TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　プライバシー保護と偽音声検出を統合する音声データ処理基盤
英文：Unification of speech deepfake detection and speaker privacy protection

利用課題責任者: WANG XIN
First name Surname: Xin Wang

所属: 国立情報学研究所
Affiliation: National Institute of Informatics
URL https://www.nii.ac.jp/

邦文抄録　本研究は、高度化する音声ディープフェイクに対し、利便性とセキュリティを両立する検知技術の確立を目的とする。今年度は主に2つの成果を得た。第一に、実環境の会話等を含む新データセット「WildSpoof」を構築し、国際コンペを通じて大規模データと基盤モデルの重要性を立証した。第二、MLOps の観点からデータドリフトを監視し、検知器を動的に微調整する運用手法を提案した。今後は、未知の生成モデルに対するさらなる汎用性向上を目指す。

英文抄録　This research establishes speech deepfake detection technologies that balance convenience and security. Key achievements include: 1) Developing the WildSpoof dataset and hosting a competition to validate the role of foundation models; 2) Proposing an MLOps framework to monitor data drift and maintain detector accuracy. While effective on in-domain data, future work will focus on improving robustness against unknown generative models.

Keywords: speech deepfake detection, watermark, deep learning, speech information processing

背景と目的[1]

近年、音声生成 AI 技術の飛躍的な進歩により、特定の個人の声を極めて精巧に模倣・再現できる「音声ディープフェイク」の生成が容易になりました。これに伴い、なりすまし詐欺やフェイクニュースによる情報捏造といったセキュリティ上の脅威が世界的に深刻化しており、信頼性の高い検知技術の確立が急務となっています。

しかし、従来の検知技術には大きな課題がありました。無響室のような理想的な環境で収録された音声には高い精度を発揮するものの、街中の騒音や残響が含まれる複雑な実環境、あるいは学習データに含まれない未知の最新攻撃に対しては、検知性能が著しく低下するという点です。また、一度構築した検知器を固定的に運用する「静的な検知」では、日々進化する新たな攻撃手法に対して次第に脆弱になります。運用面における持続的な精度維持や、大規模な学習データの収集コストも、実用化を阻む大きな障壁となっていました。

本プロジェクトでは、これらの課題を解決するため、実環境を模した大規模データセット「WildSpoof」を構築しました。さらに、機械学習の運用サイクル（MLOps）の観点から、未知の攻撃によるデータの乖離を監視する「データドリフト監視手法」を導入しました。これにより、複雑な背景音や未知の生成モデルに対する脆弱性を克服し、実社会の多様なシーンにおいて高い汎化性能と持続的な防御能力を持つ、利便性とセキュリティを両立した音声データ処理技術を確立しました。

本プロジェクトの研究成果の一環として、国際学会 ICASSP にて「WildSpoof Challenge」と題した国際コンペティションを実施しました。本稿では、同会議に採択された最新の研究成果について紹介いたします。

WildSpoof 評価用データと国際コンペティション

音声ディープフェイクの生成と検知という 2 つの音声処理タスクにおいて、実環境（in-the-wild）データの活用を促進することを目的として WildSpoof Challenge を行いました（図1）。

[1] 本文を作成する際、Gemini 3 Pro を用いて添削を行いました。

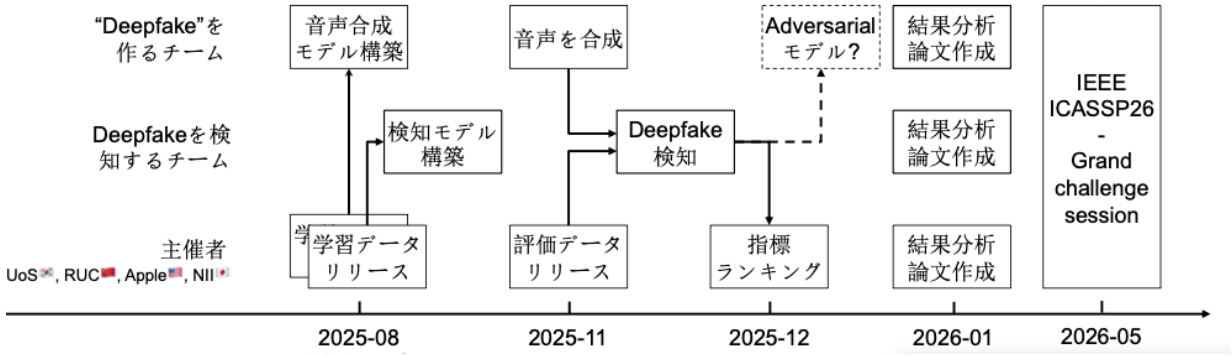

図 1：WildSpoof Challenge の概要

密接に関連しつつ高度化が進む「テキスト読み上げ生成（TTS）」と「なりすましを考慮した自動話者照合（SASV）」の 2 つの重要課題に対し、多くの参加者から協力を得ました。昨年度開発した ASVspoof5 を含む既存のデータセットと、TTS 側の参加者が新たに生成した音声データを組み合わせ、WildSpoof 評価データを構築して SASV の性能評価を行いました。

コンペの結果を図 2 に示します。評価指標である A-DCF は、値が低いほど検知性能が優れていることを示します。2019 年の学習データで構築した検知システム（B2）の結果から、音声合成技術の進化が検知側にとって新たな脅威となっていることが分かります。ここに新しいデータを追加することで検知性能が改善され（B1）、さらに Fake 音声検知の基盤モデル[1]を導入（P1）することで、性能が大幅に向上することを確認しました。また、多くの参加チームがベースライン（B1、B2）を上回り、特に T01 チームは 4 つの評価サブセットのうち 3 つで首位を獲得しました。

WildSpoof Challenge の参加チームが作成した技術レポートや結果の詳細は、公式サイト[2]で公開されています[2]。今回の結果から、単一のデータによる評価では汎用性を正確に測定できず、多様なドメインを用いた評価が不可欠であることも示唆されました。

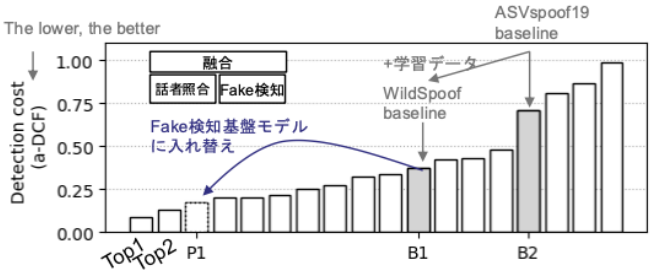

図 2：WildSpoof Challenge の結果

[2] https://wildspoof.github.io/

データドリフト監視による MLOps 手法

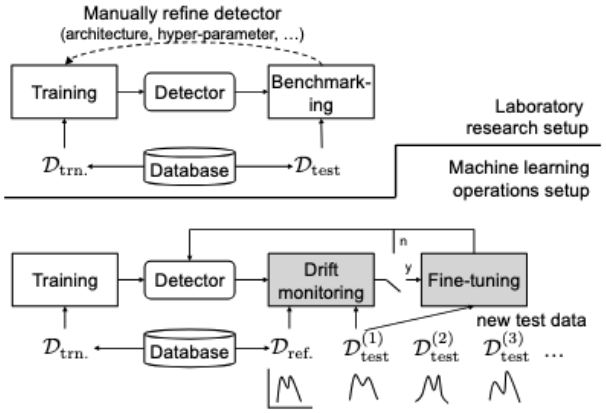

図 3：データドリフトを監視できる音声ディープフェイク検知スシテムの仕組み

WildSpoof Challenge をはじめとする本プロジェクトの多くの研究では、特定の学習データセットと評価セットを用いて性能検証を行っています。しかし、実社会のクラウドサービスやアプリケーションに導入された検知器が、モデルを更新しない「静止」した状態のままであれば、日々新たに生み出される未知の攻撃に対して次第に脆弱になるという課題があります。この脆弱性に対処するため、本研究では機械学習オペレーション（MLOps）の観点から、既存の参照データセットから乖離していく新しい攻撃データを監視・適応できるかという問いを検証しました。

具体的には、未知の攻撃データが参照データからどれほど離れているかを示す「データドリフト」に着目しました。トイデータセットおよび大規模な MLAAD データセットを用いた実験により、最新の音声合成（TTS）攻撃によるドリフトは、新データと参照データの分布間の距離を算出することで監視可能であることを示しました [3]。

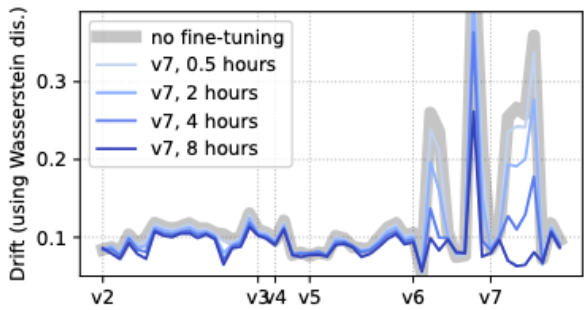

図4：MLAAD（v2-v7）におけるデータドリフト結果

さらに、ドリフトが検知された際に、同様の特性を持つデータを用いて検知器を微調整（ファインチューニング）することで、性能を維持できるかを深掘りしました。解析の結果（図４）、新しい TTS データ (MLAAD v7) [4] による微調整はドリフトを抑制し、検知誤り率を大幅に低減することを実証しました。例えば、特定のバージョン（v7）のデータを 8 時間分用いて微調整を行うだけで、高いドリフト値が顕著に減少し、その効果は微調整に使用するデータ量が増えるほど明確になります。

本研究の成果は、一度構築した検知器を使い続けるのではなく、MLOps を通じて未知の攻撃を早期に捉え、動的にモデルを更新し続ける運用の重要性を示しています。今後は、このサイクルを自動化し、より多様な攻撃環境下での堅牢性を追求していきます。

まとめ、今後の課題

本年度の成果に基づき、今後は以下の課題に取り組みます。

第一に、大規模言語モデル（LLM）と強化学習を統合し、検知の根拠を提示できる「解釈可能な音声ディープフェイク検知モデル」を構築します。これにより、ブラックボックス化しがちな検知プロセスの信頼性を高めます。第二に、MLOps の手法をさらに発展させ、データの分布だけでなくメタ情報も活用した高度なモデル微調整手法を確立し、未知の攻撃への即応性を強化します。

また、利便性とプライバシー保護の両立を目指し、音声の質を保ちつつ個人性を隠蔽する「話者匿名化」技術に関する国際コンペティションの開催も予定しています。これらの取り組みを通じて、安全で信頼できる音声社会の実現に貢献します。

参考文献

[1] Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for Deepfake Speech Detection. In *Proc. ASRU*, 2025. (published)

[2] Yihan Wu, Jee-weon Jung, Hye-jin Shim, Xin Chen, and Xin Wang. 2026. WildSpoof: Advancing In-the-wild Data in Text-to-speech Generation and Spoofing-aware Automatic Speaker Verification. In *Proc. ICASSP*, 2026. (accepted)

[3] Xin Wang, Wanying Ge, and Junichi Yamagishi. 2026. Towards Data Drift Monitoring for Speech Deepfake Detection in the context of MLOps. In *Proc. ICASSP*, 2026. (accepted)

[4] Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. In Proc. IJCNN, June 30, 2024. IEEE, 1–7.