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

不正収録音声から合成されたディープフェイク音声による なりすまし攻撃

古林嵯羽仁† 　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情報工学科

あらまし　近年，深層学習の発展によりディープフェイクメディアはますます現実との見分けがつきにくくなってきている．音声分野でも特定話者を高精度に模倣した合成音声を作成可能になるなど著しい発展が見られるが，悪意のある利用が問題視されており，未知のディープフェイクメディアを用いた攻撃に対する頑健性向上は重要なタスクとなっている．本研究では，詐称者が不正に収録した一般的には音声合成の学習データとして用いられないような背景雑音を含む音声から，音声強調と Zero-shot TTS を用いることにより複数の実話者を模倣したディープフェイク音声データを高品質に作成する手法を提案する．なりすまし音声検出 (CM) や話者照合 (ASV) の結果から本研究で作成した音声は既存の音声セキュリティシステムを高精度に突破できたことを報告する．

Spoofing attacks using deepfake speech synthesized from non-consensual recording

Sawato FURUBAYASHI†, Shinnosuke TAKAMICHI††, and Sayaka SHIOTA†

† Graduate School of Systems Design, Tokyo Metropolitan University
†† Faculty of Science and Technology, Keio University

1. はじめに

近年，深層学習によるディープフェイクメディアが問題視される機会は増えてきている．これまでに政治家などの実在の人物に，現実世界では行なっていない発言や行動を行わせるような悪意のあるディープフェイク音声，画像，動画によって風評被害をもたらすといった事例が数多く報告されており，対策が急務となっている [1].

ディープフェイク音声はテキスト音声合成（Text-to-speech; TTS）や声質変換によって生成されることが主流であり [2]，対策としてはなりすまし音声検出や話者照合を用いることが考えられる．なりすまし音声検出や話者照合に関しては，近年深層学習を用いてなりすまし音声や本人判定を高精度に識別する技術が提案されている [3], [4] が，一方でなりすまし音声検出や話者照合の突破を試みることを想定した合成音声生成に関する研究も行われている [5], [6]．また，これまでに公開されてきたなりすまし音声検出に関するデータベースでは攻撃者の立場を考慮した不正収録音声を用いる研究はあまりなされていない [7].

合成音声によるなりすまし音声攻撃として，これまでの研究では背景雑音の少ない高音質な音声を学習データとする合成音声を用いることが主流であった．しかしながら，攻撃者の観点からは，目標話者の音声を背景雑音が少なく，高音質な状態で十分に集めることは困難である．そこで本研究では攻撃者の視点から，日常的な環境で攻撃者が不正収録を行い実話者の音声を入手した場合のなりすまし音声生成を考える．具体的な方針として，不正収録音声を用いた合成音声の生成，および不正収録音声に音声強調を適用し，合成音声を生成することを行う．その際，攻撃者が不正収録できるデータ量は限られていると想定できることから，少量の参照音声から合成

が可能な Zero-shot TTS を用いる．作成したなりすまし音声について音声品質評価，なりすまし音声検出，話者照合という 3 つの観点から評価を行うことで不正収録からのなりすまし音声攻撃の難度を検証した．実験結果より，不正収録音声を用いた場合でも最先端の音声セキュリティシステムを突破できる音声が生成できることを報告する．

2. 関 連 研 究

2.1 ディープフェイク音声の検出技術

ディープフェイク音声への対策技術として，なりすまし音声検出や話者照合がある．一般的になりすまし音声検出を通過した音声に対し後段で話者照合が行われる．なりすまし音声検出では入力音声が人間が実際に発声している実発話音声であるか，機械で生成された音声であるかを判定する．入力音声がなりすまし音声検出にて実発話と判定された場合に，当該音声の話者性を判定するために話者照合が行われる．話者照合では照合対象の音声とシステム登録話者の実発話音声との類似度に基づき同一人物であるか否かが判定される．このような検出技術では多様化・高精度化する合成音声を正確に識別し，確実に棄却する性能が求められている．

2.2 Zero-Shot TTS

Zero-Shot TTS は，モデルの学習時に含まれていない未知の話者の音声を，わずか数秒程度の短い参照音声のみを用いて合成する技術である．従来，特定話者の声質を持つ音声合成システムを構築する手法として，対象話者の数十分から数時間にわたる高品質な収録音声を再学習させる話者適応が用いられていた．これに対し，Zero-Shot TTS では，大規模な多話者データセットを用いて TTS モデルの事前学習を行い，数秒の参照音声から抽出した話者埋め込みベクトルや音響特徴量をモデルの条件付けとして入力する．これにより，追加の学習を行うことなく，対象話者の声質や韻律特徴を模倣することが可能となる [8]〜[11].

2.3 音 声 強 調

音声強調は，雑音などが混入した観測信号から，目的となる音声信号を推定し相対的に強調する技術である．従来はスペクトルサブトラクション法などの統計的手法が用いられてきたが，近年では高品質に目的の信号を抽出できるモデルである DCCRNet [12]・DPTNet [13] をはじめとした深層学習を用いた手法が主流となり，非定常な雑音に対しても高い除去性能を示している．

2.4 J-SPAW の不正収録音声

J-SPAW [14] はなりすまし音声検出と話者照合のために作成された日本語音声データセットである．40 話者が日常環境として設定された 4 つの環境下にて各 1-5 秒程度の文を 50 種類読み上げている．J-SPAW に含まれる不正収録音声は，攻撃者が正規話者の発声をやや離れた地点から収録することを想定して作成されており，周囲の環境音や BGM などの背景雑音が含まれた音声となっている．

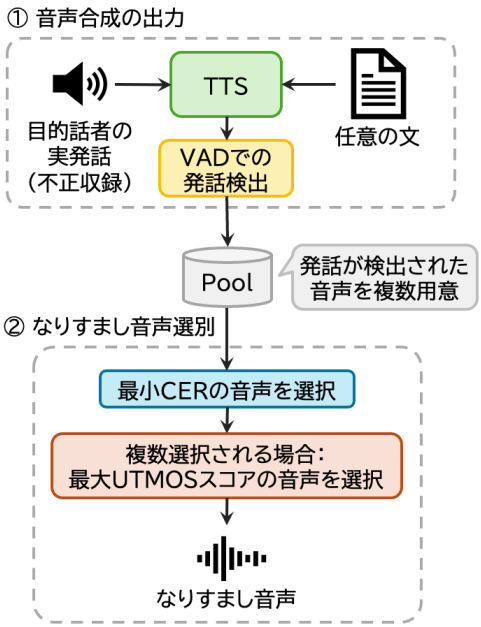

図 1: 不正収録音声を用いたなりすまし音声の処理フロー

3. 不正収録音声を用いたなりすまし音声生成

3.1 音声生成および選定工程

本研究では，数秒から数十秒程度の短い不正収録音声から，目標話者の声質を高精度に模倣したなりすまし音声を作成する手法を提案する．提案システムは図 1 に示す処理フローのように，まず Zero-shot TTS を通して目標話者を模倣した合成音声を得る工程①を経たのち，発話内容の明瞭さや音声品質をはじめとした客観的および主観的に自然と判断されやすい音声を採用するための工程②を行う．具体的な手法としては，まず工程①に示すように参照音声となる目標話者の実発話音声と，発声させたいテキストを Zero-shot TTS システムに入力し，音声波形を生成する．本研究ではこの実発話音声が不正収録音声によって入手されたものだと想定している．Zero-shot TTS は音声の合成過程にランダム性を含む手法が多いため，生成するごとに品質や韻律が変動するほか，発話区間の前後に不自然な無音区間やノイズが付与される場合がある．そこで本手法では，TTS の出力に対して音声区間検出 (Voice Activity Detection; VAD) を適用する．VAD により有音区間が検出されなかった場合は破棄し，検出された場合には合成音声のデータプールに保存する．この工程①を繰り返し，発話区間が正常に含まれる合成音声候補を複数個生成する．工程②では工程①で生成された候補の中から，最もなりすましに適した音声を選定する．

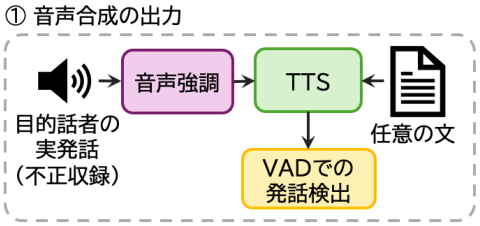

図 2: 実発話音声への音声強調適用

本研究でのなりすましに適した音声とは，機械と人間両方を騙すことが可能である音声を意味している．そのため，音声認識により適切な認識結果が得られ，かつ自然性が高いことを満たす音声を選択することを目指している．そこで，工程①で保存した全ての候補音声に対して自動音声認識を行い，入力テキストとの文字誤り率 (Character Error Rate; CER) を算出する．ここで，CER が最小となる音声を機械が読み上げ内容を最も正確に認識できる音声として採用する．なお，CER の最小値を記録する候補が複数存在する場合は，第 2 の指標として UTMOS [15] を用いた品質評価を行いスコアが最も高い音声を最終的なりすまし音声として採用する．UTMOS は入力音声に対して 1.00（音質が悪い）から 5.00（音質が良い）のスコアで人間が主観評価を行った場合の MOS 値を推定するため，UTMOS が高い音声は人間にとって自然に感じられ騙すことができる可能性が高いことを意味している．

3.2 雑音環境下における前処理

Zero-Shot TTS において，参照音声の音質は，出力される合成音声の品質および話者再現性に大きな影響を与える．正規話者の発話を不正収録した音声データには，空調音や周囲の話し声といった背景雑音が混入している場合があるため，そのまま参照音声として用いた場合，モデルが雑音成分を話者の特徴の一部として学習したり，抽出される話者埋め込みベクトルの精度が低下したりすることで，合成音声の明瞭性や話者類似度が損なわれる問題がある．そこで本研究では，Zero-Shot TTS への入力の前処理として図 2 に示すように音声強調を適用し，合成品質の向上を図る．音声強調を適用することにより，一般的に高品質な音声合成に用いられないような雑音環境下で収録された音声であっても話者性を損なうことなく明瞭な特徴量を抽出し，高い精度でなりすましが可能な音声を作成できると期待される．

4. 実 験 条 件

4.1 なりすまし音声の作成

本研究では Zero-shot TTS モデルの学習に J-SPAW の 50 発話分の不正収録音声のうち，評価に用いない 45 発話分の音声を用いた．学習に用いた不正収録音声の内訳として，4 つの収録環境（E1：静かな室内，E2：空調動作の屋内，E3：音楽の流れている屋内，E4：静かな屋外）があり，話者数は 40 となっている．不正収録は正規ユーザの発声を 1 メートルほど離れた場所で iPad により収録されている．比較検証を行う Zero-shot TTS モデルとして，日本語対応しており，かつ自然な合成が可能な CosyVoice2 [16]，ElevenLabs [17]，VALL-E X [18] の 3 種を用いた．TTS モデルごとに適切な入力長や特性が異なるため，参照音声の入力長を表 1 に示す．入力音声は長さを確保するため Silero VAD [19] を用いて実発話音声から無音区間を除去した後，複数の発話を連結して作成した．出力音声には Silero-VAD を用い，各モデルで表 1 に示す数をデータプールに保存した．CosyVoice2 は出力結果が比較的安定しており 5 つの候補で十分であったが，ElevenLabs，VALL-E X は出力音声の安定性が低いため，より多くの候補数を用意する必要があった．工程②で CER を計算する際には Whisper [20] を用いた．不正収録音声へ音声強調を適用する際には，DPTNet，DCCRNet の事前学習モデル [21], [22] を用いた．最終的に，40 話者 × 4 環境 × 5 種類のテキスト（計 800 音声）を，各 TTS モデルおよび音声強調の有無の組み合わせごとに作成し，評価データとした．

表 1: 工程での各 TTS モデルにおける入出力設定

モデル名	参照音声長 (入力)	生成候補数
CosyVoice2	約 5 秒	5
ElevenLabs	約 90 秒	15
VALL-E X	5 ~ 10 秒	10

4.2 評 価 指 標

本節では，作成したなりすまし音声の品質および攻撃性能を多角的に評価する．まず，音声の品質評価のために UTMOS strong を用い，参照音声への音声強調の有無（なし，DCCRN，DPTNet）および各 Zero-shot TTS モデルの組み合わせについて，作成した合成音声の平均 UTMOS スコアを算出した．次に，なりすまし検出に対する攻撃性能の評価として，最先端モデルの一つである wav2vec2+AASIST [3] を用いた．なりすまし音声検出の評価指標には，なりすまし音声誤受理率と実発話音声誤棄却率が等しくなる点である Equal Error Rate for Countermeasure（$\mathrm{EER}_{\mathrm{cm}}$）を用いた．本実験において，$\mathrm{EER}_{\mathrm{cm}}$ が高いほどなりすまし音声攻撃が成功しており，検出が困難な脅威のある音声であると解釈できる．

また，話者照合モデルには，最先端モデルの一つである ECAPA-TDNN [4] を用いた．評価用のトライアルデータとして，J-SPAW の話者照合用の 800 個の実発話音声と本実験で作成した合成音声 800 音声を用い，同一の実話者音声同士：7,600 ペア，異なる実話者音声同士：30,000 ペア，合成音声と目標話者の実発話音声：16,000 ペアの 3 種類のペアを作成した．これら合計 53,600 ペアを用い，本人誤棄却率と他人誤受理率が等しくなる点である EER for Automatic Speaker Verification（$\mathrm{EER}_{\mathrm{asv}}$）を算出した．実発話音声のみのトライアルデータと比較して $\mathrm{EER}_{\mathrm{asv}}$ が

表 2: UTMOS による主観評価結果

合成手法		CosyVoice2			ElevenLabs			VALL-E X		
	音声強調	なし	DCCR	DPT	なし	DCCR	DPT	なし	DCCR	DPT
J-SPAW	静かな室内 (E1)	2.56	3.01	3.21	2.47	3.05	3.20	2.08	2.39	2.52
	空調のある室内 (E2)	2.11	3.30	3.25	2.02	3.27	3.16	1.68	2.60	2.64
	音楽のある室内 (E3)	1.73	3.23	2.93	1.63	3.18	2.90	1.47	2.50	2.55
	静かな屋外 (E4)	2.90	3.74	3.68	2.68	3.62	3.46	2.17	2.83	2.90
	全環境	2.32	3.32	3.27	2.20	3.28	3.18	1.85	2.58	2.65

表 3: EER_{cm} （%）↑（参考）ASVspoof2021LA データ：0.82

合成手法		CosyVoice2			ElevenLabs			VALL-E X		
	音声強調	なし	DCCR	DPT	なし	DCCR	DPT	なし	DCCR	DPT
J-SPAW	静かな室内 (E1)	15.56	13.94	14.50	32.00	19.00	16.00	10.50	12.00	16.00
	空調のある室内 (E2)	13.94	12.50	14.50	27.56	16.38	13.38	2.88	10.50	13.38
	音楽のある室内 (E3)	12.94	9.50	10.00	27.56	20.88	15.00	7.00	4.50	15.00
	静かな屋外 (E4)	15.56	15.00	15.50	31.50	28.13	19.56	3.50	7.00	19.56
	全環境	14.75	12.88	13.75	29.00	21.13	15.63	6.75	9.25	15.63

上昇していれば，話者照合モデルが合成音声を本人であると誤認していることを示し，なりすまし攻撃として有効であるという結果となる．

最後に，なりすまし音声検出と話者照合を統合したシステム全体の性能評価として，tandem Detection Cost Function (t-DCF) [23] を算出した．t-DCF は，なりすまし音声検出システムと話者照合システムが直列に結合された運用シナリオを想定した指標である．本実験では，各条件における t-DCF を算出することで，提案手法によるなりすまし音声がセキュリティシステム全体に対してどの程度の脅威を与えるかを総合的に評価する．

5. 実 験 結 果

5.1 合成音声の自然性評価 (UTMOS)

表 2 に，各条件における合成音声の UTMOS スコアを示す．各列に音声合成手法と合成に用いた実発話音声への音声強調の有無，各行に評価対象の合成音声の作成に用いた実発話音声の収録環境を記載している．まず，音声強調を用いない場合の結果に着目する．TTS システム間での比較では，CosyVoice2 と ElevenLabs が高い品質を達成した一方，VALL-E X はやや低い水準に留まった．これはモデルの事前学習データの違いや，日本語処理能力の差に起因すると考えられる．環境間では，E1 (静かな室内)，E4 (静かな屋外) で全ての合成手法で比較的高いスコアを記録しているが，背景雑音の多い E2 (空調のある室内)，E4 (音楽のある室内) ではスコアが低下する傾向がみられた．これは TTS システムが入力音声の背景雑音も学習することで出力された合成音声にも背景雑音が混ざり，明瞭性が低下しているためであると考えられる．次に，音声強調（DCCRN, DPTNet）を適用した場合の結果を見ると，全ての TTS システムおよび環境において，スコアの大幅な向上が確認された．特筆すべきは，背景雑音を多く含む E2 や E3 であっても，音声強調を適用することで背景雑音の少ない E1，E4 と同等，あるいはそれ以上のスコアまで品質が改善されている点である．以上の結果より，実環境での不正収録音声を Zero-shot TTS の入力とする場合，前処理としての音声強調が自然性に極めて有効であることが示された．

5.2 なりすまし音声検出

表 3 に，なりすまし検出システムに対する EER_{cm} の評価結果を示す．本指標は値が高いほど，なりすまし音声検出システムを欺くことに成功している（攻撃性能が高い）ことを意味する．まず，音声強調を用いない条件における結果について述べる．TTS モデル間での比較を行うと，CosyVoice2 および ElevenLabs を用いた場合，全環境の音声を評価した場合でそれぞれ 14.75%，29.00% という高い EER を記録した．ASVspoof2021LA 評価データによる wav2vec2+AASIST の EER_{cm} は 0.82 と報告されている．評価対象のデータが異なるので単純な比較はできないものの，本研究の提案手法によるなりすまし音声が高いなりすましの精度を有していることを示している．一方，VALL-E X は平均 6.75% と相対的には低い値に留まった．環境間に着目すると，E1（静かな室内）や E4（静かな屋外）といった背景雑音の少ない環境では，EER が高くなる（検出が困難になる）傾向が見られた．続いて，音声強調を適用した条件における結果について述べる．全体的な傾向として，UTMOS による品質評価は向上したにもかかわらず，音声強調を用いることで EER_{cm} は低下する（検出されやすくなる）傾向が見られた．TTS モデル間で見ると，音声強調適用によるスコアの低下幅は ElevenLabs で特に顕著であった．環境間での比較においても，全ての環境条件で一貫してスコアの低下が見ら

表 4: EER_{asv} （%）↑（参考）J-SPAW 実発話音声：1.69

	合成手法	CosyVoice2			ElevenLabs			VALL-E X		
	音声強調	なし	DCCR	DPT	なし	DCCR	DPT	なし	DCCR	DPT
J-SPAW	静かな室内 (E1)	9.87	8.17	7.67	6.66	6.86	6.73	3.19	3.09	3.14
	空調のある室内 (E2)	9.57	7.17	7.16	6.18	6.33	6.26	2.86	3.07	3.13
	音楽のある室内 (E3)	8.30	5.94	5.67	4.88	5.34	5.22	2.52	2.76	2.95
	静かな屋外 (E4)	10.02	8.17	8.08	6.21	6.86	6.65	3.09	3.25	3.24
	全環境	13.98	10.78	10.50	8.78	9.23	9.03	3.67	3.85	3.91

れた．特に，音声強調によって聴覚上のノイズが除去されクリアになったはずの E2 や E3 においても，EER は改善せずむしろ低下している場合がある．この要因については第 6. 節にて詳細な考察を行う．

5.3 話 者 照 合

表 4 に，話者照合システムに対する EER_{asv} の結果を示す．4. 2 節で述べたトライアルデータのうち，J-SPAW の実発話音声 37,600 ペアのみの評価結果は EER_{asv} = 1.69 である．音声強調を用いない条件について着目する．TTS モデル間での比較では，CosyVoice2 が最も高い傾向を示し，次いで ElevenLabs，VALL-E X の順となり，どの合成手法とも J-SPAW に比べて高い EER_{asv} を示していることから話者照合に対しても突破能力が高いことがわかる．環境間での比較では，なりすまし音声検出の結果と同様に，E1 や E4 といった静かな環境の方がスコアが高く，E3 などの雑音環境ではスコアが低い傾向が見られた．これは，参照音声に背景雑音が多く含まれる場合，生成された音声の話者特徴が背景雑音に埋もれ，照合精度が低下するためと考えられる．音声強調を適用した条件について着目する．TTS モデル間で見ると，ElevenLabs および VALL-E X においては，音声強調を用いることで EER_{asv} のわずかな上昇が見られた．一方，CosyVoice2 においては，逆に EER_{asv} が低下する傾向が見られた．環境間での比較を見ると，ElevenLabs と VALL-E X は音声強調によって E2 や E3 の EER_{asv} が改善しているが，環境間でのスコアの大小の傾向はあまり変わらないことが確認できる．音声強調により背景雑音が低減しているが，背景雑音低減の際に話者性を歪ませてしまう処理が含まれてしまい，話者情報を維持できていないことを示唆している．

最後になりすまし音声検出と話者照合を組合せた際の性能評価指標である t-DCF で評価した結果を表 5 に示す．wav2vec+AASIST の ASVspoof2021LA 評価データでの min t-DCF は 0.21 と報告されている．評価データの違いから単純比較はできないものの，提案手法による合成音声の t-DCF は全条件で条件でも ASVspoof2021LA 評価データのスコアを上回っていることから，システム全体に対して突破能力が高いことが確認された．

6. 検証実験：音声強調がなりすまし検出に与える影響

前節の結果より，音声強調の適用は UTMOS スコアを向上させる一方で，なりすまし検出に効果的でないことが判明した．この原因として，以下の 2 つの仮説が考えられる．

仮説 (1) なりすまし音声検出システムが，背景雑音の有無そのものを自然性の指標として利用しており，強調処理によるノイズ除去を不自然と判断した．

仮説 (2) 音声強調により背景雑音が除去されたことで，合成音声らしさを決定づける何らかの特徴が顕在し，なりすまし音声検出システムが検知しやすくなった．

この要因を解明するため，CosyVoice2 を用いて追加実験を行った．音声強調を行った実発話音声から合成したなりすまし音声に対し，強調前の実発話音声の背景雑音を再付加し，EER_{cm} の変化を測定した．もし仮説 1 が主たる要因であれば，ノイズの再付加により EER_{cm} は音声強調なしの場合と同程度まで回復すると予想できる．実験結果を表 6 に示す．結果として，音声強調ありのなりすまし音声にノイズを再付加した場合の EER_{cm} は，音声強調なしの場合の水準まで回復せず，多くの環境で音声強調あり，ノイズ再付加なしの場合に比べ低い値となった．単純にノイズを付加するだけではなりすまし音声検出システムを突破することは困難である考えられる．例外として，E3（音楽のある室内）ではノイズ付加によるスコア改善が見られた．要因として音楽が合成音声の波形に生じている何らかの特徴をマスキングしており，なりすまし検出の精度を低下させる方向に働いたと考えられる．以上のことから，音声強調による EER_{cm} 低下の主要因は仮説 2 であり，かつ音声強調なしの場合における高い EER_{cm} は，Zero-shot TTS が生成過程で背景雑音と発話を一体として生成することにより，合成音声らしい特徴がマスキングされていたと結論付けられる．

7. お わ り に

本研究では，日常的な環境からの不正収録を想定した

表 5: min t-DCF（%）の比較結果．（参考）ASVspoof2021LA：0.21

	合成手法	CosyVoice2			ElevenLabs			VALL-E X		
	音声強調	なし	DCCR	DPT	なし	DCCR	DPT	なし	DCCR	DPT
J-SPAW	全環境	0.45	0.45	0.46	0.99	0.69	0.56	0.41	0.47	0.40

表 6: ノイズ付加有無による EER_{cm}（%）↑

合成手法	CosyVoice2				
ノイズ付加	なし			あり	
音声強調	なし	DCCR	DPT	DCCR	DPT
静かな室内 (E1)	15.56	13.94	14.50	9.06	10.44
空調のある室内 (E2)	13.94	12.50	14.50	9.38	9.00
音楽のある室内 (E3)	12.94	9.50	10.00	12.50	11.63
静かな屋外 (E4)	15.56	15.00	15.50	14.50	13.00
All	14.75	12.88	13.75	11.75	11.00

音声から Zero-shot TTS を基盤としてなりすまし音声を作成し，なりすまし音声検出，話者照合，音声の品質評価という 3 つの観点から有用性を評価した．また不正収録を想定した音声に対し音声強調を適用し同様に評価することでノイズ除去が評価結果に及ぼす影響についても考察した．結果として強力な音声セキュリティシステムに対しても本研究で作成した音声が高い効果を発揮することが確認された．一方音声強調を適用することによりなりすまし音声のとしての有用性が一様高まるとはいえないことも確認できた．今後の展望として，本研究で作成した音声を用いてなりすまし音声検出モデルや話者照合モデルを学習し，音声セキュリティ向上を目指すことなどが挙げられる．

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] G. Pei, J. Zhang, M. Hu, G. Zhai, C. Wang, Z. Zhang, J. Yang, C. Shen, and D. Tao, “Deepfake Generation and Detection: A Benchmark and Survey,” in Proc. *CoRR*, 2024.

[2] M. Li, Y. Ahmadiadli, and X.-P. Zhang, “A Survey on Speech Deepfake Detection,” in Proc. *ACM Comput. Surv.*, vol. 57, no. 7, pp. 1–38, 2025.

[3] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, “Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation,” in Proc. *Odyssey*, 2022.

[4] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, “ECAPA-TDNN embeddings for speaker diarization,” in Proc. *INTERSPEECH*, pp. 3560–3564, 2021.

[5] E. Jamdar and A. K. Belman, “SyntheticPop: Attacking Speaker Verification Systems With Synthetic VoicePops,” *arXiv:2502.09553*, 2025.

[6] A. Kassis and U. Hengartner, “Breaking Security-Critical Voice Authentication,” in Proc. *SP*, pp. 951–968, 2023.

[7] Y. Wu, J. weon Jung, H. jin Shim, X. Cheng, and X. Wang, “WildSpoof Challenge Evaluation Plan,” 2025.

[8] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, Y. Liu, S. Zhao, and N. Kanda, “E2 TTS: Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS,” in in Proc. *SLT*, pp. 682–689, 2024.

[9] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” pp. 2709–2720, 2022.

[10] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, *et al.*, “XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model,” in Proc. *INTERSPEECH*, 2024.

[11] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, *et al.*, “CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens,” in Proc. *CoRR*, 2024.

[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” Proc. *INTERSPEECH*, 2020.

[13] J. Chen, Q. Mao, and D. Liu, “Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation,” Proc. *INTERSPEECH*, 2020.

[14] S. Shiota, S. Horie, K. Kanno, S. Takamichi, “J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset,” in Proc. *INTERSPEECH*, 2025. (accepted).

[15] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “UTMOS: UTokyo-Sarulab System for VoiceMOS Challenge 2022,” in Proc. *INTERSPEECH*, pp. 4521–4525, 2022.

[16] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, *et al.*, “CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models,” in Proc. *CoRR*, 2024.

[17] https://elevenlabs.io//.

[18] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, “Speak foreign languages with your own voice: Cross-lingual neural codec language modeling,” *arXiv preprint arXiv:2303.03926*, 2023.

[19] https://github.com/snakers4/silero-vad/.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[21] https://huggingface.co/JorisCos/DPTNet_Libri1Mix_enhsingle_16k.

[22] https://huggingface.co/JorisCos/DCCRNet_Libri1Mix_enhsingle_16k.

[23] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification,” in in Proc. *Odyssey*, pp. 312–319, 2018.