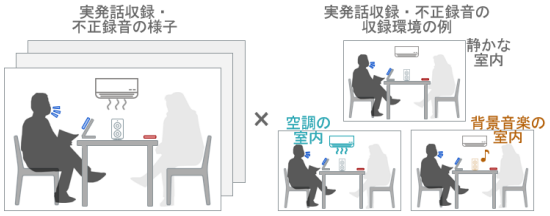


図 2: 実発話の収録条件の組み合わせ

築できたといえる．また，なりすまし音声検出と話者照合を直列に用いた統合認証システムを想定した評価指標である t-DCF を用いることで実運用環境における J-SPAW2 の録音再生攻撃の脅威を分析した結果を報告する．

2. 関 連 研 究

2.1 録音再生攻撃に対するなりすまし音声検出とその課題

録音再生攻撃によるなりすまし音声とは，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を用いて再生することで，話者照合システムを欺く攻撃手法である．なりすまし音声検出は，入力音声が実際に人間が発声した実発話であるか，あるいは録音再生や音声合成などによって生成されたなりすまし音声であるかを識別するタスクである．これまでの研究において，録音再生によるなりすまし音声検出の性能は，不正録音時の収録環境や，再生攻撃時の環境条件の影響を大きく受けることが報告されている [8]–[10]．特に，使用されるマイクやスピーカーの特性，話者とマイクの距離，収録空間の残響特性などが検出性能に影響を与えることが知られており，これらの要因が検出器の頑健性を左右する重要な要素であるといえる．

このような背景から，録音再生攻撃に対するなりすまし音声検出の頑健性向上を目的として，様々な条件下で収録された音声を含む評価用データベースの整備が進められてきた．しかし，既存のなりすまし音声検出用コーパスの多くは，音声合成や声質変換による論理的攻撃を主な対象としており，録音再生による攻撃を体系的に扱ったものは限られている．

2.2 J-SPAW

J-SPAW は，日本語音声を用いた話者照合評価セットとなりすまし音声検出評価セットから構成されるコーパスである．話者照合評価セットにおいては図 1(a) に示すように，正規ユーザによる実発話音声の収録に加え，攻撃者がやや離れた位置から不正録音する状況を想定した収録が行われている．さらに，なりすまし音声検出評価セットでは，図 1(b) に示すように，不正録音された音声を再生機器で再生し，再収録することで録音再生攻撃音声が作成されている．J-SPAW では，不正録音のマイク位置が話者の口元から離れているなど，実際の攻撃シナリオを意識した現実的な録音再生攻撃を想定している点に特徴がある．一方で，公開されている評価結果では，比較的最先端なりすまし音声検出モデルを用いた場合になりすまし音声検出率が高く，録音再生攻撃に対する検出タスクとしては難易度が十分でないという課題もあった．また，録音再生攻撃の種類が少ないため収録環境や再生条件が検出性能に与える影響を分析することは難しく，環境要因の影響を体系的に評価するためのデータとしては限定的である．さらに，収録距離や再生音量といった個々の収録条件を比較することを目的とした設計にはなっていないため，どのような収録条件がどの程度影響を及ぼしているかを分析することは困難である．

3. J-SPAW2

3.1 J-SPAW との共通点

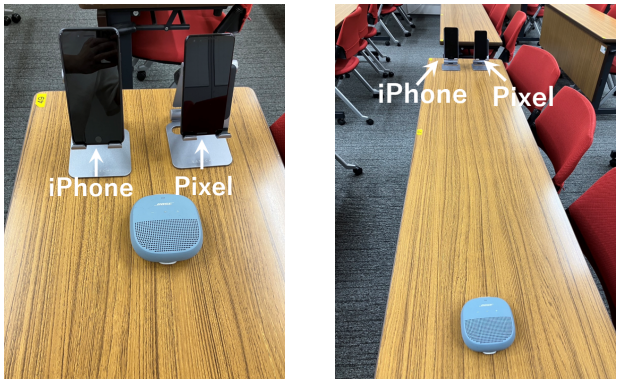
J-SPAW2 は，既存の日本語音声コーパスである J-SPAW の設計方針を踏まえ，話者照合およびなりすまし音声検出の評価を目的として構築されている．J-SPAW2 では，J-SPAW に収録された正規話者の実発話音声をそのまま使用することを前提とし，なりすまし音声（PA および LA）のみを新たに再収録・再構成している．したがって，実発話音声に関する話者構成，発話内容，および収録環境の設計は J-SPAW と共通である．実発話音声は，静かな室内，空調が動作している室内，背景音楽が流れている室内，および屋外の 4 種類の収録環境において収録されている（図 2）．収録機器は一般的な個人利用端末を想定し，Pixel3 (”Pixel”)，iPhone8 Plus (”iPhone”)，iPad mini (第 5 世代) (”iPad”) の 3 種類を使用している．このうち Pixel および iPhone は正規話者の実発話音声を収録するための正規収録機器として，iPad は攻撃者が不正に音声を録音する状況を想定した不正録音機器として用いている．J-SPAW に含まれる実発話の収録条件を表 1 に示す．各話者は 50 文の定型文を収録環境ごとに発話しており，各発話の長さは 1-5 秒程度である．収録話者は男性 21 名，女性 19 名の計 40 名で構成され，実発話の合計発話数は 24,000 発話（50 発話 × 40 話者 × 4 収録環境 × 3 収録機器）である．

3.2 収録環境を分析可能とするなりすまし音声攻撃

J-SPAW2 は，J-SPAW と同一の話者および定型文を用いながら，録音再生攻撃における再生・収録条件の違いがなりすまし音声検出性能に与える影響をより詳細に分析することを目的として新たに構築したコーパスである．J-SPAW2 において実発話収録時に iPad により不正に録音することを想定して収録した音声を再生音源として用い，再生機器から攻撃収録機器へ再生・再収録することで，新たな録音再生攻撃によるなりすまし音声を生成している．J-SPAW では，各話者 50 発話のうち 5 発話のみを評価用に用い，残りの 45 発話は攻撃者が入手可能であることを想定している．J-SPAW2 のなりすまし音声の生成には，iPad により不正に録音した想定の実発話音声である 45 発話の

表 1: 実発話の収録条件

ラベル	
正規録音用マイク	
M1	Google Pixel 3（話者からの距離：1.0m）
M2	Apple iPhone 8（上記と同一位置）
不正録音用マイク	
M3	Apple iPad mini（第 5 世代）
正規／不正録音環境（録音部屋）	
R1	室内 1（4.4(W) × 7.4(L) × 2.5(H) [m]）
R2	研究機関 1 の道路近接の屋外環境
R3	室内 2（10.8(W) × 2.0(L) × 2.8(H) [m]）
R4	研究機関 2 の芝生の屋外環境
録音時の環境条件	
E1	静かな環境（R1, R3）
E2	空調稼働環境（R1, R3）
E3	スピーカから音楽を再生する環境（R1, R3）
E4	屋外環境（R2, R4）



(a) 近距離　　(b) 遠距離

図 3: 近距離・遠距離での収録状況 (Bose)

中から 25 発話を用いた．不正録音した音声の再生環境はすべて 静かな室内環境で実施し，空間条件を固定したうえで，再生機器や再生条件の違いに着目した．再生機器は J-SPAW と同じ iPad，Mac，Bose，Sony の 4 種類を使用している．収録時のサンプリング周波数は 48 kHz とし，評価時には 16 kHz にダウンサンプリングして使用している．攻撃収録機器には，実発話音声の収録で使用した Pixel および iPhone の 2 種類を使用している．本研究で作成した録音再生攻撃音声の条件を 表 2 に示す．J-SPAW との大きな違いとして，再生機器と攻撃収録機器の距離（近距離・遠距離）および再生音量（音量大・音量小）を組み合わせた 4 種類の再生条件を設定し，すべての組み合わせで録音再生攻撃音声を収録した．J-SPAW2 に含まれるなりすまし音声は 128,000 発話（40 話者× 25 発話× 4 実発話収録環境× 4 再生機器× 2 攻撃収録機器× 4 再生条件）である再生機器と攻撃収録機器の距離は，近距離では約 10 cm，遠距離では約 1 m とした．図 3(a) に近距離での収録状況，図 3(b) に遠距離における Bose を用いたなりすまし音声攻撃の収録状況を示す．これらの距離条件と再生音量（音量大・音量小）を組み合わせた 4 種類の収録条件は，音声信号の強度および環境ノイズの影響が異なる状況を想定して設定したものである．

表 2: なりすまし音声作成時の収録条件

ラベル	
再生環境（再生部屋）	
r1	室内 3（11.0(W) × 8.0(L) × 2.6(H) [m]）
再生用スピーカ	
s1	Bose Soundlink Micro Bluetooth Speaker Bundle
s2	iPad
s3	MacBook Pro
s4	Sony SRS-ZR7
再生時の環境条件	
e1, e2, e3	それぞれ E1, E2, E3 に対応．e3 と E3 では使用楽曲が異なる
再生条件（距離）	
p1	近距離（約 10cm）
p2	遠距離（約 1m）
再生条件（音量）	
v1	音量大
v2	音量小
再収録用マイク	
m1, m2	それぞれ M1, M2 に対応

4. 実 験 条 件

4.1 なりすまし音声検出

J-SPAW2 に含まれる 128,000 発話のなりすまし音声と J-SPAW に収録されている実発話のうち 800 発話 (40 話者× 5 発話× 4 収録環境) を用いて再生機器と再生条件ごとに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 [4] のベースラインとして公開されている事前学習済みモデル Linear frequency cepstral coefficients gaussian mixture model (LFCC-GMM) [11], [12] と ASVspoof5 のベースラインとして公開されている事前学習済みモデル AASIST [13], [14]，文献 [15] で発表されているなりすまし音声研修の最先端モデルの 1 つであり，ASVspoof2021 の評価セットで非常に高い性能が得られていることが報告されている wav2vec2.0+AASIST (w2v2+AASIST) の合計 3 つのモデルを使用した．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 （$\mathrm{EER}_{\mathrm{cm}}$）を使用した．$\mathrm{EER}_{\mathrm{cm}}$ 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，$\mathrm{EER}_{\mathrm{cm}}$ が高くなることを目指す．

4.2 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．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

表 3: 収録条件と再生機器ごとの平均 SNR (dB)（カッコ内は標準偏差）**太字**は各機器における最大値を，下線は最小値を表す

音量	距離	iPad	Mac	Bose	Sony	全体
音量大	近距離	**17.24**	**25.58**	**31.13**	**30.39**	**26.09**
		(4.50)	(6.68)	(54.87)	(15.19)	(20.31)
	遠距離	8.78	11.85	17.84	22.85	15.33
		(3.24)	(7.43)	(8.48)	(7.76)	(6.73)
音量小	近距離	5.82	7.30	16.43	13.27	10.71
		(5.93)	(8.09)	(19.11)	(10.79)	(10.98)
	遠距離	2.68	1.88	6.78	5.50	4.21
		(1.92)	(0.96)	(10.70)	(10.53)	(6.03)

し音声の照合ペア数が 632,800 の合計 670,400 ペアを使用した．話者照合の評価に使用したモデルは，Voxceleb2 [16] で学習された Emphasized channel attention, propagation and aggregation in time delay neural network（ECAPA-TDNN）[17] 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．

4.3 統 合 評 価

さらに，なりすまし音声検出と話者照合を組み合わせたシステム全体の性能を評価するため，tandem detection cost function (t-DCF) [4], [18], [19] を用いた評価も行った．t-DCF は，話者照合システムの前段に配置されたなりすまし音声検出器の誤りが，話者照合における認証結果に与える影響を考慮した評価指標であり，なりすまし音声検出と話者照合を直列に用いた実運用に近い性能評価を可能にする．t-DCF の値は 0 から 1 の範囲を取り，値が小さいほどなりすまし音声検出と話者照合を結合したシステム全体の性能が高いことを示す．本研究では，検出困難ななりすまし音声を作ることを目標としており，なりすまし音声検出および話者照合の双方を突破する状況を重視するため，EER と同様に t-DCF も高くなることを目指す．

5. 実 験 結 果

5.1 再生条件および再生機器による SNR 特性の分析

表 3 に，再生機器および収録条件ごとの発話平均 SNR と標準偏差を示す．全体として，音量大・近距離条件では音圧レベルが高く，話者の音響的特徴が比較的明瞭に記録されやすいため，高い SNR が得られる．音量小・遠距離条件では音声信号に対して空間的反響や環境雑音の影響が相対的に大きくなり，話者固有の特徴が劣化する傾向があるため，すべての再生機器に共通して SNR が低下することが確認される．他の 2 条件（音量小・近距離，音量大・遠距離）は，これら 2 条件の中間的な特性を示している．特に，音量大・近距離条件における平均 SNR は，Bose で 31.13 dB，Sony で 30.39 dB と高く，音量小・遠距離条件ではそれぞれ 6.78 dB，5.50 dB まで低下しており，再生条件の違いが音声の信号品質に大きな影響を与えていることが分かる．一方，iPad や Mac の内蔵スピーカーでは，音量大・近距離条件においても平均 SNR がそれぞれ 17.24 dB，25.58 dB にとどまり，全条件を通して比較的低い SNR を示す傾向が確認された．これは，専用スピーカーと比較して再生能力の性能に制限があることが影響していると考えられる．

標準偏差に着目すると，音量大・近距離条件において Bose で 54.87 dB と極めて大きなばらつきが観測されており，再生機器の音響特性や発話内容，および再収録環境との作用が SNR に強く影響している可能性が示唆される．一方，音量小・遠距離条件では標準偏差が比較的小さく，全体として一様に低品質な音声となる傾向が見られる．これらの結果から，J-SPAW2 には，高 SNR から低 SNR まで幅広い音響条件下で収録された録音再生攻撃音声が体系的に含まれており，実環境における多様ななりすまし音声の検出性能を評価・分析するための有用なコーパスであることが確認できる．

表 4: LFCC-GMM の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	35.00	12.50	50.62	**63.27**
Mac	31.13	17.90	66.12	**82.99**
Bose	50.50	27.25	53.77	**58.64**
Sony	18.50	17.49	48.10	**63.12**
All	34.50	19.25	54.75	**66.38**

表 5: AASIST の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	31.76	36.88	44.25	**49.28**
Mac	31.15	28.34	42.00	**53.87**
Bose	**42.47**	18.25	21.00	36.12
Sony	**51.38**	29.40	20.88	42.88
All	40.10	28.75	35.12	**46.12**

表 6: w2v2+AASIST の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	3.99	4.74	5.12	**14.23**
Mac	2.88	3.41	3.38	**37.40**
Bose	0.62	1.00	1.12	**5.12**
Sony	0.50	0.50	2.73	**14.36**
All	2.53	2.62	3.60	**18.62**

5.2 なりすまし音声検出

表 4, 5, 6 に，LFCC-GMM，AASIST，および w2v2+AASIST を用いて，J-SPAW2 に収録されているなりすまし音声に対する検出実験を行った結果を示す．

表 4 に示す LFCC-GMM の結果を見ると，すべての再生機器および収録条件において EER_{cm} が 15% 以上と高い値を示して

表 7: 収録条件ごとの EER_{asv}(%) ↑

なりすまし音声	音量大		音量小	
なし	近距離	遠距離	近距離	遠距離
1.69	**23.91**	15.28	20.54	9.79

おり，なりすまし音声攻撃が広範な条件下で成立していることが分かる．特に，4 種類すべての再生機器において，音量小・遠距離条件で EER_{cm} が最大となっており，Mac では 82.99%，iPad でも 63.27% に達している．これらの結果は，LFCC-GMM が ASVspoof データセットで学習されていたとしても，実環境に近い録音再生攻撃条件に対して十分な頑健性を持たないことを示している．

次に，表 5 に示す AASIST の結果を見ると，LFCC-GMM と同様に，多くの再生機器および収録条件において EER_{cm} が依然として高い値を示しており，深層学習ベース手法であっても録音再生攻撃に対して十分な頑健性を有していないことが分かる．特に，音量小・遠距離条件では EER_{cm} が高くなる傾向が共通して見られ，iPad および Mac ではそれぞれ 49.28%，53.87% に達している．一方で，Bose や Sony では音量大・近距離条件において EER_{cm} が相対的に高くなる場合も確認された．これらの結果から，再生機器および検出モデルの特性に応じて，検出性能が劣化する収録条件が異なることがわかる．

さらに，表 6 に示す w2v2+AASIST の結果では，LFCC-GMM や AASIST と比べて EER_{cm} が大幅に低下しており，全体として高い検出性能を有することが分かる．音量大・近距離条件では，すべての再生機器で EER_{cm} が 5% 未満に抑えられている．しかし，LFCC-GMM や AASIST と同様に，音量小・遠距離条件では検出性能が著しく低下し，Mac では 37.40%，iPad や Sony でも 14% 程度の高い EER_{cm} が観測される．一方で，既存の日本語音声コーパス J-SPAW を用いた先行研究として，文献 [6] では，J-SPAW2 とは発話数や録音再生条件が異なるものの，録音再生攻撃に対する w2v2+AASIST の検出性能が評価されている．同文献では，すべての攻撃収録環境および再生機器において EER_{cm} が 6% 以下であったと報告されている．これと比較すると，本研究で作成したなりすまし音声，特に音量小・遠距離条件の音声は，w2v2+AASIST のような最新のなりすまし音声検出手法に対しても検出困難であり，攻撃として高い有効性を持つことが分かる．また，再生機器の違いに着目すると，Bose や Sony といった専用スピーカーに比べて，iPad や Mac の内蔵スピーカーでは EER_{cm} が比較的高くなる傾向が一貫して見られた．これは，内蔵スピーカーによる再生音声が環境雑音や反射音の影響を受けやすく，なりすまし音声検出を困難にしているためだと考えられる．

5.3 話者照合

本収録で行った 4 種類の攻撃収録条件において話者照合実験を行った結果の EER_{asv} を 表 7 に示す．比較のため，既存の J-SPAW の実発話のみを用いた話者照合結果（なりすまし音声なし）も合わせて示す．表 7 より，なりすまし音声なしの条件では EER_{asv} が 1.69% と低く，ECAPA-TDNN による話者照合が高い性能を示していることが確認できる．一方，なりすまし音声を含む 4 条件すべてにおいて EER_{asv} が大きく上昇しており，今回作成したなりすまし音声が本人音声として誤認識され，話者照合システムの識別性能を低下させていることが分かる．特に，平均 SNR の高い音量大・近距離条件では EER_{asv} が 23.91% と最も高くなっており，話者の音響的特徴が比較的明瞭に再収録される環境において，なりすまし音声が話者照合システムを突破しやすいことが示された．これは，音声の劣化が比較的少ない条件において，話者の音響的な特徴が十分に保持されることで，登録された本人音声との類似度が高くなりやすいためだと考えられる．また，音量小・遠距離条件のように，空間的反響や環境雑音の影響を強く受ける収録条件においても，EER_{asv} は 9.79% と高い値を示している．この結果は，音響的特徴が不明瞭になる条件であっても，なりすまし音声が一定の確率で本人と誤判定されることを意味しており，話者照合システムに対する攻撃の有効性が幅広い収録条件において維持されていることを示している．以上より，J-SPAW2 に含まれるなりすまし音声は，攻撃収録条件によって話者照合性能への影響の度合いは異なるものの，全体として話者照合システムを突破可能な難易度の高い攻撃音声となっていることが確認できた．

表 8: モデルと収録条件ごとの t-DCF ↑

	音量大		音量小	
	近距離	遠距離	近距離	遠距離
LFCC-GMM	0.755	0.682	**0.983**	0.975
AASIST	0.905	0.939	0.976	**1.000**
w2v2+AASIST	0.103	0.095	0.160	**0.619**

これまでの結果より，なりすまし音声検出実験においては，なりすまし音声の平均 SNR が低下するにつれて EER_{cm} が上昇する傾向が確認されたのに対し，話者照合では必ずしも同様の傾向は見られなかった．つまり，なりすまし音声検出において検出が困難となる低 SNR 条件では，話者照合は比較的成功しやすくなる場合があり，両者の間にはトレードオフの関係が存在すると考えられる．この結果は，両者を同時に突破する攻撃音声を生成・評価することの難しさと重要性を改めて示す結果である．

5.4 統合評価

表 8 に，なりすまし音声検出として LFCC-GMM，AASIST，w2v2+AASIST を，話者照合として ECAPA-TDNN を用いた場合の t-DCF の結果を示す．t-DCF は，なりすまし音声検出と話者照合を直列に用いた際のシステム全体の性能を表す指標であり，値が大きいほど両システムを同時に突破できていることを意味する．

LFCC-GMM を用いた場合，すべての再生条件において t-DCF が高い値を示しており，特に音量小・近距離および音量小・遠距離の条件ではそれぞれ 0.983，0.975 と 1 に近い値となった．これは，LFCC-GMM においては，音量小・遠距離のようななりすまし音声検出が困難な条件で，話者照合を含めたシステム全体としても攻撃を十分に防げていないことを示している．一方，AASIST を用いた場合においても，全体的に t-DCF は高い値を示す傾向が確認され，特に音量小条件では 0.976〜1.000

と極めて高い値となった．この結果は，なりすまし音声検出として一定の性能を有するモデルであっても，話者照合と組み合わせた際には，攻撃収録条件によってシステム全体の脆弱性が顕在化する可能性があることを示している．これに対し，w2v2+AASIST を用いた場合，音量大・近距離および音量大・遠距離の条件では t-DCF がそれぞれ 0.103，0.095 と低い値に抑えられており，高性能ななりすまし音声検出器が統合システム全体の防御性能を大きく向上させていることがわかる．しかし，音量小・遠距離の条件では t-DCF が 0.619 と大きく上昇しており，高性能な検出器を用いた場合であっても，録音再生条件によっては，なりすまし音声検出と話者照合の双方を突破可能であることが示された．w2v2+AASIST において既存の J-SPAW では t-DCF が 0.096 となっており，音量大・遠距離以外の収録条件においては J-SPAW2 コーパスの検出困難性を十分に示すことができる．

これらの傾向は，SNR および話者照合実験における $\mathrm{EER}_{\mathrm{asv}}$ の結果とも整合的である．音響的劣化が大きい条件では，なりすまし音声検出または話者照合のいずれか一方，あるいは両方の性能が低下し，その結果として t-DCF が大きく悪化する傾向が確認された．特に，音量小・遠距離条件のように SNR が低下する環境では，検出性能と照合性能のバランスが崩れやすく，統合システム全体が攻撃に対して脆弱になる可能性が高いと考えられる．以上の結果から，なりすまし音声検出単体で検出が困難な条件は，話者照合を含めた統合的な認証システムにおいても，脆弱性を生じさせる可能性が高いことが明らかとなった．つまり，生体認証システムの安全性を評価する際には，個別の性能指標に加えて，t-DCF のような統合的指標に基づく評価が不可欠であることを示している．また，本研究で作成したなりすまし音声は，実運用を想定した統合認証システムに対しても高い攻撃成功率を有しており，録音再生攻撃における収録条件の違いを考慮した評価の重要性を改めて示す結果となった．

6. おわりに

本研究では，実環境における録音再生攻撃条件が，なりすまし音声検出および話者照合性能に与える影響を明らかにすることを目的とし，J-SPAW2 を構築した．そのために，複数の再生機器，収録距離，音量条件を組み合わせた録音再生攻撃音声を新たに収録し，既存のなりすまし音声検出モデルおよび話者照合モデルを用いて評価を行った．実験の結果，全てのモデルにおいて収録条件によってはなりすまし音声検出性能が大きく低下する傾向が確認された．特に w2v2+AASIST は既存研究において高い検出性能が報告されているにもかかわらず，本研究で収録した一部の条件下では $\mathrm{EER}_{\mathrm{cm}}$ や t-DCF が大きく悪化しており，実環境を想定した録音再生攻撃の難しさを示す結果となった．一方で話者照合実験では，音量大・近距離の条件において $\mathrm{EER}_{\mathrm{asv}}$ が最も高く，話者の音響的特徴が比較的明瞭に再収録される条件では，話者照合システムが突破されやすいことが示された．今後の課題としては，より多様な室内音響特性や雑音環境を含む収録条件の拡張に加え，異なる話者照合モデルや最新のなりすまし音声検出手法を用いた包括的な評価が挙げられる．また，本研究で収録した録音再生攻撃音声を学習データとして活用することで，実環境における攻撃条件に対してより頑健ななりすまし音声検出モデルの構築が可能になると考えられる．

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．

文　献

[1] Zhang Rui et.al., “A survey on biometric authentication: Toward secure and privacy-preserving identification,” *Transaction on IEEE Access*, vol. 7, pp. 5994–6009, 2019.
[2] 俵直弘, “話者認識システムとなりすまし対策,” 日本音響学会誌, 78 巻 6 号, pp.338-346, 2022.
[3] https://www.asvspoof.org/.
[4] Héctor Delgado et.al., “ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” *arXiv preprint arXiv:2109.00535*, 2021.
[5] Jee-weon Jung et. al., “Spoofceleb: Speech deepfake detection and sasv in the wild,” *IEEE OPEN JOURNAL OF SIGNAL PROCESSING*, vol. 6, pp. 68–77, 2025.
[6] Sayaka Shiota et. al., “J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset,” Proc. *Interspeech*, pp. 3913–3917, 2025.
[7] 堀江 涼花 et.al., “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,” Proc. 日本音響学会秋季研究発表会, 2025.
[8] Kinnunen et. al., “The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” Proc. *Interspeech*, pp. 2–6, 2017.
[9] 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.
[10] Lantian Li et.al., “A study on replay attack and anti-spoofing for automatic speaker verification,” Proc. *Interspeech*, 2017.
[11] Sahidullah et. al., “A comparison of features for synthetic speech detection,” Proc. *Interspeech*, pp. 2087–2091, 2019.
[12] Hemlata Tak et. al,, “Spoofing attack detection using the non-linear fusion of sub-band classifiers,” Proc. *Interspeech*, pp. 1106–1110, 2020.
[13] Jee-weon Junget. al., “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371, IEEE, 2022.
[14] Xin Wang et.al., “ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale,” in *Proc. ASVspoof 2024*, pp. 1–8, 2024.
[15] Hemlata Tak et. al., “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” Proc. *Odyssey*, 2022.
[16] Joon Son Chung et. al., “VoxCeleb2: Deep Speaker Recognition,” *Proc. Interspeech*, pp. 1086–1090, 2018.
[17] Desplanques et. al., “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” *Proc. Interspeech, pp. 3830–3834, 2020*.
[18] Tomi Kinnunen et al., “t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification,” in *Speaker Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.
[19] Tomi Kinnunen et.al., “Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.