TSUBAME 共同利用　令和6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　インフォデミックの時代において、AI によって生成された音声ディープフェイクの検出や声紋などの個人情報を保護できる音声データ処理基盤は不可欠です。しかし、音声ディープフェイク技術が進化するにつれ、その検出はますます困難になっています。本プロジェクトでは、音声ディープフェイクの検出に着目し、より実環境に近い音声データを用いて、大規模かつ最新の音声ディープフェイクを網羅したデータセットを作成しました。そのデータセットを用いた検証の結果、最新の音声ディープフェイク検出器でも性能が低下することが確認されました。また、音声透かし技術に基づく新しい音声ディープフェイク検出手法についても検討しました。

英文抄録　In the era of infodemics, a speech data processing infrastructure capable of detecting AI-generated fake speech and protecting personal voiceprints is essential. As speech deepfake technology advances, however, detecting speech deepfakes is becoming increasingly challenging. This project focuses on speech deepfake detection and has created a large-scale dataset that covers both the latest and legacy speech deepfake algorithms, using speech data with more realistic acoustic conditions. Through this dataset, it was confirmed that the performance of existing speech deepfake detectors deteriorates. Additionally, we explored a new speech deepfake detection method based on speech watermarking technology.

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

背景と目的[1]

ディープラーニング技術の進歩により、特定の人物の声を模倣し、本物と区別がつかないほど精巧な音声を生成できる AI 技術が数多く開発されました。これらの技術は、ナレーションや音声アシスタントなどの分野で有益に活用される一方、詐欺やなりすましといった悪用のリスクも高まっています。

音声ディープフェイク検出技術は、このようなリスクに対応するために開発されています。主な手法として、畳み込みニューラルネットワーク（CNN）、トランスフォーマーモデル、自己教師あり学習などの深層学習アプローチがよく利用されています。

近年では、より高度なディープフェイク音声が登場し、それに対応する検出技術の精度向上が求められています。そのため、高度なディープフェイク音声を網羅した学習データが必要となります。また、実際の利用環境に近い条件で収録された音声データは、モデルの検出性能向上にも重要であると考えられます。

本プロジェクトでは、大規模な音声ディープフェイクデータセット **ASVspoof5** を作成し、国際的なコンペティションを開催してディープフェイク検出の性能を比較しました [1]。その結果、最新のシステムでも一部のディープフェイクアルゴリズムの検出が依然として困難であることが明らかになりました。Mp3 などによって圧縮されたデータの検出精度が落ちたことも確認されました。

音声ディープフェイク検出に加え、それと組み合わせることが可能な音声透かし（watermarking）を提案しました [2]。この新しい手法では、透かしを埋め込んだ合成音声の検出が可能であることが確認されましたが、背景雑音や圧縮を伴う環境では、検出性能が低下することが判明しました。

本プロジェクトの研究成果は、国際学会で発表されています。また、開発された ASVspoof5 データセットおよび音声透かしの実装も公開されております。

[1] 本文を作成する際、ChatGPT を用いて文法や語彙の添削を行いました。

表1：既存データセットと開発した ASVspoof5 データセットの比較

	話者数	発話数	ディープフェイクの種類	データー収録環境
ASVspoof 2015	106	263,151	10	録音スタジオ
ASVspoof 2019 LA	107	121,461	19	録音スタジオ
ASVspoof 2021 LA	67	164,612	13	録音スタジオ
WaveFake	2	136,085	9	録音スタジオ
ADD 2023	未知	517,068	未知	録音スタジオ
HAD	218	160,836	2	録音スタジオ
CFAD	1302	347,400	12	録音スタジオ
MLAAD	未知	76,000	54	Youtube
ASVspoof 5	**1922**	**1,004,081**	**32**	様々な室内環境

概要

ASVspoof 5データセットの開発

本プロジェクトの成果の一つは ASVspoof 5 データセットの作成です。このデータベースは、LibriVox プロジェクトにおいて適切な許可を得て公開されている、約 2000 人の話者の録音を使用しました。合成音声を生成するためのプロトコルは慎重に設計されており、32 種類の音声合成アルゴリズムを用いて各目標話者の音声データを合成しました。その際、音声クローニングに必要な学習データ量、話者間のデータバランス、さまざまなエンコーディングや圧縮の影響などの要素も考慮しました。

表 1 に示されているように、既存のデータベースと比較して、ASVspoof 5 は話者数だけでなくデータ量の面でも最大規模を誇ります。 また、録音条件もレコーディング・スタジオで作成されたデータより、より実環境に近いものとなっています。

データセットの詳細をまとめた論文は、 *Computer Speech & Language* 誌に投稿済みであり、ArXiv でも公開されています [3]。 また、データセット自体は Zenodo と Hugging Face に公開されており、現時点ですでに 5,000 回以上ダウンロードされています。[2]

音声ディープフェイク検出ための透かし技術の開発

本プロジェクトでは、音声ジェネレーターに埋め込む電子透かし技術を開発しました。 この手法は、協調的フレームワークを用いて音声生成器を訓練することに基づいています。敵対的ネットワーク（GAN）とは異なり、本手法では入力音声が本物か偽物かを識別するための協調的識別器が追加されます。この識別器は音声生成器と同時に Fine-tuning され、学習の目標は、音声生成器によって生成された音声が協調識別器によって適切に識別されるようにすることです。したがって、電子透かしの情報量は 1 ビットとなります。協調識別器は、透かし検出器として機能します。

学習フレームワークは ICASSP で公開され、github でオープンソース化されている。

採択済み論文では、一つの音声生成器を用いて透かしの検出性能を測りましたが、複数の生成モデルにおける透かしの検出性能や識別器の改善について、雑誌論文を作成しています。その結果も公開される予定です。

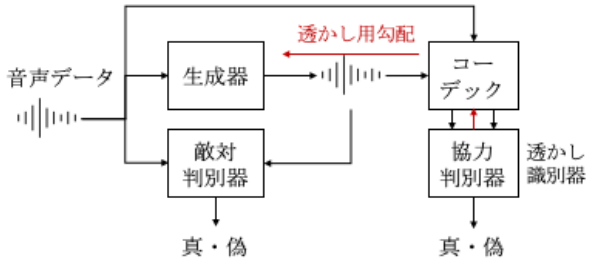

図1：音声生成モデル向けの協力的透かし技術の概要

[2] Zenodo に公開されたダウンロード回数

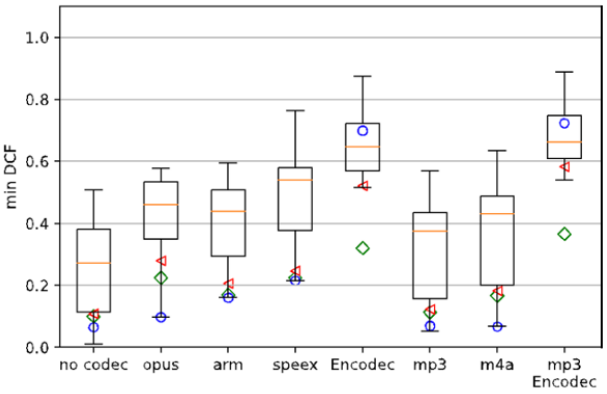

図2：コーデックにより圧縮されたデータに対する、ASVspoof5 Challenge 上位システムの識別性能（min Detection cost function, min-DCF）

codec	bitrate	None - o	None - c
None	-	1.53±0.3	0.29±0.1
DAC	8	35.85±1.0	24.29±0.9
OGG-Opus	16	53.42±1.0	41.53±1.0
OGG-Opus	32	21.75±0.8	9.18±0.6
OGG-Opus	64	4.55±0.4	0.45±0.1
OGG-Opus	128	2.27±0.3	0.21±0.1
MP3	16	49.70±1.0	48.42±1.0
MP3	32	31.96±1.0	11.49±0.7
MP3	64	6.26±0.5	0.79±0.2
MP3	128	1.92±0.3	0.29±0.1
OGG-Vorbis	q=1	46.50±1.0	27.08±0.9
OGG-Vorbis	q=2	42.17±1.0	17.39±0.8
OGG-Vorbis	q=3	38.10±0.9	9.59±0.7
All	-	28.58±0.2	16.38±0.2

図3：音声透かし技術を用いた合成音声の識別性能（エラー率 [%]）。DAC や OGG などはテストデータにかけたコーデック。列 c は協力的音声透かしの識別エラー率、列 o は音声透かしない場合の識別エラー率。

結果および考察

ASVspoof 5 を用いて ASVspoof 5 Challenge を主催しました。世界中の 50 組織が音声ディープフェイク検出システムを開発し、チャレンジに参加しました。結果の一つとして、図2にしめされたように、人間の音声と合成音声データが MP3 や Opus などで圧縮された場合、上位システムでも合成音声データの識別性能が劣ることがわかりました。その中でも、最新の深層学習を用いたコーデック（Encodec）で圧縮された場合、識別エラー率が最も高くなることがわかりました。

音声透かしの実験でも、コーデックが識別性能に影響を与える結果が得られました。コーデックなし（None）の場合、提案された音声透かし（c）は 0.29% のエラー率で合成音声を識別できました（図3列 c）。しかし、コーデックを適用すると、人間音声と合成音声を区別するための特徴量が失われ、識別性能が低下しました。また、コーデックの圧縮率が高いほど、識別エラー率が高くなることもわかりました。

まとめ、今後の課題

本年度のプロジェクトで得られた結果に基づき、以下の課題に取り組む予定です：

1) 圧縮された音声データに対するディープフェイク検出精度の向上
2) 協調音声透かしを用いた音声生成モデルの学習時に、最新のコーデックを適用できるかどうかの検討

また、本年度のプロジェクトでは、話者匿名化に関する研究も進める予定です。

参考文献

[1] Xin Wang, Héctor Delgado, Hemlata Tak, et al. ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. **In ASVspoof Workshop 2024**, 2024. 1--8. https://doi.org/10.21437/ASVspoof.2024-1

[2] Lauri Juvela and Xin Wang. 2025. *Audio codec augmentation for robust collaborative watermarking of speech synthesis*. **In Proc. ICASSP, 2025**. https://arxiv.org/abs/2409.13382

[3] Xin Wang, Héctor Delgado, Hemlata Tak, et.al. 2024. *ASVspoof 5: Design, collection and validation of public resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech*. **Submitted to Computer Speech & Language** https://arxiv.org/abs/2502.08857