TSUBAME 共同利用　令和7年度　学術利用　成果報告書

利用課題名　実環境における音声と非音声の頑健なディープフェイク検出
英文：Toward a Universal Deepfake Audio Detector: Robust Deepfake Detection of Speech and Non-Speech in Real-World Conditions

利用課題責任者: GE WANYING
First name Surname: Wanying Ge

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

英文抄録 This research aims to develop robust speech deepfake attribution technologies. This year, we proposed "FakeMark," a watermarking framework that injects artifact-correlated cues to enable reliable source attribution even under malicious attacks or distortions. Our results provide a foundation for more resilient and adaptive speech security systems.

Keywords: speech deepfake attribution, watermarking, robustness, speech processing

Background and Objectives

The rapid evolution of generative AI has made it easy to create highly realistic "speech deepfakes," posing severe security risks such as fraud and misinformation. Traditional detection systems often struggle to generalize to unseen generation models or remain vulnerable to signal distortions (e.g., codec compression). To address these limitations, this project focuses on enhancing "attribution" (identifying the source system) through proactive watermarking. Using TSUBAME's computing resources, we developed methods to maintain high accuracy in diverse, real-world scenarios.

FakeMark: Deepfake Attribution with Watermarked Artifacts

We developed "FakeMark," a novel framework that injects watermarks correlated with the intrinsic artifacts of specific deepfake systems, as illustrated in Figure 1.

Unlike conventional methods that embed arbitrary bitstrings, FakeMark allows a detector to identify the source system by leveraging both the injected watermark and the generator's natural fingerprints.

Experimental results showed that FakeMark maintains high attribution accuracy even when samples undergo heavy codec compression or removal attacks. This hybrid approach ensures that if the watermark is partially removed, the intrinsic artifacts still provide sufficient cues for detection, significantly improving robustness over existing solutions.

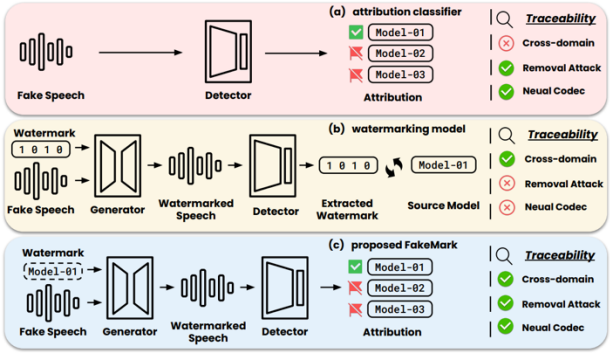

Figure 1：Overview of (a) deepfake detection and attribution, (b) classifier-based attribution, (c) watermarking-based attribution, and (d) proposed FakeMark.

Future Work

Moving forward, we will investigate a new

perspective on MLOps: the impact of data quality on training performance. We aim to develop strategies to select high-value data and remove less useful samples during out-of-domain training, thereby improving both detection accuracy and computational efficiency.

Additionally, we plan to examine how to extend these speech-based MLOps outcomes to general audio and music data. This expansion will help establish a more universal framework for audio deepfake detection and attribution across diverse acoustic domains.

参考文献

[1] Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. FakeMark: Deepfake Speech Attribution With Watermarked Artifacts. Under review.