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

SUBJECT:

Generalization of the synthetic speech detection.

SUPERVISOR:

prof. dr hab. inż. Michał Choraś
e-mail: Michal.Choras@pbs.edu.pl
Wydział Telekomunikacji, Informatyki i Elektrotechniki
Politechnika Bydgoska
ul. Profesora Sylwestra Kaliskiego 7, 85-796 Bydgoszcz

dr inż. Ewelina Bartuzi-Trokielewicz
e-mail: ewelina.bartuzi@nask.pl
NASK - Państwowy Instytut Badawczy
ul. Kolska 12, 01-045 Warszawa

DESCRIPTION:

The rapid and widespread development of machine learning models for generating human speech presents a plethora of significant ethical and safety concerns for society. These models facilitate voice cloning and text-based speech generation or real-time voice conversion from one individual to another, creating an utterance that becomes difficult for the human ear to distinguish from real speech.

In response to this threat, a number of methods have been proposed to detect crafted voice samples. Many of these methods are based on deep learning models and frame this task as a binary classification problem, which has shown extremely high performance. However, further work indicates that, the performance of these methods notably declines when evaluated across different datasets that were not considered during training. We face the challenge of increasing the robustness of these methods so that they can effectively detect new manipulations.

The purpose of this study will be to explore the potential of generalizing synthetic speech detection methods to data obtained by different types of synthesis methods that were not used during the training phase. This aspect involves the investigation of the influence of the training process (training set selection, augmentation methods), network architecture, classification strategies and other aspects that can affect the reliability of the solution.

The author will propose and test their own methods for detecting synthetic speech to prove the thesis that it is possible to detect synthetic speech (audio deepfake) using machine learning methods that incorporate biometric features. The author will also consider other modalities, including the analysis of text derived from speech (known as text to speech) with particular emphasis on sentiment analysis.

BIBLIOGRAPHY:

1. Hutiri, W., Papakyriakopoulos, O., & Xiang, A. (2024). Not my voice! a taxonomy of ethical and safety harms of speech generators. arXiv (Cornell University).
2. Gupta, G.; Raja, K.; Gupta, M.; Jan, T.; Whiteside, S.T.; Prasad, M. A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. Electronics 2024, 13, 95.
3. Bhangale, K.B., Kothandaraman, M. Survey of Deep Learning Paradigms for Speech Processing. Wireless Pers Commun 125, 1913–1949 (2022).
4. 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
5. Jordan J. Bird and Ahmad Lotfi. Real-time detection of AI-generated speech for deepfake voice conversion, 2023.
6. Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, 245–263
7. Aakash Varma Nadimpalli and Ajita Rattani. On improving cross-dataset generalization of deepfake detectors, 2022
8. Piotr Kawa, Marcin Plata, Michal Czuba, Piotr Szymański, and Piotr Syga. Improved deepfake detection using whisper features, 2023
9. Piotr Kawa, Marcin Plata, and Piotr Syga. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. In Interspeech 2022,interspeech2022.ISCA, September2022
10. Zhang, Y., Jiang, F., Zhu, G., Chen, X., & Duan, Z. (2023). Generalizing voice presentation attack detection to unseen synthetic attacks and channel variation. In Advances in computer vision and pattern recognition.