なりすまし音声検出に対する話者適応を用いた音声合成攻撃*

☆古林 嵯羽仁（東京都立大）, 高道 慎之介（慶応義塾大/東京大）,
塩田 さやか （東京都立大）

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が増えてきている．例えば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報告されており，対策が急務となっている．

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 [1]，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 [2, 3] 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 [4, 5]．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．

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てきた．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としての視点から，日常的な環境で不正録音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，不正録音を想定してなりすまし音声が収録された音声コーパスである J-SPAW [6] を用いて話者適応を含む TTS モデルを追加学習することで目標話者を模倣した合成音声を作成し，なりすまし音声検出，話者照合，音声品質評価という 3 つの観点から評価を行った．実験結果より，特定の環境下で不正録音した音声から合成した音声を用いた場合，追加学習によりなりすまし音声としての有用性を高めることができた．また不正録音音声を音声合成に用いた場合でも目標話者へわずかに話者性を近づけることが可能であることが確認できたことを報告する．

2 関連研究

2.1 なりすまし音声攻撃となりすまし音声検出

合成音声によるなりすまし音声攻撃は，TTS・声質変換 [7] などにより特定話者の声質を模倣した音声を生成し，システムに対してなりすましを行う攻撃のことである．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，なりすまし音声であるかを判定する．なりすまし音声の話者性を評価するためには話者照合が用いられ，照合対象の音声と特定話者の実発話音声との類似度から同一人物であるか否かが判定される．

2.2 話者適応を用いたテキスト音声合成

合成音声を作成する手法の一つとして TTS がある．TTS とは，所望のテキストを音声合成モデルへ入力することでテキストに従った合成音声を作成する技術である．TTS における話者性は合成モデルをどのデータで学習したかに大きく依存するため，話者性についての自由度は低い．そこで，話者適応技術により合成モデルの持つ話者性から所望の話者への話者性の変換を行うことが多い．話者適応は，目標話者の発話データが少量でも話者性を再現できるため，なりすまし音声攻撃でも重要な技術となっている．深層学習に基づく合成モデルでの話者適応では，x-vector [8] のような話者照合で用いられる話者埋め込みベクトルを用いた手法 [9] などが提案されている [10].

2.3 J-SPAW

J-SPAW 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 種類読み上げてもらい，スマートフォンで録音した音声が収録されている．本研究では J-SPAW に収録された不正録音音声を音声合成に用いることで，なりすまし音声攻撃を行う攻撃者が不正録音音声から合成されたなりすまし音声による攻撃を行うことを再現している．

3 提案手法

先行研究 [9] では，x-vector を用いて TTS モデルである FastSpeech2 [11] に話者適応を行い，目標話者に似せた合成音声を生成する手法が取られている．ここで，x-vector とは話者照合に用いられる，深層学習に基づいた話者埋め込み抽出器から出力されるベクトルであり，x-vector を用いることで目標話者に固有な話者情報を得ることが可能となる．本研究では先行研究で用いられた音声合成手法を基にした 2 つのなりすまし音声作成手法を用いる．１つ目の手法は先行研究にて事前学習された FastSpeech2 に対して J-SPAW の不正収録音声を用いて追加学習を行い，

---

*Attacks Based on Speaker Adaptation-Based Speech Synthesis Against Spoofing Speech Detection Systems. by Furubayashi Sawato (Tokyo Metropolitan University, Takamichi Shinnosuke (Keio University/Universityof Tokyo), Shiota Sayaka (Tokyo Metropolitan University)

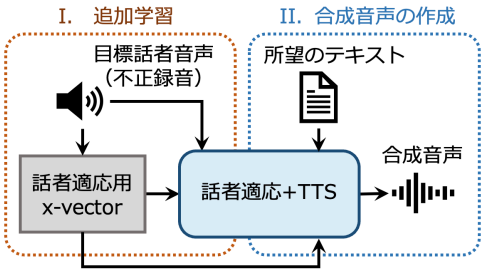

Fig. 1　話者適応による音声合成（ノイズ除去なし）

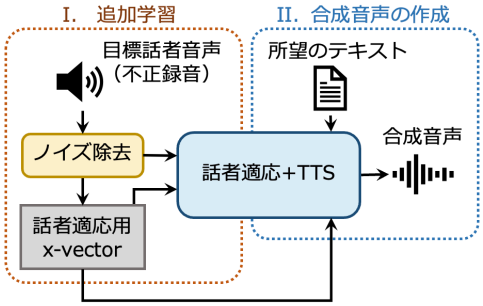

Fig. 2　話者適応による音声合成（ノイズ除去あり）

追加学習済モデルに対して不正録音音声から抽出した x-vector を用いて話者適応を行うことで実話者の話者性を模倣した合成音声を作成する手法（ノイズ除去なし：Fig. 1）である．2 つ目は 1 つ目の手法のうち追加学習に用いる不正録音音声にノイズ除去を適用した音声を用いる手法（ノイズ除去あり：Fig. 2）である．2 つ目の手法では話者適応に用いる x-vector の抽出もノイズ除去適用後の音声を用いる．x-vector を用いた TTS モデルの話者適応は通常，モデルパラメータを固定したままで目標話者の x-vector を入力することを指す．一方で上記の 2 手法は，目標話者の音声がある程度の量だけ入手できることを前提に，モデルパラメータを更新する．ノイズ除去を行うことで高品質な合成音声生成には用いられないような背景雑音が多く含まれる音声からも品質の高い合成音声を作成することができるのかを検証する．

4　実験条件

4.1　合成音声の作成

合成音声の作成条件について述べる．話者適応用 x-vector の作成，FastSpeech2 の追加学習には J-SPAW の不正録音音声を用いた．不正録音音声の内訳として，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，発話数は 45 となっている．不正録音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．このうち本研究では音声合成に使用する実発話音声の収録環境の組み合わせとして 4 環境全て（All），比較的静かな環境である E1 と E4 の 2 環境（E1+E4），E1 のみを用いた場合（E1），E4 のみを用いた場合（E4）の 4 パターンを用

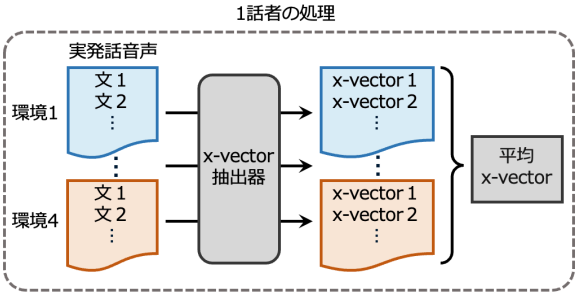

Fig. 3　1 話者分の x-vector の計算方法

いた．ノイズ除去を行う場合には不正録音音声に対して深層学習に基づく音声復元手法である Miipher [12] のオープンソースモデル[1]あるいは信号処理に基づくノイズ除去手法として spectral subtoraction (SS) [13] を適用した．話者適応に用いる x-vector の作成フローを Fig. 3 に示す．まず，ある目標話者の不正録音音声について各収録環境，各発話それぞれの音声から xvector_jtubespeech[2]によって x-vector を抽出し，話者内全ての発話の平均を計算することで 1 話者分の平均 x-vector とする．他の話者においても同様に平均 x-vector を作成し，各話者それぞれの平均 x-vector と FastSpeech2 で話者適応を行う．FastSpeech2 の事前学習モデルと追加学習手法は文献 [9] に従う．事前学習の 400000Step から 800000Step まで学習を進めた内，評価ロスが最も低くなる段階のモデルを追加学習モデルとした．FastSpeech2 では平均 x-vector の作成，追加学習に含まれていない 5 文を合成する文として入力し，200 個（40 話者× 5 文）の合成音声をなりすまし音声として作成した．

4.2　合成音声の評価

なりすまし音声検出モデルとしては最先端モデルの一つである文献 [2] のモデルを用いた． 3 章で述べた話者適応による音声合成のノイズ除去なし，ありそれぞれに対して 4.1 節で述べたそれぞれの不正録音音声のパターンごとに作成した合成音声と J-SPAW の実発話音声を用いてなりすまし音声検出の性能評価を行った．評価指標はなりすまし音声誤受理率と実発話音声誤棄却率の等しい点である Equal error rate for countermeasure（EER_{cm}）である．本実験においては，なりすまし音声検出の EER_{cm} が高いほどなりすまし音声攻撃に成功しており，なりすまし音声攻撃を想定した検出が困難な合成音声として期待される結果となる．

話者照合モデルとしては最先端モデルの一つである ECAPA-TDNN [3] 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 200 音声を用い，同一の実話者音声どうしによる 7,600 ペア，異なる実話者音声どうしによる 30,000 ペアに合成音声と目標話者の実発話音声による 4,000 ペアを加え，合計

[1]https://github.com/Wataru-Nakata/miipher
[2]https://github.com/sarulab-speech/xvector_jtubespeech

Table 1　EER_{cm}（%）と UTMOS スコア

ノイズ除去	環境	EER_{cm} ↑ 追加学習 なし	EER_{cm} ↑ 追加学習 あり	UTMOS ↑ 追加学習 なし	UTMOS ↑ 追加学習 あり
なし	All	2.50	2.13	2.52	1.49
	E1+E4	2.56	2.13	2.69	1.73
	E1	3.00	**6.56**	2.52	1.63
	E4	2.13	2.50	2.47	1.58
あり Miipher	All	3.00	2.50	2.88	2.98
	E1+E4	3.94	3.63	2.89	3.06
	E1	3.00	3.00	2.89	**3.10**
	E4	2.13	3.00	2.82	2.97
あり SS	All	2.44	2.00	2.63	1.58
	E1+E4	2.06	2.00	2.68	1.94
	E1	2.94	2.00	2.62	1.80
	E4	2.50	1.06	2.68	1.77

41,600 ペアを作成した．評価指標は本人誤棄却率と他人誤受理率の等しい点である EER for automatic speaker vertification（EER_{asv}）である．本実験においては EER_{asv} について，実発話音声のみのトライアルデータで話者照合を行なった場合と比較して値が上昇していれば話者照合モデルが合成音声と実話者が同一人物であると判定したペアが存在することになり，なりすまし音声としては望ましい結果となる．加えて話者照合を行った際の類似度を表すスコアの分布を示すことで，実発話同士でのスコア分布と合成音声と実発話を照合した際のスコア分布を比較し分析する．

本実験では音声の品質評価のために合成音声品質自動評価システムである UTMOS [14] を用いた．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 値を推定するものである．作成した合成音声に対しなりすまし音声検出を行った結果の EER_{cm} と UTMOS により出力されたスコアの関係も合わせて分析する．

5　実験結果

Table 1 に本実験で作成した合成音声に対するなりすまし音声検出の EER_{cm} と UTMOS のスコアを示す．事前学習済 FastSpeech2 に話者適応のみを行い作成した合成音声の評価結果を追加学習なし，4.1 節で示した 4 つのパターンごとに追加学習した FastSpeech2 に話者適応を行い作成した合成音声の評価結果を追加学習ありとして示している．また，ノイズ除去なしと Miipher，SS それぞれをノイズ除去に用いた場合の結果も示している．Table 1 より，全環境の音声を用いた All と E1+E4 はノイズ除去の有無に関わらず全ての合成手法で追加学習によって EER_{cm} が低下することが確認された．一方でノイズ除去なしとノイズ除去あり (Miipher) において E1，E4 それぞれ 1 環境のみの音声を用いた場合は EER_{cm} が上昇する傾向にあることが確認できる．ノイズ除去に SS を適用した場合はいずれの環境においても追加学習によって EER_{cm} が低下している．SS は Miipher と比較してノイズ除去性能が高くないため，追加学習の効果も低いことがわかる．ノイズ除去の適用により EER_{cm} が上昇する傾向が見られないことから，本実験で用いたノイズ除去手法ではノイズを含む音声から高精度なりすまし音声を作成することは難しいと考えられる．UTMOS のスコアに着目すると，ノイズ除去あり（Miipher）ではノイズ除去なしに比べて追加学習ありなしの共にスコアが高くなることから，Miipher をノイズ除去を適用することによる合成音声の品質の向上が確認できるが，ノイズ除去あり（SS）では低下することがわかる．EER_{cm} と UTMOS スコアの関係からなりすまし音声検出では音声品質が直接なりすまし音声としての有用性に影響しないことが読み取れる．

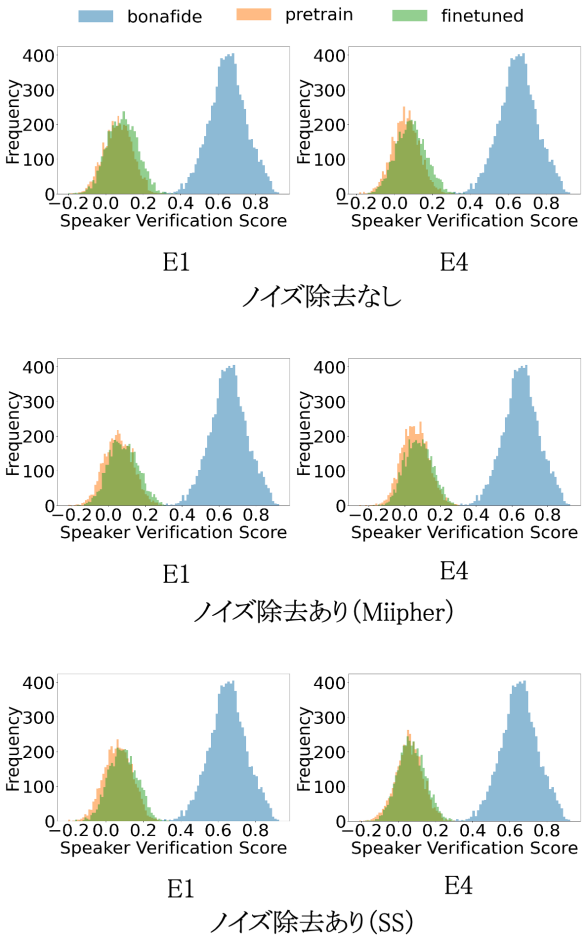

Fig. 4　話者照合スコア分布

話者照合の評価結果について，実発話のみのトライアルデータで話者照合を行なった時と比較して EER_{asv} はほとんど変化が見られなかったことから，Fig. 4 にノイズ除去なし，ノイズ除去あり（Miipher），ノイズ除去あり（SS）それぞれにおいて追加学習あ

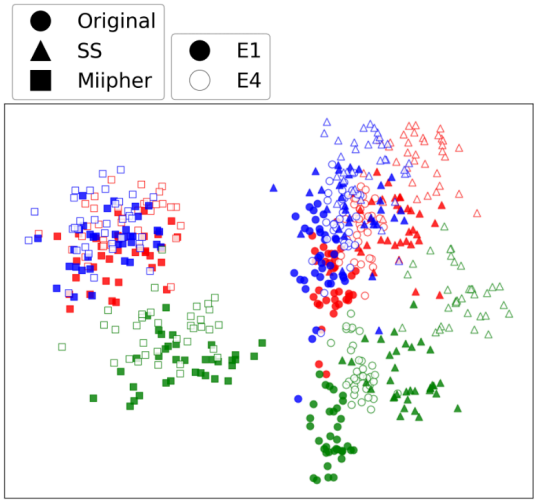

Fig. 5　主成分分析による x-vector の分布

りなしでの合成音声と目標話者の実発話音声のペア，同一話者の実発話ペアそれぞれの類似度を表すスコアの分布を可視化したヒストグラムを示す．音声合成に用いた不正録音音声は E1 と E4 である．追加学習した FastSpeech2 を用いて作成した合成音声のスコア分布は事前学習モデルを用いて作成された合成音声のスコア分布に比べて実発話音声のスコア分布に若干接近している傾向があることがわかる．このことから不正録音を想定した音声を用いた場合でも FastSpeech2 の追加学習により実話者に話者性を近づけることができると言える．

J-SPAW の 3 話者の不正録音音声について，ノイズ除去なし，Miipher 適用後の音声，SS 適用後の音声それぞれから抽出した話者適応用 x-vector に主成分分析を行った結果を Fig. 5 に示す．不正録音音声に Miipher を適用した音声から抽出した x-vector はノイズ除去なしの場合と SS を適用した音声から抽出した x-vector と異なる位置に分布している傾向が見られることから，Miipher を適用することによりノイズは非常に高精度に除去されるが話者性が大きく変化してしまっていることがわかる．しかし Fig.4 でノイズ除去あり（Miipher）のスコア分布がノイズ除去なし，ノイズ除去あり（SS) のスコア分布と比較し大きな違いがないことから，Miipher による不正録音音声の話者性の変化によりもたらされる影響を本実験の結果からは確認できなかった．

6　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音声から TTS モデルへの追加学習と深層学習による話者埋め込み x-vector を用いた話者適応を行うことで合成音声を作成し，なりすまし音声検出，話者照合，音声の品質評価という 3 つの観点からなりすまし音声としての有用性を評価した．また不正収録を想定した音声に対しノイズ除去を行い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．結果として追加学習によるなりすまし音声攻撃の成功可能性の上昇は一様な傾向として確認することは出来なかったが，追加学習により目標話者へ話者性を近づけることができることが確認された．またノイズ除去を行うことによる合成音声の品質の改善が見られた．今後の課題として話者照合モデルやなりすまし音声検出の突破を目的とした新たな損失関数の設定を行い追加学習によりなりすまし音声攻撃の成功可能性を安定して上昇させることや，多様なノイズ除去システムを適用し効果を比較することが挙げられる．

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．

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] E. Jamdar and A. K. Belman, “SyntheticPop: Attacking speaker verification systems with synthetic voicepops,” *arXiv:2502.09553*, 2025.

[5] A. Kassis and U. Hengartner, “Breaking security-critical voice authentication,” in Proc. *SP*, pp. 951–968, 2023.

[6] 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).

[7] X. Zhang, Y. Wan, and W. Wang, “Dimensional affective speech synthesis based on voice conversion,” Trans on. *Intelligent Computing*, vol. 3, 2024.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proc. *ICASSP*, pp. 5329–5333, 2018.

[9] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, “Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,” in Proc. *ICASSP*, pp. 1–5, 2023.

[10] Y. Fan, Y. Qian, F. K. Soong, and L. He, “Multi-speaker modeling and speaker adaptation for DNN-based tts synthesis,” in Proc. *ICASSP*, pp. 4475–4479, 2015.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in Proc. *ICLR*, 2020.

[12] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, “Miipher: A robust speech restoration model integrating self-supervised speech and text representations,” pp. 1–5, 2023.

[13] R. Scheibler, E. Bezzam, and I. Dokmanić, “Pyroomacoustics: A python package for audio room simulation and array processing algorithms,” in Proc. *ICASSP*, pp. 351–355, 2018.

[14] 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.