

劣環境下における Deepfake 音声検出のためのドメイン適応

堤 歩斗¹ 後藤 晃² 齊藤 裕子² 松浦 廣樹² 塩田 さやか¹

概要: 近年のゼロショット音声クローン技術の発展により、少量の参照音声から話者忠実度の高い音声生成が可能となり、Deepfake 攻撃の脅威が高まっている。既存の Deepfake 検出モデルはスタジオ品質の英語音声で高い検出性能を達成しているが、異なる言語や劣環境への適用可能性は十分に検証されていない。そこで本研究では、電話・Telegram・対面インタビューという劣環境で収録された日本語音声を用いて、日本語対応の最先端ゼロショット音声合成モデル 11 種類による攻撃を検証した。CosyVoice3 などの最先端モデルによる攻撃では、話者照合システムの等価エラー率 (EER) を 0.91% から最大 11.63% まで上昇させ、Deepfake 検出モデルでも EER が 50% に達する等、最先端モデルが脅威となることを示した。そこで、劣環境データを用いたドメイン適応を行うことで、どこまで精度を向上できるかを評価した。実験結果から、既知の攻撃に対しては劣環境下においても EER 2% 以下を達成し、未知の攻撃に対しても検出精度の改善が見られたことを報告する。

1. はじめに

近年の深層学習技術の発展により、人間の声を高精度に模倣した合成音声を容易に作成できるようになった。従来の音声合成技術では、話者適応を用いる場合でも高品質な録音データと追加の学習が必要であった [1], [2]。しかし、ゼロショット音声クローン (Zero-Shot Voice Cloning) 技術の登場により、わずか数秒から数十秒の参照音声があれば、合成音声モデルの再学習を行わずとも任意のテキストをその話者にかなり似せた声で読み上げる音声を生成することが可能となった [2], [3]。このような技術の普及は、音声を用いた詐欺や、著名人になりすましたフェイクニュースの拡散など、社会的な悪用リスクを増大させており、対策は急務となってきている。

Deepfake 音声への対策技術として、自動話者照合 (Automatic Speaker Verification; ASV) と、なりすまし音声検出 (Countermeasure; CM) がある [4]。近年主流となっている深層学習に基づく ASV では、入力音声から抽出した話者情報となる話者埋め込みと事前に登録された発話の話者埋め込みを比較し、本人かどうかを判定している。一方、CM は入力音声が入音の実発話 (bonafide) か、合成・変換されたなりすまし音声 (spoof) かを判定する技術である。実用的な話者認証システムでは、ASV と CM を組み合わせることで、頑健性を向上させることができる。

既存の CM 研究はスタジオ品質の英語音声を前提としており、実環境で収録された音声や言語の依存性は十分に検証されていない。しかし、実際の法科学分野や金融機関などで求められる話者照合システムの運用形態としては、電話回線やメッセージアプリを経由した音声を扱うことが想定される。これらの音声には背景雑音や伝送路の歪み、帯域制限などの劣化が含まれており、このような実運用を見据えた劣環境においては、既存のシステムは性能が大幅に低下するとされている。そこで本研究では、劣環境下における Deepfake 攻撃の脅威を評価し、またその対策について提案する。使用するデータは、実際の電話回線・Telegram アプリ音声通話・対面インタビューの 3 条件で収録された音声から構成され、実運用環境での収録条件を備える。これらのデータに対し、11 種類の音声合成モデルを用いて Deepfake 音声を生成し、ASV および CM モデルに対する最先端 Deepfake の脅威を広範に評価する。実験結果より、最先端の音声合成モデルによる攻撃は劣環境下においても ASV の等価エラー率 (Equal error rate; EER) を 0.91% から最大 11.63% まで上昇させ、システムを欺く脅威となることを示した。また、Deepfake 検出モデルは高品質音声条件では検出に成功したが、劣環境下では EER が 50% に達し、検出が機能しないことを示した。これに対し、劣環境データを用いたドメイン適応を行った結果、学習データに含まれる既知の攻撃に対しては劣環境下においても EER 2% 以下を達成し、未知攻撃に対しても検出精度の大幅な改善が見られるなど、ドメイン適応の有効性が示された。

¹ 東京都立大学 システムデザイン学部
〒191-0065 東京都日野市旭が丘 6-6

² NEC 第一官庁システム開発統括部 研究開発グループ
〒108-8001 東京都港区芝 5-7-1

2. 関連研究

2.1 なりすまし音声検出

CM は、入力音声 が bonafide か spoof か を判定する二値分類タスクであり、CM のシステムを評価する国際的なコンペティションである ASVspoof Challenge [5], [6] などを通して近年活発に研究が行われている技術である。2019 年に開催された ASVspoof Challenge (ASVspoof2019) では VCTK Corpus [7] を用いたスタジオ品質の英語音声に対して、テキスト音声合成 (Text-to-Speech; TTS) と声質変換 (Voice Conversion; VC) による攻撃の検出が評価された。評価データには未知の攻撃に対する汎化性能をはかるために 11 種類の学習データに含まれない攻撃手法が用意された [8]。2 年後に開催された ASVspoof2021 ではさらに多くの未知条件の攻撃が評価データに用いられ、CM モデルの汎化性能が評価された [5]。近年では ASVspoof5[6] や SpoofCeleb[9] など、実環境で収集された音声やゼロショット攻撃を含むデータセットも登場しており、CM では次々と公開される最先端技術へ対応するために汎化性能の向上が強く求められている。

代表的な CM モデルとしては、RawNet2[10] や、自己教師あり学習特徴量を用いた wav2vec2+AASIST[11] などがある。これらのモデルは ASVspoof データの評価セットで高い性能を示すが、学習データと異なる言語・チャネル・攻撃手法への汎化性能については検証が不十分である。ASVspoof2021 では電話回線を経由した音声の評価データに追加されたが、これはシミュレーションで劣化を付与したものであり、実環境で収録された音声とは特性が異なる可能性がある。また、これらのデータセットは英語音声を対象としており、他言語への汎化性能は十分に検証されていない。

2.2 Deepfake 音声攻撃

話者の声を模倣する技術には、TTS と VC の 2 種類がある。TTS はテキストを入力として音声を生成する技術であり、VC は入力された音声の声質を別の話者のものに変換する技術である。従来の TTS による Deepfake 攻撃は、対象話者のまとまった高品質録音データを用いた適応学習が必要であった [1], [2]。しかし、近年登場したゼロショット TTS では、数秒から数十秒程度の参照音声から話者特徴を抽出し、対象話者の声を模倣して任意のテキストを読み上げる音声を生成できるようになってきている。初期のゼロショット TTS である VALL-E X [12] は、Neural Codec Language Model を用いて多言語対応のゼロショット合成を実現したが、自然性や話者忠実度に課題があった。その後発表された、CosyVoice3 [13] や XTTS [14] などでは、大規模データによる学習やモデル設計の改良により、高

表 1 劣環境データのデータ量と収録方法

条件	話者数	時間	収録方法・劣化要因
Interview	78	2.8h	対面会話, IC レコーダ, 背景雑音
Phone	78	7.7h	携帯電話 (3G 回線), 帯域制限
Telegram	78	4.1h	VoIP 音声通話, コーデック圧縮

い話者忠実度と自然性を両立している。また、従来のゼロショット TTS モデルは英語が中心であったが、近年は日本語を含む多言語に対応したモデルも多く登場している。VC においても同様の進展があり、わずかな参照音声のみから話者性を抽出するゼロショット VC が登場している。代表例として、OpenVoice [15] や Seed-VC [16] があり、いずれもわずかな参照音声に基づくゼロショットの声質変換を実現している。TTS が任意のテキストから音声を生成するのに対し、VC は既存の発話内容を保持したまま声質のみを変換するため、より自然な韻律を維持しやすいという特徴があり、なりすまし音声攻撃としての難易度がさらに高くなってきている。

3. 提案法

3.1 劣環境音声

本研究では、小澤ら [17] が収録した日本語音声データ (以下、劣環境データ) を使用した。表 1 に各条件の詳細を示す。Interview は対面会話を IC レコーダで収録しており、話者とマイクの距離に起因する背景雑音や残響の影響を受ける。Phone は携帯電話の 3G 回線を経由するため、帯域制限による高周波成分の減衰が生じる。Telegram は VoIP アプリの音声コーデックによる圧縮アーティファクトが含まれる。音声は電話帯域を想定し、8 kHz にダウンサンプリングして用いた。これらにはシミュレーションでは再現できない実環境の劣化も含まれる。

3.2 Deepfake 音声の生成

表 2 に本研究で使用した 11 種類の TTS/VC モデルを示す。音響モデルでは音声波形を生成するアーキテクチャを示しており、使用したモデルには FlowMatching [18] や GAN [19], VITS [20] など多様なアプローチが含まれていることがわかる。話者条件付けでは、参照音声からどのように話者情報を抽出するかを示しており、Ref は参照音声のトークン列やメルスペクトログラムを直接使用して条件付けや文脈内学習を行う方式、CAM++ [21] や Style Enc は話者埋め込みを抽出する方式であることを示している。これらのゼロショット TTS/VC を使い、3.1 節で述べた劣環境データを参照音声として Deepfake 音声を生成した。劣環境データの 78 話者のうち、63 話者を評価用、15 話者を CM モデルのファインチューニング用 (4.4 節) として使用した。評価用 63 話者に対して、各話者の bonafide 音声を 20 秒ごとのセグメントに分割し、各話者の最初のセ

表 2 使用した音声合成モデルのアーキテクチャ。

モデル名	音響モデル	話者条件付け
TTS		
CosyVoice2[22]	Flow+HiFT	Ref, CAM++
CosyVoice3[13]	Flow+HiFT	Ref, CAM++
XTTS-v2[14]	HiFi-GAN	H/ASP
GPT-SoVITS[23]	VITS	Ref
VALL-E X[12]	Vocos	Ref
Chatterbox[24]	HiFT	Ref, Style Enc
FishAudio s1-mini[25]	HiFT	Ref
VC		
CosyVoice2-VC[22]	Flow+HiFT	Ref, CAM++
CosyVoice3-VC[13]	Flow+HiFT	Ref, CAM++
OpenVoice V2[15]	VITS	Style Enc
Seed-VC[16]	Diffusion+BigVGAN	Style Enc

グメントを参照音声として使用して Deepfake 音声を生成した。図 1 に Deepfake 生成の概要を示す。TTS モデルでは、各セグメントを音声認識モデルにより書き起こしたテキストを入力として、対象話者のなりすまし音声を生成した。VC モデルでは、ファインチューニング用 15 話者の発話音声をソース音声とし、対象話者の参照音声の声質に変換することで、なりすまし音声を生成した。いずれのモデルも 24~44 kHz で音声を出力するため、実発話に合わせて 8 kHz にダウンサンプリングを行った。

4. 実験条件

4.1 合成音声の品質評価

合成音声の自然性評価には UTMOS[26] を使用した。UTMOS は MOS 評価データで学習されたニューラルネットワークモデルであり、1~5 点で音声の自然性を自動予測する指標である。全ての音声は 8 kHz にダウンサンプリングされた後、モデルの入力に合わせて 16 kHz にリサンプリングされており、以降の実験でも 16 kHz で使用されている。

4.2 ASV 評価

Deepfake 音声 ASV を欺いて対象話者として認識されるかを検証するため、ASV による評価を行った。ASV モデルとして VoxCeleb1 および VoxCeleb2 で学習された ECAPA-TDNN [27], [28] を使用した。評価データとしては劣環境データの 3 種類に加えて、スタジオ収録で収集された JVS コーパス [29] を用いた。JVS コーパスでは 20 話者の発話は VC のソース音声として使用し、80 話者分の合成音声を生成し、評価に使用した。これら 4 つの条件で独立に以下の 3 種類のトライアルを作成した：

- **target:** bonafide 同士の同一話者ペア
- **non-target:** bonafide 同士の異なる話者ペア
- **spooftrial:** Deepfake 音声と対象話者の bonafide 音

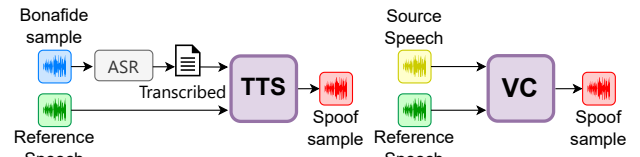


図 1 ゼロショット TTS/VC による Deepfake 生成のフロー。

声のペア

spooft trial では、Deepfake 音声は対象話者の声を模しているが、合成音声であるため判定上は異なる話者ペアとして扱う。ASV モデルが Deepfake 音声を対象話者と判定すれば Deepfake 攻撃は成功し、異なる話者と判定すれば攻撃は失敗となる。評価指標には EER_{ASV} を用いる。 EER_{ASV} は、異なる話者を同一話者と誤判定する誤受理率と、同一話者を異なる話者と誤判定する誤棄却率が等しくなる点での誤り率である。bonafide のみ (target および non-target) の EER_{ASV} を基準として、spooft trial を加えた際の EER_{ASV} 上昇により Deepfake 攻撃の脅威を評価する。そのために、各収録条件において、これら 3 種類のトライアルが同数になるようにアンダーサンプリングを行った。各条件の合計トライアル数は Interview 5,652 組、Phone 33,009 組、Telegram 8,913 組である。

4.3 CM 評価

CM モデルが日本語音声および各収録条件に対して Deepfake 音声を検出できるかを検証するため、CM の評価を行った。CM モデルとして ASVspooft2019 データセットで学習された RawNet2[10] および wav2vec2+AASIST[11] を使用した。評価指標には EER_{CM} を用いる。CM は ASV と異なり、各音声サンプルが bonafide か spooft かを判定するサンプル単位の評価を行う。 EER_{CM} は、spooft 音声を bonafide と誤判定する誤受理率と、bonafide 音声を spooft と誤判定する誤棄却率が等しくなる点での誤り率である。各条件のサンプル数は Interview で bonafide 5,709 件、spooft 9,004 件、Phone で bonafide 13,706 件、spooft 22,194 件、Telegram で bonafide 7,205 件、spooft 9,928 件である。

4.4 ドメイン適応

劣環境下における Deepfake 音声検出のために、CM モデルに対してドメイン適応を行った。ドメイン適応として、ASVspooft2019 データセットで事前学習済みの RawNet2 および wav2vec2+AASIST のファインチューニングを行った。学習データとして、評価から除外した 15 話者の bonafide 音声 (1.3 時間) と、既知攻撃として 6 種類の TTS/VC モデル (CosyVoice2, GPT-SoVITS [23], VALL-E X, XTTS-v2 [14], CosyVoice2-VC, OpenVoice) で合成した合計 12.4 時間の Deepfake 音声を使用した。5 種類のモデル (CosyVoice3, FishAudio [25], Chatterbox [24],

表 3 合成音声の評価結果. UTMOS (自然性, 1~5 点), EER_{ASV} (%), EER_{CM} (%). Int: Interview, Ph: Phone, Tel: Telegram. 太字はカテゴリ内で最も優れた合成品質・攻撃成功率を示す.

合成手法	UTMOS \uparrow				EER_{ASV} (%) \downarrow				EER_{CM} (%) \downarrow								
	Int	Ph	Tel	JVS	Int	Ph	Tel	JVS	RawNet2				wav2vec2+AASIST				
									Int	Ph	Tel	JVS	Int	Ph	Tel	JVS	
bonafide	1.37	1.65	1.67	2.97	0.53	0.58	0.91	0.10	-	-	-	-	-	-	-	-	-
CosyVoice2	2.25	2.36	2.51	2.84	8.78	8.05	6.11	10.63	41.3	60.0	44.1	17.1	57.9	62.8	58.1	11.1	
XTTS-v2	1.69	1.79	1.99	2.20	2.47	1.97	3.57	2.17	19.3	39.9	27.4	2.2	19.9	24.2	22.0	0.6	
GPT-SoVITS	1.28	1.24	1.27	1.99	0.37	0.39	0.72	0.17	94.9	93.8	93.6	15.2	43.5	46.9	49.3	8.3	
VALL-E X	1.34	1.51	1.62	1.75	1.75	2.05	2.66	0.45	49.2	66.2	59.2	33.2	48.1	43.7	48.7	21.9	
CosyVoice3	2.43	2.38	2.60	2.70	8.70	7.55	9.49	9.83	46.1	62.5	57.3	17.6	66.9	67.9	70.2	11.8	
Chatterbox	1.87	2.15	2.26	3.00	4.09	4.85	5.49	1.16	40.3	54.4	33.5	4.1	35.4	33.7	28.3	0.5	
FishAudio	2.47	2.54	2.78	3.12	7.96	6.38	8.58	8.59	37.4	41.9	31.0	3.3	41.0	48.6	47.1	0.3	
CosyVoice2-VC	1.81	2.12	2.08	2.88	8.06	5.33	7.80	8.51	44.9	55.7	49.3	46.9	50.2	66.5	60.4	43.3	
OpenVoice	1.29	1.68	1.60	2.17	0.63	0.52	1.13	0.07	46.8	60.6	63.0	43.4	39.7	48.0	45.3	30.8	
CosyVoice3-VC	1.95	2.08	1.95	2.73	9.31	10.07	11.63	10.54	47.4	60.0	54.1	51.6	55.7	64.6	58.6	36.3	
Seed-VC	1.41	1.57	1.60	2.31	2.74	6.78	6.00	3.25	37.3	50.7	42.8	49.4	23.1	26.4	29.3	10.2	

CosyVoice3-VC, Seed-VC) は評価のみに使用し, 未知攻撃に対しての汎化性能を評価するために使用した. 学習では, 元のドメインでの性能を維持するため, ASVspooof2019の学習データと本研究のデータを混合して学習し, ドメイン適応を行った. 学習は 50 エポック実施し, 検証データでの性能が最も良いエポックのモデルを採用した. RawNet2 はバッチサイズ 32, 学習率 $1e-4$, wav2vec2+AASIST はバッチサイズ 16, 学習率 $1e-4$ で学習した.

4.5 t-DCF 評価

ASV と CM を組み合わせたシステムの総合的な性能を評価するため, t-DCF (tandem Detection Cost Function) [30] による評価を行う. 本研究では文献 [30] に基づいたコストパラメータ ($P_{tar} = 0.9405$, $P_{non} = 0.0095$, $P_{spooof} = 0.05$, $C_{miss} = 1$, $C_{fa} = 10$) で min t-DCF を求めた. min t-DCF が 1.0 の場合は Deepfake 攻撃に脆弱であることを意味し, 0 に近いほどシステム全体として頑健であることを示す.

5. 実験結果

5.1 合成音声の品質評価

表 3 に UTMOS による bonafide と各合成音の自然性評価の結果を示す. まず bonafide の UTMOS スコアを見ると, JVS が 2.97 であるのに対して劣環境では 1.37~1.67 と低い値になっており, 収録条件による音声品質の違いが確認できる. 次に, TTS では FishAudio (2.47~3.12) や CosyVoice3 (2.38~2.70) が高いスコアを示していることがわかる. これらのモデルは劣環境において bonafide を上回るスコアを達成しており, チャンネル劣化を含む参照音声からでも自然性の高い音声を生成できることを示している. 一方, GPT-SoVITS は劣環境音声では合成音声の出力

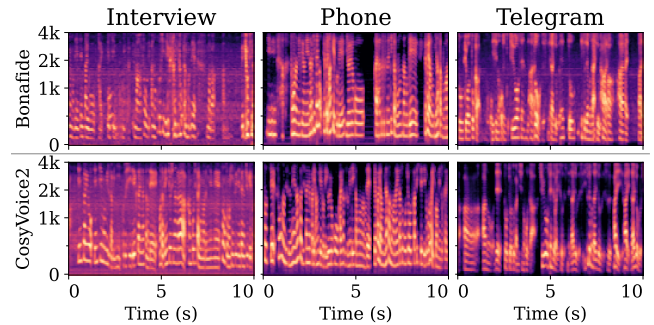


図 2 各収録条件における bonafide (上) と CosyVoice2 (下) のメルスペクトログラム比較.

結果が破綻しており, UTMOS スコアとしても極めて低い値 (1.24~1.28) を示した. 対して JVS では 1.99 と良好な品質で生成できており, 劣環境音声からの合成は難しいことを示している. また, VC ではどの合成手法においても bonafide に近いスコアとなった. これはソース音声の音韻情報を引き継いでいるためと考えられる.

次に, 図 2 に劣環境の bonafide と CosyVoice2 による合成音のメルスペクトログラムを示す. スペクトログラムから, 各収録条件に帯域制限, ノイズフロア, コーデックアーティファクト等のスペクトル特性が観察される. CosyVoice2 はこれらのチャンネル固有の特徴も再現していることが分かる.

5.2 ASV 評価

表 3 の EER_{ASV} に ASV の評価結果を示す. bonafide のみではどの条件でも EER_{ASV} は 0.10~0.91% と低く, ECAPA-TDNN が高い話者識別性能を持つことが確認できる. 一方, Deepfake 攻撃を含む評価では, GPT-SoVITS と OpenVoice を除く全ての合成手法で EER_{ASV} の上昇がみられた. 特に CosyVoice2, CosyVoice3-VC などの最先

表 4 ドメイン適応後 CM モデルの評価結果. EER_{CM} は CM 単体の性能 (%), min t-DCF は統合システムの耐攻撃性を示す. Int: Interview, Ph: Phone, Tel: Telegram. 太字はカテゴリ内で最も高い検出性能・頑健性を示す.

合成手法	RawNet2						wav2vec2+AASIST						
	EER_{CM} (%) ↓			min t-DCF ↓			EER_{CM} (%) ↓			min t-DCF ↓			
	Int	Ph	Tel	Int	Ph	Tel	Int	Ph	Tel	Int	Ph	Tel	
既知	CosyVoice2	7.3	4.6	5.4	0.16	0.11	0.07	1.3	0.2	0.2	0.03	0.01	0.01
	XTTS-v2	22.7	7.7	11.0	0.07	0.05	0.09	0.4	0.1	0.1	0.01	0.00	0.01
	GPT-SoVITS	4.7	0.6	1.6	0.00	0.00	0.01	0.1	0.0	0.2	0.00	0.00	0.01
	VALL-E X	16.4	8.4	13.6	0.06	0.05	0.08	2.0	0.4	1.0	0.02	0.01	0.03
	CosyVoice2-VC	5.6	5.2	6.6	0.13	0.09	0.14	2.0	0.0	0.6	0.03	0.00	0.02
	OpenVoice	8.5	4.3	6.8	0.02	0.01	0.03	0.3	0.1	0.6	0.01	0.00	0.01
未知	CosyVoice3	11.7	7.1	9.7	0.24	0.15	0.19	3.2	0.5	1.0	0.07	0.01	0.03
	Chatterbox	18.6	14.1	18.5	0.13	0.13	0.16	5.6	0.4	0.6	0.06	0.01	0.02
	FishAudio	30.6	27.3	31.9	0.25	0.19	0.27	6.3	15.8	19.4	0.14	0.18	0.26
	CosyVoice3-VC	10.6	6.9	9.4	0.26	0.17	0.24	1.7	0.4	0.1	0.03	0.01	0.01
	Seed-VC	36.2	38.9	41.3	0.09	0.22	0.19	7.5	8.0	8.2	0.05	0.10	0.13

端モデルは大幅に EER_{ASV} を上昇させ Deepfake 攻撃の高い脅威を示した. 例えば, Telegram 条件では bonafide のみの EER_{ASV} 0.91% に対し, CosyVoice3-VC 攻撃時には 11.63% と 12 倍に上昇しており, 最先端モデルが生成する Deepfake 音声が高い話者忠実度を持ち, すべての条件で ASV を効果的に欺くことを示している.

次に, Deepfake 音声の品質と ASV 性能の関係を分析する. 高い UTMOS スコアを示した合成手法である FishAudio, CosyVoice3, CosyVoice2 などのモデルはいずれも EER_{ASV} の上昇が大きい. これは ASV モデルが品質の良い Deepfake 音声を本人の音声だと判定してしまうことを示唆している.

図 3 に CosyVoice2 攻撃時の ASV モデルの出力するスコアの分布のヒストグラムを示す. Target のスコア分布と spoof のスコア分布が重なっており, ASV モデルでは Deepfake 音声の対象話者として判定されてしまうことが確認できる.

5.3 CM 評価

表 3 の EER_{CM} に CM の評価結果を示す. なお, 参考として RawNet2 および wav2vec2+AASIST は ASVspoof2021 評価データではそれぞれ EER_{CM} が RawNet2: 5.72%, wav2vec2+AASIST: 0.96% を達成している. JVS では, RawNet2 と wav2vec2+AASIST 共に, TTS 攻撃に対して ASVspoof2021 と同程度の高い検出性能を発揮した. しかし, VC 攻撃については検出が困難であり, VC による Deepfake 攻撃の検出が難しいことが分かる. 一方, 劣環境では TTS・VC 共に EER_{CM} が非常に高く, 多くの場合で 50% 前後となった. これは判定がほぼ機能していないことを意味し, CM モデルのドメイン外のチャンネルに対する

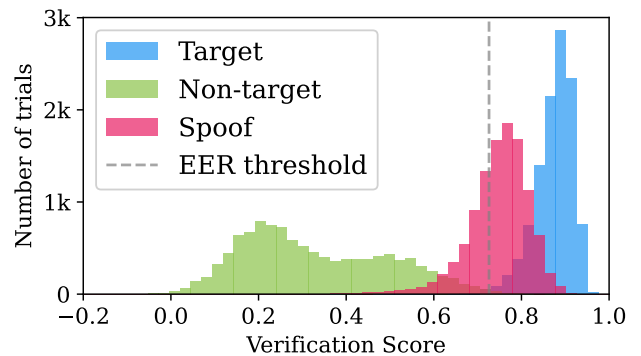


図 3 ASV スコア分布 (Phone 条件, CosyVoice2 攻撃). target (青), non-target (緑), spoof (赤) の 3 種類のスコア分布を示す. 点線は EER 時の判定閾値.

脆弱性が顕在化した. 特に, RawNet2 は GPT-SoVITS に対して 93~95% という極端に高い EER_{CM} を示した. これは GPT-SoVITS の合成破綻により, 通常の合成音声とは異なる特徴を持つ音声生成され, CM モデルが誤って bonafide と判定したためと考えられる. これらの結果は, CM の汎化性能低下の主因が言語ミスマッチではなくチャネル劣化であることを示している.

5.4 ドメイン適応

表 4 にドメイン適応後の CM の評価結果を示す. 両モデルともに, ドメイン適応により表 3 の結果から大幅に検出性能が改善した. wav2vec2+AASIST は特に顕著な改善を示し, 学習データに含まれていた 6 種類の既知攻撃に対して EER_{CM} が 2% 以下となった. 学習データに含まれていなかった 5 種類の未知攻撃についても性能が大幅に改善した. CosyVoice3 と CosyVoice3-VC は, wav2vec2+AASIST で 3.2% 以下と高精度で検出が可能であった. これは表 2 に

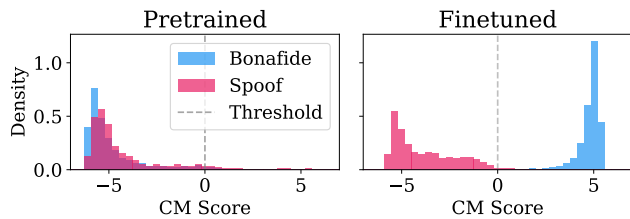


図 4 CM スコア分布の比較 (wav2vec2+AASIST, CosyVoice2 攻撃). bonafide (青), spoof (赤). 点線は閾値. 左: 事前学習済みモデル, 右: ドメイン適応後.

示す通り, CosyVoice3 が CosyVoice2 と同様の構造を持つモデルである事に由来している可能性がある. また, 完全に未知攻撃である Chatterbox も EER_{CM} 5.6%以下と良好な検出精度を示した. 一方で FishAudio (6.3~19.4%) と Seed-VC (7.5~8.2%) では, ドメイン適応前から性能は改善したものの, 依然として EER_{CM} は高く, 攻撃手法自体の検出の難しさを示す結果となった.

一方, ドメイン適応の弊害として, ASVspoof2021 評価データでの性能劣化が確認された. RawNet2 は ASVspoof2021 での EER_{CM} が 5.72%から 22.23%に劣化した. 一方で, wav2vec2+AASIST は 0.96%から 1.30%と影響は軽微であった.

図 4 に CosyVoice2 における wav2vec2+AASIST のドメイン適応前後のモデルの CM スコア分布の比較を示す. 事前学習済みモデルでは bonafide と spoof の CM スコア分布が完全に重なっていたものが, ドメイン適応により分離していることが確認できる. ここで, ドメイン適応により分布が変化しているのは主に bonafide 側であることが分かる. これは, モデルが特定の攻撃手法の特徴を学習したのではなく, 劣環境の bonafide 音声に適応出来たことを示している. 学習に含まれていなかった未知攻撃に対しても性能が改善したことから, ドメイン適応が意図した通り機能していることが確認できる.

5.5 t-DCF 評価

表 4 の min t-DCF に t-DCF での評価結果を示す. wav2vec2+AASIST は既知攻撃に対して min t-DCF 0.01~0.03 を達成し, すべての環境で高い頑健性を持つことが確認された. 特に Phone 条件では多くの攻撃で 0.01 以下の min t-DCF を示し, なりすまし攻撃をほぼ完全に防御できることが確認された. 未知攻撃への汎化については, 攻撃手法により結果が分かれた. CosyVoice3 (0.01~0.07) や Chatterbox (0.01~0.06) では良好な汎化性能を示した一方で, FishAudio (0.14~0.26) や Seed-VC (0.05~0.13) は比較的高い値を示した. FishAudio は ASV・CM 両方に対して効果的な攻撃手法であるため, min t-DCF も高くなっており, 最先端モデルの脅威は未だ健在である.

6. まとめ

本研究では, 実環境で収録された日本語音声に対し, 11 種類のゼロショット TTS/VC による Deepfake 攻撃の脅威を評価した. 実験では攻撃により EER_{ASV} が最大 11.6% まで上昇した. 事前学習済みの CM モデルは高品質音声では検出に成功したが, 劣環境下では EER_{CM} が 50% を超え検出に失敗し, CM 性能低下の主因は言語ではなくチャネル劣化であることが明らかになった. 同一チャネルデータによるドメイン適応で性能は改善したが, FishAudio や Seed-VC など一部の未知攻撃への汎化は依然として課題である.

参考文献

- [1] Joshi, R. and Garera, N.: Rapid Speaker Adaptation in Low Resource Text to Speech Systems using Synthetic Data and Transfer learning, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 267–273 (2023).
- [2] Azzuni, H. and Saddik, A. E.: Voice Cloning: Comprehensive Survey, *arXiv preprint arXiv:2505.00579* (2025).
- [3] Chen, S., Wang, C., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S. and Wei, F.: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 33, pp. 705–718 (online), DOI: 10.1109/TASL-PRO.2025.3530270 (2025).
- [4] Li, M., Ahmadiadli, Y. and Zhang, X.-P.: A Survey on Speech Deepfake Detection, *ACM Comput. Surv.*, Vol. 57, No. 7 (online), DOI: 10.1145/3714458 (2025).
- [5] Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A. et al.: Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 2507–2522 (2023).
- [6] Wang, X., Delgado, H., Tak, H., weon Jung, J., jin Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T. H., Evans, N., Lee, K. A. and Yamagishi, J.: ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale, *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pp. 1–8 (online), DOI: 10.21437/ASVspoof.2024-1 (2024).
- [7] Yamagishi, J., Veaux, C. and MacDonald, K.: CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive*. (2019).
- [8] Nautsch, A., Wang, X., Evans, N., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J. and Lee, K. A.: ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 3, No. 2, pp. 252–265 (2021).
- [9] Jung, J.-w. et al.: SpoofCeleb: Speech deepfake detec-

- tion and SASV in the wild, *OJSP* (2025).
- [10] Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N. and Larcher, A.: End-to-end anti-spoofing with rawnet2, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6369–6373 (2021).
- [11] Tak, H., Todisco, M., Wang, X., Jung, J.-w., Yamagishi, J. and Evans, N.: Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, *The Speaker and Language Recognition Workshop* (2022).
- [12] Zhang, Z., Zhou, L., Wang, C., Chen, S., Wu, Y., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J. et al.: Speak foreign languages with your own voice: Cross-lingual neural codec language modeling, *arXiv preprint arXiv:2303.03926* (2023).
- [13] Du, Z., Gao, C., Wang, Y., Yu, F., Zhao, T., Wang, H., Lv, X., Wang, H., Ni, C., Shi, X. et al.: Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, *arXiv preprint arXiv:2505.17589* (2025).
- [14] Casanova, E., Davis, K., Gölge, E., Gökner, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S. et al.: Xtts: a massively multilingual zero-shot text-to-speech model, *arXiv preprint arXiv:2406.04904* (2024).
- [15] Qin, Z., Zhao, W., Yu, X. and Sun, X.: Open-voice: Versatile instant voice cloning, *arXiv preprint arXiv:2312.01479* (2023).
- [16] Liu, S.: Zero-shot voice conversion with diffusion transformers, *arXiv preprint arXiv:2411.09943* (2024).
- [17] 小澤茂樹, 後藤 晃, 斎藤裕子, 松浦廣樹, 越仲孝文: 法科学分野への応用を想定したテキスト独立話者照合の精度評価, *SPEASIP* (2025).
- [18] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. and Le, M.: Flow Matching for Generative Modeling, *11th International Conference on Learning Representations, ICLR 2023* (2023).
- [19] Kong, J., Kim, J. and Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in neural information processing systems*, Vol. 33, pp. 17022–17033 (2020).
- [20] Kim, J., Kong, J. and Son, J.: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, *Proceedings of the 38th International Conference on Machine Learning* (Meila, M. and Zhang, T., eds.), Proceedings of Machine Learning Research, Vol. 139, PMLR, pp. 5530–5540 (2021).
- [21] Wang, H., Zheng, S., Chen, Y., Cheng, L. and Chen, Q.: CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking, *Interspeech 2023*, pp. 5301–5305 (online), DOI: 10.21437/Interspeech.2023-1513 (2023).
- [22] Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H. et al.: Cosyvoice 2: Scalable streaming speech synthesis with large language models, *arXiv preprint arXiv:2412.10117* (2024).
- [23] RVC-Boss: GPT-SoVITS: 1 min voice data can also be used to train a good TTS model, <https://github.com/RVC-Boss/GPT-SoVITS> (2024). accessed 2026-01-20.
- [24] Resemble AI: Chatterbox-TTS, <https://github.com/resemble-ai/chatterbox> (2025). GitHub repository.
- [25] Liao, S., Wang, Y., Li, T., Cheng, Y., Zhang, R., Zhou, R. and Xing, Y.: Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, *arXiv preprint arXiv:2411.01156* (2024).
- [26] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S. and Saruwatari, H.: UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022, *Interspeech 2022* (2022).
- [27] Desplanques, B., Thienpondt, J. and Demuynck, K.: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, *Interspeech 2020*, pp. 3830–3834 (online), DOI: 10.21437/Interspeech.2020-2650 (2020).
- [28] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D. and Bengio, Y.: SpeechBrain: A General-Purpose Speech Toolkit (2021). arXiv:2106.04624.
- [29] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JVS corpus: free Japanese multi-speaker voice corpus, *arXiv preprint arXiv:1908.06248* (2019).
- [30] Kinnunen, T., Delgado, H., Evans, N., Lee, K. A., Vestman, V., Nautsch, A., Todisco, M., Wang, X., Sahidullah, M., Yamagishi, J. and Reynolds, D. A.: Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2195–2210 (online), DOI: 10.1109/TASLP.2020.3009494 (2020).