Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (TIB PAN)

SUBJECT:

Segment-Level Detection of Partially Synthesized Audio Deepfakes with Insight into Model Decisions

SUPERVISORS:

Piotr Szczuko PhD DSc (main supervisor);
piotr.szczuko@pg.edu.pl;
Gdańsk University of Technology;

Ewelina Bartuzi-Trokielewicz PhD (auxiliary supervisor);
ewelina.bartuzi@nask.pl;
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa

DESCRIPTION:

The proposed PhD research focuses on **segment-level detection of partially synthesized audio deepfakes**, with an emphasis on **interpreting model decisions**. Audio deepfakes – synthetic or manipulated voice recordings – are becoming increasingly realistic and widely accessible due to advances in neural speech synthesis. These technologies, while promising for accessibility and content creation, also raise significant ethical and safety concerns, including impersonation attacks, misinformation campaigns, voice spoofing, and violations of privacy and consent [1].

Most existing research targets the detection of fully synthesized utterances; however, **partially manipulated audio**, in which only short segments are altered – e.g., a few words replaced within otherwise authentic speech – poses a more nuanced and difficult problem [2], [3]. A subtle change such as altering “You owe me five hundred dollars” to “I owe you five hundred dollars” can entirely reverse the meaning of a sentence [4], and such manipulations may evade detection by current systems. This research aims to address that gap by advancing detection techniques for these nuanced cases and providing tools to safeguard against such attacks.

In parallel, it is essential to examine the **internal representations and learned features** that guide model decisions across different architectures, including convolutional, recurrent, and transformer-based networks. A focus on **explainability** can provide valuable insight into these mechanisms. Techniques like **Layer-Wise Relevance Propagation (LRP)** [5] and **SHAP (SHapley Additive exPlanations)** [6] are promising tools for revealing whether specific acoustic cues (e.g., artifacts, prosody, or irregular speech rate) are systematically responsible for detection triggers.

Another key research direction involves studying **the robustness of detection methods to real-world distortions** such as lossy compression, downsampling, background noise, or reverberation – factors that significantly affect performance in practical scenarios [2], [7]. Understanding how these operations alter both the input signal and model explanations could lead to more resilient designs. Furthermore, it is important to investigate the impact of other potential sources of bias, such as language, speaker gender, or accent, which may also influence detection outcomes and model behaviour.

Recent advancements in audio deepfake detection highlight the effectiveness of **RawNet-based architectures**, which process raw audio waveforms to capture intricate temporal features. Architectures such as AASIST further enhance detection by integrating both time and frequency domain information through spectro-temporal graph attention networks [8]. Additionally, incorporating features from large-scale pre-trained models like Whisper has been shown to improve generalization and detection performance [9]. The proposal also considers the potential of incorporating **advanced speech processing techniques**, including speaker diarization [10], and deep clustering for source separation [11], especially in complex acoustic environments.

Recent advancements in audio deepfake detection highlight the effectiveness of **RawNet-based architectures**, which operate directly on raw audio waveforms to capture intricate temporal features. Architectures such as **AASIST** further improve detection performance by integrating information from both the time and frequency domains using spectro-temporal graph attention networks [8]. Moreover, leveraging features from large-scale pre-trained models like **Whisper** has been shown to enhance generalization and overall detection accuracy [9]. This proposal also explores the potential of **advanced speech processing techniques** – such as speaker diarization [10] and deep clustering for source separation [11], particularly in challenging acoustic environments.

Overall, the research aims to contribute toward more **granular, interpretable, and robust detection of audio manipulations**, with implications for forensics, security, and media authentication.

REQUIREMENTS:

- MSc degree in computer science, mathematics or a related field
- Programming skills (preferably in Python)
- Experience with knowledge engineering
- Knowledge about current research on automatic speaker verification (ASV)
- Advanced Level in English (speaking and writing)

BIBLIOGRAPHY:

[1] Hutiri W., Papakyriakopoulos O., & Xiang A. (2024). *Not my voice! A taxonomy of ethical and safety harms of speech generators*. arXiv:2402.09989.

[2] Gupta G., Raja K., Gupta M., Jan T., Whiteside S. T., & Prasad M. (2024). *A comprehensive review of deepfake detection using advanced machine learning and fusion methods*. Electronics, 13(1), 95.

[3] Mcuba M., Singh A., Ikuesan R. A., & Venter H. S. (2023). *The effect of deep learning methods on deepfake audio detection for digital investigation*. Procedia Computer Science, 219, 466–473.

[4] Liu L., Wei H., Liu D., & Fu Z. (2024). *HarmoNet: Partial DeepFake Detection Network based on Multi-scale HarmoF0 Feature Fusion.* In Proceedings of Interspeech 2024. isca-archive.org

[5] Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., & Samek W. (2015). *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*. PloS One, 10(7), e0130140.

[6] Lundberg S. M., & Lee S.-I. (2017). *A unified approach to interpreting model predictions*. In Advances in Neural Information Processing Systems (NeurIPS).

[7] Nadimpalli A. V., & Rattani A. (2022). *On improving cross-dataset generalization of deepfake detectors*.

[8] Jung J. et al. (2021). *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks.* arXiv e-prints.

[9] Kawa P., Plata M., Czuba M., Szymański P., & Syga P. (2023). *Improved deepfake detection using whisper features*.

[10] Aperdannier R., Schacht S., & Piazza A. (2024). *A Review of Common Online Speaker Diarization Methods*. arXiv:2406.14464.

[11] Kim J., Kindt S., Madhu N., & Kang H.-G. (2024). *Enhanced Deep Speech Separation in Clustered Ad Hoc Distributed Microphone Environments*. arXiv:2406.09819.

[12] Xie Z., Li B., Xu X., Liang Z., Yu K., & Wu M. (2024). *FakeSound: Deepfake General Audio Detection.* arXiv:2406.08052.

[13] Luong H. -T., Li H., Zhang L., Lee K. A., & Chng E. S. (2025). *LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation*. Speech and Signal Processing (ICASSP)

[14] Zeng S. et al. (2025). *Adversarial Training and Gradient Optimization for Partially Deepfake Audio Localization*. Speech and Signal Processing (ICASSP) .

[15] He J., Yi J., Tao J., & Zeng S. (2025). *PET: High-Frequency Temporal Self-Consistency Learning for Partially Deepfake Audio Localization*. Speech and Signal Processing (ICASSP).

[16] Alali A., & Theodorakopoulos G. (2025). *Partial Fake Speech Attacks in the Real World Using Deepfake Audio*. Journal of Cybersecurity and Privacy, *5*(1), 6.