J-SpAW:話者照合となりすまし音声検出のための日本語音声コーパス

菅野 滉大[1,a)]　高道 慎之介[2,3,b)]　塩田 さやか[1,c)]

概要：本論文では，話者照合となりすまし音声検出のための日本語音声コーパス Japanese Spoofing Attack, recorded in-the-Wild (J-SpAW) の概要について紹介する．声の生体認証技術である話者照合は深層学習の発展に伴い大幅な性能改善がなされた．一方で，録音再生技術の簡易化や音声合成技術の発展にともない，話者照合へのなりすまし音声攻撃に対する対策についても急務となってきている．特に深層学習に基づく手法が研究の主流となってきている状況では，研究促進に必要となるのは多様な音声コーパスであるといえる．しかしながら，話者照合となりすまし音声検出のための音声コーパスはなりすまし音声検出のコンペティションである ASVspoof challenge で公開されているものがほとんどであり，収録状況や言語の多様性などが不十分であった．そこで本研究では，日本語話者による話者照合およびなりすまし音声検出の評価が可能となる音声コーパス J-SpAW を構築し，音声コーパスとしての性能を調査した．実験では J-SpAW を用いて話者照合を行い，話者照合としては高い性能が得られることを確認した．また，J-SpAW を用いてなりすまし音声を作成し，なりすまし音声検出の様々なベースラインモデルでなりすまし音声検出の評価を行ったところ，なりすまし音声検出の検出性能が非常に低く，既存のモデルでは話者照合と組み合わせたときにもなりすまし音声検出の精度はまだ不十分であることが確認された．これらの結果から，J-SpAW は，話者照合となりすまし音声攻撃の検出両方を同時に評価できる新たな音声コーパスとして使用可能であることを報告する．

キーワード：話者照合，なりすまし音声検出，音声コーパス，J-SpAW

1. はじめに

近年，電子決済やクラウドサービスの普及によりセキュリティの確保がますます重要となるなか，顔画像や指紋を用いた生体認証はユーザー認証の手段として広く採用されてきている．生体認証には様々な生体パターンを鍵とする手法が存在するが，その中でも音声を用いてユーザー認証を行う技術が話者照合 (Automatic Speaker Verification; ASV) である．話者照合はマイクのみで音声データを入手できるため実用化のハードルが低く，また携帯電話などの通信を経由することに関しても親和性が高いことから需要が高まっている．

深層学習は様々な分野に導入され目覚ましい発展を遂げている．話者照合に関しても，x-vector[1] や ECAPA-TDNN（Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network）[2], RawNet3[3] に代表される話者埋め込みネットワークに基づく手法が提案され，それぞれ非常に高い性能が報告されている．しかし話者照合技術を産業展開するにはいくつか問題があり，その一つになりすまし音声攻撃への対策がある．なりすまし音声攻撃とは事前に収録した登録話者の音声や，登録話者の音声データを用いて音声合成や声質変換技術により合成した音声を話者照合に入力することで，登録話者になりすまして話者照合の突破を図る攻撃のことである．なりすまし音声攻撃は話者照合の実用化に向けて考慮しなければならない重要課題であり，なりすまし音声攻撃を検出するためのなりすまし音声検出という技術が活発に研究されている [4]．2015 年よりなりすまし音声検出の研究のために ASVspoof challenge（以下 ASVspoof）というなりすまし音声検出技術の精度を競うコンペティションが開催されている [5]．なりすまし音声攻撃への対策を講じる上で，話者照合のなりすまし音声攻撃に対する耐性を調べ，頑健性を改善していくことは重要である．しかしながら現在公開されているなりすまし音声攻撃に対する対策用の音声コー

[1] 東京都立大学 システムデザイン研究科
[2] 慶應義塾大学 理工学部
[3] 東京大学 情報理工学系研究科
[a)] kanno-kouta@ed.tmu.ac.jp
[b)] shinnosuke_takamichi@keio.jp
[c)] sayaka@tmu.ac.jp

パスの多様性が少ないという問題がある．ASVspoof から公開された英語圏の音声コーパスは存在するが，日本語がメインとなっているなりすまし対策用の音声コーパスは存在しない．システムのなりすまし対策の汎用性を評価するためには，様々なドメインの音声コーパスが必要であり，様々な評価コーパスを公開することが分野の適切な成熟に重要となっている．

そこで，本研究では話者照合及びなりすまし音声攻撃への対策法を評価することができる日本語音声コーパス J-SpAW (Japanese Spoofing Attack, recorded in-the-Wild)[*1]を作成した．本コーパスは，話者照合の評価及びなりすまし音声検出の評価，さらに両方のシステムの性能をあわせて評価することが可能となるよう設計されている．また攻撃者が使用できる音声の収録方法についても，実際に正規ユーザが話者照合を利用するシーンやなりすまし音声を作成する攻撃者の実際の立ち位置などを考慮している．具体的には，盗聴が行われる際の収録機器の位置を再現するために，発話者が話しかける収録機器と盗聴用の収録機器の位置を離して収録をおこない，また収録環境も実際の環境を想定した 4 種類に分けて収録を行った．なりすまし音声攻撃の作成においては ASVspoof で想定されているなりすまし音声攻撃のタスクである論理的アクセス (Logical access; LA) と物理的アクセス (Physical access; PA) それぞれを評価できるように両方のタスクに使用可能なデータや音素バランスなどを考慮している．実験では，作成した音声コーパスの話者照合タスクとしての性能となりすまし音声検出タスクとしての性能を調査した．話者照合の評価には最先端手法である ECAPA-TDNN および ResNetSE34V2[6]，RawNet3 を用いた．また，なりすまし音声検出としては LA タスク及び PA タスクそれぞれ，ASVspoof challenge 2021 でベースラインとして用意されている LFCC-LCNN[7], RawNet2[8], LFCC-GMM[9][10], CQCC-GMM[11][12] の 4 つを用いた評価を行った．実験結果から，作成した音声コーパス J-SpAW は話者照合モデルの評価にも利用でき，かつなりすまし音声検出の LA タスクと PA タスクの耐性評価にも利用できることを報告する．

2. 関連研究

2.1 話者照合

話者照合は入力された音声が事前に登録された話者本人の音声であるか，そうでないかを判断する二値分類タスクである．近年主流となっている深層学習による話者埋め込みネットワークを用いた話者照合のフローを図 1 に示す．話者照合は，登録部と照合部の 2 つのフローと，両方で利用される話者埋め込み抽出モデルの学習部で構成されている．話者照合システム構築の際にはまず大量の話者を含む学習データを用いて話者埋め込み抽出モデルを学習する．登録部では，学習済みの話者埋め込み抽出モデルに登録音声を入力することで登録音声に対する話者埋め込みベクトルを抽出する．照合部においても同様に照合音声に対する話者埋め込みベクトルを抽出し，登録音声および照合音声のそれぞれから抽出された話者埋め込みベクトルの類似度を計算する．判定では算出された類似度が閾値以上であれば本人，閾値以下であれば他人であると判定する．

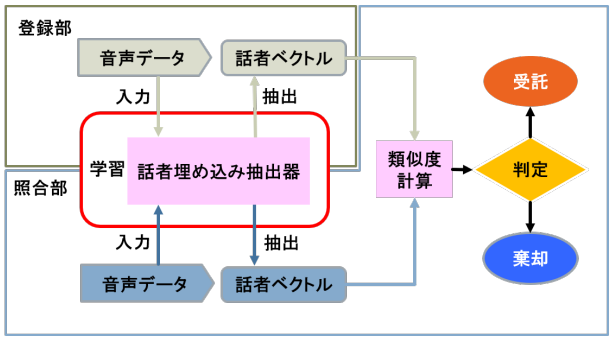


図 1: 話者照合のフロー図

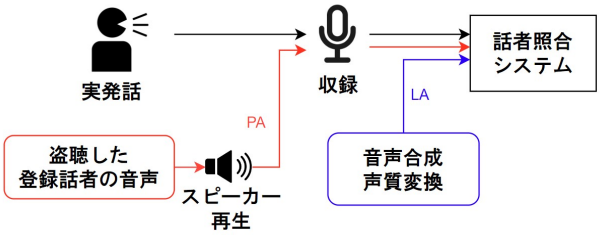


図 2: なりすまし音声攻撃の種別

2.2 なりすまし音声検出

なりすまし音声攻撃とは攻撃者が別のユーザーに成りすまして話者照合を突破しようとする攻撃である．図 2 に ASVspoof で想定しているなりすまし音声攻撃の流れを示す．ASVspoof ではなりすまし音声攻撃として，LA，PA の二つのタスクを定義している．LA タスクは，収録機器を介さず話者照合システムに直接なりすまし音声を入力する方法で，音声合成・声質変換で生成した合成音声による攻撃を想定している．PA タスクは，事前に実発話を盗聴で録音した音声 (盗聴音声) を再生機器から再生し，ユーザーが生声を入力する手順と同様に，システムの収録機器を通じて入力する方法である．なりすまし音声攻撃に対する対策法（Countermesure; CM）として様々ななりすまし音声検出法が提案されているが [13]，多くの手法が ASVspoof で公開されたデータを用いた機械学習を行い，閾値による判定を行っている [14].

[*1] https://github.com/takamichi-lab/j-spaw

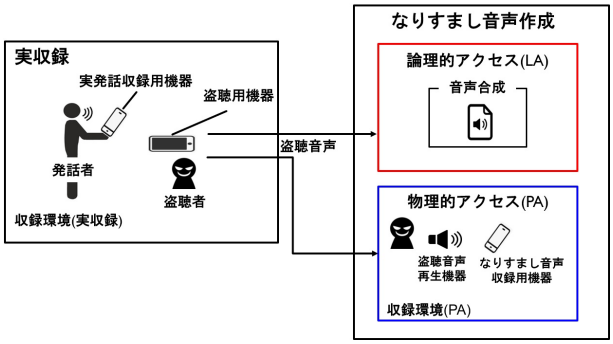


図 3: J-SpAW の話者照合用評価セットとなりすましなりすまし音声検出用評価セットの作成の流れ

3. 音声コーパス J-SpAW

話者照合となりすまし音声攻撃に関する研究は世界中で活発に行われているがコンペティション主体で研究が行われていることや，なりすまし音声を作成するコストが高いことから収録環境やドメインなどが異なる音声コーパスがあまり存在していない．そこで本章では，分野の多様性を促進するために話者照合及びなりすまし音声検出の評価を統一的に行うことができる音声コーパス J-SpAW の内容を説明する．J-SpAW は話者照合用評価セットとなりすまし音声検出用評価セットの 2 つで構成されている．各評価セットについて詳細を述べる．

3.1 話者照合用評価セット

J-SpAW コーパスの話者照合評価セットは，実環境での話者照合の使用を想定して静かな室内，空調が動作している室内，背景音楽が流れている室内，及び室外の 4 つの収録環境で収録を行った．収録機器には一般的な使用シーンで想定されるスマートフォンを用いた．発話内容はウェイクアップワードを含む短い命令文 25 種と音素バランスを考慮した定型文 25 種を用意し，一人あたり 50 文の定型文を収録環境ごとに収録する．各発話の発話長は 2 秒程度となっている．収録の流れは図 3 の実収録の部分に該当する．話者照合の評価セットは実収録で収録された音声を話者照合における登録音声と照合音声に分けて作成した．収録話者は男性 21 名女性 19 名であり，収録のサンプリング周波数は 48kHz，話者照合の評価時には 16kHz にダウンサンプリングしている．

3.2 なりすまし音声検出用評価セット

J-SpAW コーパスのなりすまし音声検出評価セットは，LA タスクおよび PA タスクの 2 種類のなりすまし音声攻撃を想定し構成されている．なりすまし音声を作成するためにまず，攻撃者が登録者の音声を入手する状況を考える必要がある．そこで，本コーパスでは，3.1 説の話者照合

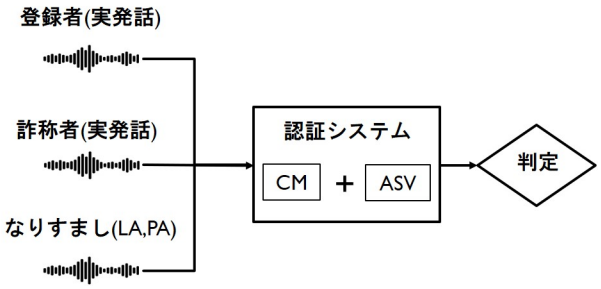


図 4: t-DCF の計算の流れ

評価用音声の収録である実収録の際に，攻撃者が盗聴音声を入手することを想定し，図 3 に示すように登録者から約 1 メートル程度離れたところに置かれた盗聴用機器で盗聴音声の収録を行った．LA タスクについては盗聴音声を用いて登録者の音声合成モデルを学習することで任意の文章について登録話者に似た音声を合成することでなりすまし音声を生成し，なりすまし音声検出の評価を行った．PA タスクについては攻撃方法が録音再生攻撃となるため，盗聴音声をもう一度盗聴音声再生機器で再生し，それをなりすまし音声収録用機器で再収録したものをなりすまし音声として用いた．この時のなりすまし音声収録用機器は実収録で使用したものと同じ機器である．盗聴音声の再生時には，実収録と異なる室内での収録を行ったが，収録環境としては静かな室内，空調が動作している室内，背景音楽が流れている室内の 3 つの収録環境を用意した．

3.3 話者照合となりすまし音声検出の評価

話者照合の性能評価の指標には本人棄却率 (False rejection rate; FRR) と他人受入率 (False acceptance rate; FAR) が等しくなる点である，等価エラー率 (Equal error rate; EER) を用いる．EER は値が小さい程照合性能が高いことを意味する．なりすまし音声検出の性能評価の指標としては，なりすまし音声受入率と実発話棄却率が等しくなる点である等価エラー率（以降 CM-EER）を用いる．さらに，話者照合となりすまし音声検出両方の判定を考慮した指標としてタンデム検出コスト関数 (Tandem detection cost function; t-DCF)[15] も用いる．t-DCF による計算考えるために図 4 に各音声の処理の流れを示す．図 4 に示す通り入力音声は登録者 (実発話)，詐称者 (実発話)，なりすまし (LA，PA) の 3 種類を想定しており，CM と ASV はなりすまし音声検出器と話者照合器を表している．t-DCF の計算式は以下のようになる [16][17].

$$\text{t}-\text{DCF}(\tau_{cm}) = C_0 + C_1 P^{cm}_{miss}(\tau_{cm}) + C_2 P^{cm}_{fa}(\tau_{cm}) \quad (1)$$

ここで P^{cm}_{miss} と P^{cm}_{fa} はなりすまし音声受入率と実発話棄却率が等しくなるような閾値 τ_{cm} を選んだときに求まる実発話棄却率となりすまし音声受入率である．また C_0，C_1，C_2 は次の式から求まる．

$$C_0 = \pi_{tar} C_{miss} P_{miss}^{asv} + \pi_{non} C_{fa} P_{fa}^{asv} \quad (2)$$

$$C_1 = \pi_{tar} C_{miss} - (\pi_{tar} C_{miss} P_{miss}^{asv} + \pi_{non} C_{fa} P_{fa}^{asv}) \quad (3)$$

$$C_2 = \pi_{spoof} C_{fa,spoof} P_{fa,spoof}^{asv} \quad (4)$$

ここで π_{tar}，π_{non}，π_{spoof} は，図 4 における登録者 (実発話)，詐称者 (実発話)，なりすまし (LA，PA) の事前確率に対応している．また C_{miss}，C_{fa}，$C_{fa,spoof}$ はそれぞれ本人棄却率時のコスト，他人誤受入時のコスト，なりすまし音声誤受入時のコストである．P_{miss}^{asv}，P_{fa}^{asv}，$P_{fa,spoof}^{asv}$ はそれぞれ本人棄却率，他人受入率，なりすまし音声受入率である．t-DCF も値が小さい程なりすまし音声検出と話者照合を組み合わせた全体の認証システムの性能が高いことを意味している．

4. 話者照合評価実験

4.1 実験条件

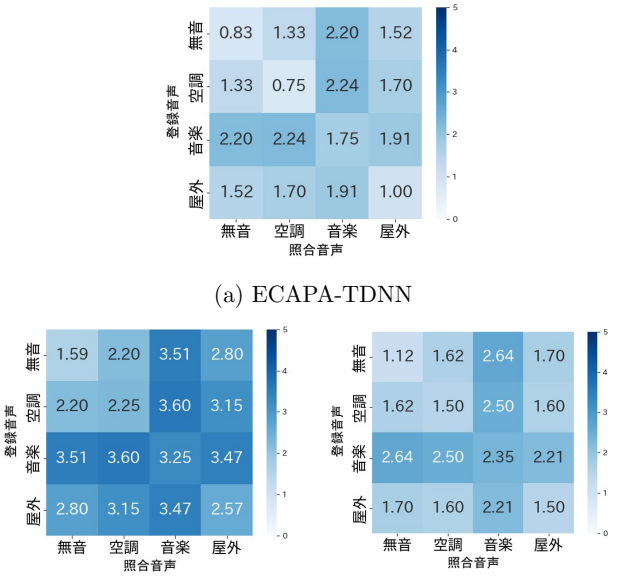
本章では，J-SpAW コーパスの話者照合評価セットの性能を調査するために，話者照合モデルとして最先端と言われる 3 手法の pretrain モデルを用いて評価を行った．用いた話者照合モデルは ECAPA-TDNN[18]，ResNetSE34V2[6][19]，RawNet3[3][19] であり，すべて大規模音声コーパス Voxceleb2[20] で学習されている．J-SpAW の実収録で収録された音声は合計で 8,000 発話 (40 話者 ×50 発話 ×4 環境 (実収録)) となっているが，話者照合用評価セットの登録音声及び照合音声には実収録した 50 個の定型文のうち 5 文を使用しており．話者数及び実収録の収録環境を加味して，合計 800 発話 (40 話者 ×5 発話 ×4 収録環境 (実収録)) となっている．話者照合の適切な評価のためには本人同士および他人同士の評価ペア（トライアル）を十分な数用意する必要がある．そこで J-SpAW では Voxceleb1 の評価タスクとほぼ同量となるよう本人同士のトライアルが 7,600 個，他人同士のトライアルが 30,000 個とし，合計で 37,600 個のトライアルを評価セットとして用意した．実収録時に使用した実発話収録用機器は Pixel3(以下 Pixel) 及び iPhone8 Plus(以下 iPhone) であり，2 台のスマートフォンを並べて同時収録を行った．本実験では Pixel で収録された音声を登録者もしくは詐称者の実発話として用い実験を行った．ただし，実験結果は iPhone の場合でもほぼ同様であった．

4.2 実験結果

表 1 に各モデルにおける EER(%) を示す．3 つのモデルすべてにおいて，J-SpAW の EER の方が Voxceleb1 より若干高いが，話者照合の精度としては十分に低い EER を得られていることがわかる．J-SpAW の照合結果を分析するために登録時及び照合時の収録環境別の EER を図 4 に示す．静かな室内を「無音」，空調が動作している環境の室内を「空調」，音楽が流れている環境の室内を「音楽」，室外での収録を「屋外」としている．図より登録音声及び照合音声がどちらも静かな室内で収録された際の EER は Voxceleb の EER とほぼ同等であることから，表 1 において J-SpAW の EER が Voxceleb よりも高くなっている要因は収録環境の違いによる影響が大きいということが推測される．また，登録時及び照合時どちらかに音楽が流れている環境の音声が用いられた場合の EER が高くなっていることがわかる．これは音声以外の雑音が大きいため，音声の話者性を捉えることが難しく，話者照合の難易度が上がったためだと考えられる．また同じ収録環境で録音された音声に対して話者照合を行ったときの EER は全体的に低いこともわかる．このことから最先端モデルを用いた話者照合において言語の違いに対する影響は小さいものの収録環境の違いが精度に関係していることが確認できた．

表 1: 各モデルにおける EER(%)

Model	J-SpAW	Voxceleb1
ECAPA-TDNN[18]	1.75	0.86
ResNetSE34V2[19]	2.99	1.02
RawNet3[19]	1.87	0.89



(a) ECAPA-TDNN

(b) ResNetSE34V2

(c) RawNet3

図 5: J-SpAW に対する登録音声と照合音声の収録環境ごとの EER(%)

5. LA タスクのなりすまし音声検出評価実験

5.1 実験条件

本章では LA タスクにおける J-SpAW コーパスのなりすまし音声検出の難易度を調査するために，ASVspoof2021 の LA タスクでベースラインとして用いられている 4 つのベースラインモデル (LFCC-GMM[9][10]，CQCC-GMM[11][12]，LFCC-LCNN[7]，RawNet2[8]) を用いた．使用した全てのモデルは ASVspoof2019 の LA タスク用

表 2: LA タスクに対する CM-EER・min t-DCF を計算するときに用いたトライアル数

音声コーパス	入力音声	ASV	CM
J-SpAW	本人 (実発話)	7600	800
	他人 (実発話)	30000	—
	なりすまし (PA)	800	800
ASVspoof	本人 (実発話)	16863	18452
	他人 (実発話)	605304	—
	なりすまし (PA)	163114	163114

表 3: LA タスクにおける CM-EER/t-DCF(V:VALL-EX, T: TTS, EER: CM-EER)

		J-SpAW		ASVspoof[24]	
モデル		EER	t-DCF	EER	t-DCF
CQCC-GMM	V	67.00	0.9913	15.62	0.4974
	T	31.50	0.8773		
LFCC-GMM	V	49.62	0.9499	19.30	0.5758
	T	3.50	0.5427		
LFCC-LCNN	V	50.00	1.0000	9.26	0.3445
	T	4.88	0.6322		
RawNet2	V	54.50	1.0000	9.50	0.4257
	T	18.25	0.9868		

に指定されている学習データを用いて学習されている．J-SpAW で用意した LA タスクのなりすまし音声について説明する．なりすまし音声は VALL-E X[21], x-vector を用いた話者適応によるテキスト音声合成 (Text to speech; TTS)[22] の 2 種類の音声合成手法により作成した．

VALL-E X による音声合成について説明する．VALL-E X は Microsoft 社が開発した再学習不要の音声合成モデルである．目標話者となる話者の数秒の音声と発話内容をテキストとして入力することで合成音声を作成する．入力音声には iPad mini(第 5 世代)(以下 iPad) で収録した盗聴音声のうち，各収録環境 (実収録) 各環境から 2 発話を選び．合計 480 発話 (40 話者 ×3 発話 ×4 環境) を用いた．VALL-E X は毎回合成される合成音声が異なるため，目標テキストごとに二回合成を行い，訓練済みの UTMOS[23] によって平均オピニオン評点 (Mean opinion score; MOS) の高い合成音声を採用する．次に x-vector を用いた話者適用による TTS について説明する．x-vector を用いて入力データの話者性を抽出し，その x-vector を用いた話者適応を施すことで少量の目標話者音声しか用いることができない場合でも目標話者の音声合成を生成できる手法である．本研究では，話者適応データとして iPad で収録した盗聴音声を用いた．最初になりすまし音声攻撃を行う対象の目標発話を設定する．目的発話と異なる発話内容の 45 音声を使用し，各収録環境 (実収録) から x-vector を抽出する．次にそれぞれの話者の各収録環境 45 発話から抽出した x-vector に対し平均を求めた．それぞれの話者に対し 4 収録環境 (実収録) の平均 x-vector が得られるため，合計 160 個 (40 話者 ×4 環境 (実収録)) となる．TTS モデルには，この平均 x-vector と合成したい発話内容を指定するテキストを入力することで，合成音声を生成した．

表 2 にトライアルの内訳を示した．参考のために ASVspoof2021 における LA のトライアルの内訳も記載した．また t-DCF を計算する際のパラメータは ASVspoof2021 と同じ値を使用し，コストと事前確率を，$\pi_{tar} = 0.9405$, $\pi_{non} = 0.0095$, $\pi_{spoof} = 0.05$, $C_{miss} = 1$, $C_{fa} = 10$, $C_{fa,spoof} = 10$, として t-DCF の計算を行った．

5.2 実験結果

表 3 に各モデルにおける CM-EER(%) と t-DCF を示す．2 種類の合成音声を比較すると TTS による合成音声に対する CM-EER が低いことが分かる．合成音声を聞いてみると，TTS による合成音声は自然性も話者性もが十分でなかったため，なりすまし音声検出が簡単にできてしまい CM-EER が低くなったと考えられる．一方，VALL-E X による合成音声に対する CM-EER はすべてのモデルで高くなっている．このことから VALL-E X による合成音声に対してはなりすまし音声検出が難しいことが分かる．また，ASVspoof2021 の評価データを用いた際の CM-EER と t-DCF と J-SpAW の結果を比較すると，CM-EER の傾向はモデルごとに異なるが，J-SpAW の t-DCF の結果かは全体的に高くなっていることがわかる．つまり J-SpAW による評価セットは難易度の高いものになっていることが確認できた．

6. PA タスクのなりすまし音声検出評価実験

6.1 実験条件

本章では PA タスクにおける J-SpAW コーパスのなりすまし音声の検出の難易度を調査するために，ASVspoof2021 の PA タスクでベースラインとして用いられている 4 つのモデルを用いた．ASVspoof2019 の PA タスク用に公開されているデータを用いて学習されている．次に J-SpAW における PA タスクのなりすまし音声について説明する．盗聴音声については LA タスクと同様の条件で入手したものとしている．PA タスクでのなりすまし音声の作成については次の通りである．実発話収録された定型文 50 発話のうち半分の 25 発話を収録環境ごとに用意し，合計 2,100 発話 (21 話者 ×25 発話 ×4 収録環境 (実収録)) を盗聴音声として盗聴音声再生機器から再生し，なりすまし音声収録用機器で再収録することにより作成した．PA タスクでは，盗聴音声再生機器として様々な種類のスピーカを用いることが可能であることから本実験では Bose Slink Micro，Sony SRS-ZR7，MacBook Pro，iPad の 4 種類を使用して録音

表 4: PA タスクに対する CM-EER・min t-DCF を計算するときに用いたトライアル数

音声コーパス	入力音声	ASV	CM
J-SpAW	本人 (実発話)	7,600	800
	他人 (実発話)	30,000	—
	なりすまし (PA)	6,300	6,300
ASVspoof	本人 (実発話)	16,863	18,452
	他人 (実発話)	605,304	—
	なりすまし (PA)	163,114	163,114

再生を行った．なりすまし音声収録用機器は実収録と同様 iPhone および Pixel の 2 種類で同時収録し作成したが，実験には pixel で再収録したなりすまし音声のみを用いた．なりすまし音声の収録環境としては静かな室内 (無音)，空調が動作している室内 (空調)，音楽が流れている室内 (音楽) の 3 種類を用意した．背景音楽は実収録とは異なる音源を用いた．本実験では，5 章と同様に実発話として実収録時の 800 発話 (40 話者 ×5 発話 ×4 収録環境 (実収録)) と BOSE Slink Micro を再生機器に用いて再収録したなりすまし音声 6,300 発話 (2,100 発話 ×3 収録環境 (PA)) とでなりすまし音声検出実験を行い，CM-EER を計算した．また，なりすまし検出と話者照合同時に評価する指標である t-DCF の算出のために，なりすまし音声を含む話者照合のトライアルを作成した．トライアルの内訳は表 4 の通りである．

6.2 実験結果

表 5 に J-SpAW を用いた PA によるなりすまし音声攻撃を行ったときの各モデルにおける CM-EER，t-DCF を示す．表より，ASVspoof の結果と同様にどのモデルでも非常に CM-EER が高いことがわかる．また，収録環境事に傾向を見ると LFCC-GMM と RawNet2 では静かな室内での録音再生より空調のかかった室内の方が CM-EER が 10 ポイント程度低下していることがわかるが，依然として CM-EER としては高い数値となっている．一方背景音楽が流れている際にはどのモデルでも CM-EER が非常に高くなっており，なりすまし音声検出のデータとして難しいタスクとなっていることがわかる．t-DCF で見てもほぼ 1 に近い値になっているものが多く，タスクの難易度の高さが確認できた．

7. 終わりに

本稿では，話者照合となりすまし音声検出のための日本語音声コーパス J-SpAW の構築と，その音声コーパスとしての性能を調査し報告した．J-SpAW は，実発話及びなりすまし音声生成用の盗聴音声の収録方法は現実的な盗聴方法を考慮して行われている．また，比較的簡易なりすまし音声生成手法を用いてのなりすまし音声攻撃を用意し公開することでまだ評価用の音声コーパスの多様性が十分でないなりすまし音声検出において有用な音声コーパスの一つとして用いれることを実験的に評価し示した．今後の課題として，PA タスクの録音再生による再収録時の多様性を増やすことや，LA タスクの合成音声の自然性を高くするなどで高度なりすまし音声攻撃を検出できるための音声コーパスとしての多様性を増やして行くことがあげられる．

表 5: PA タスクにおける CM-EER/t-DCF（EER: CM-EER）

	J-SpAW			ASVspoof[24]	
モデル		EER	t-DCF	EER	t-DCF
LFCC-GMM	無音	45.51	0.9710	39.54	0.9724
	空調	31.13	0.8075		
	音楽	47.65	1.0000		
CQCC-GMM	無音	40.76	0.9986	38.07	0.9434
	空調	41.24	0.9995		
	音楽	40.87	1.0000		
LFCC-LCNN	無音	72.24	1.0000	44.47	0.9958
	空調	67.62	1.0000		
	音楽	47.87	1.0000		
RawNet2	無音	35.00	0.9368	48.60	0.9997
	空調	27.73	0.7332		
	音楽	56.87	1.0000		

謝辞 本研究は JSPS 科研費 JP24K14993, 22H03639 の助成を受けたものです．

参考文献

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "x-vectors: Robust dnn embeddings for speaker recognition". *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE*, pp. 5329–5333, 2018.

[2] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification". *Proc. Interspeech 2020, pp. 3830–3834.ISCA, 2020.*

[3] Jee weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. "Pushing the limits of raw waveform speaker recognition". *Proc. Interspeech 2022*, 2022.

[4] 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,pp338-346, 2022.

[5] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, and Nicholas Evans. "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection". *ASVspoof 2021 Workshop, pp. 47–54*, 2021.

[6] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. "In defence of metric learning for speaker recognition". *Proc. Inter-*

speech 2020, 2020.

[7] Xin Wang and Junich Yamagishi. "A comparative study on recent neural spoofing countermeasures for synthetic speech detection". *arXiv: 2103.11326 [eess.AS]. 2021.*

[8] Massimiliano Todisco Andreas Nautsch Nicholas Evans Hemlata Tak, Jose Patino and Anthony Larcher. "End-to-end antispoofing with RawNet2". *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp. 6369–6373*, 2021.

[9] Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilci. "A comparison of features for synthetic speech detection". *Proc. Interspeech, 2015, pp. 2087–2091.ISCA*, 2015.

[10] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco. "Spoofing Attack Detection Using the Non-Linear Fusion of Sub-Band Classifiers". *Proc. Interspeech 2020, pp. 1106–1110.*, 2020.

[11] Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients". *Proc. Odyssey 2016, pp. 283–290*, 2016.

[12] Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification". *Computer Speech Language, vol. 45, pp. 516–535, 2017.*

[13] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. "Audio replay attack detection with deep learning frameworks". *Proc. Interspeech 2017,pp 82-86*, 2017.

[14] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Uamagishi, Federico Alegre, and Haizhou Li. "Spoofing and countermeasures for speaker verification: A survey". *Speech Communication,pp130-153*, 2015.

[15] Hector Delgado Nicholas Evans Massimiliano Todisco Md Sahidullah Junichi Yamagishi Douglas A. Reynolds Tomi Kinnunen, Kong Aik Lee. "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification". *Proc. Odyssey 2018*, 2018.

[16] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, and Others. "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan". *arXiv preprint arXiv:2109.00535*, 2021.

[17] Tomi Kinnunen, Hector Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, and Douglas A Reynolds. "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals". *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2195–2210, 2020.

[18] https://github.com/TaoRuijie/ECAPA-TDNN.

[19] https://github.com/clovaai/voxceleb_trainer.

[20] Andrew Zisserman Joon Son Chung, Arsha Nagrani. "VoxCeleb2: Deep Speaker Recognition". *Proc. Interspeech 2018*, 2018.

[21] Ziqiang Zhang, Long Zhou, et al. "speak foreign languages with your own voice: Cross-lingual neural codec language modeling". *arXiv:2303.03936(2023).*

[22] Kenntaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari. "text-to-speech synthesis from dark data with evaluation-in-the-loop data selection". *Proc. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[23] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022". *Proc. Interspeech 2022*, 2022.

[24] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild", Proc. 2023 IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.