SafeSpeak-2024: Audio Spoofing Detection Hackathon for Voice Security

Vyacheslav Berezhnoy, Kirill Borodin, Ivan Chernov, Nikita Koshelev
Deep Robotics Institute, Novosibirsk State University
Novosibirsk, Russia
Emails: v.berezhnoy31@gmail.com, k.borodin@g.nsu.ru, i.chernov2@g.nsu.ru, nik_koshelev03@mail.ru

***Abstract*—SafeSpeak-2024 is a hackathon aimed at advancing audio spoofing detection technologies for secure voice authentication, addressing challenges in voice anti-spoofing and presentation attack detection. The competition tasks participants with developing lightweight, high-performance models to detect audio spoofing attacks. Emphasizing computational efficiency and real-world applicability, models are evaluated using ASVspoof metrics to ensure industry-standard robustness and accuracy.**

***Index Terms*—Anti-spoofing, presentation attack detection, automatic speaker verification, deepfake detection, self-supervised learning, wav2vec 2.0**

I. Introduction

We initiated our work for the SafeSpeak-2024 hackathon by reviewing recent advancements in audio anti-spoofing, guided by a comprehensive survey [1]. Most spoofing detection systems comprise two components: a feature extractor and a classifier.

II. Research

A. Feature Extractor

Traditional systems employed deterministic feature extractors, such as spectrograms. Recent approaches leverage deep learning for more informative, domain-specific feature representations, including supervised methods like RawNet2 [2] and self-supervised methods like wav2vec 2.0 [3].

B. Classifier

Many classifiers in the survey are convolutional neural networks (CNNs), originally designed for image classification. For audio-specific classification, architectures like Graph Attention Networks (GANs), e.g., AASIST [4], have been developed.

C. Complete Model

Our solution is based on a top-performing model from the survey [5], utilizing wav2vec 2.0 as the feature extractor and AASIST as the classifier, augmented with RawBoost data augmentation [6]. Due to its high parameter count (560M), we opted for the lightweight PSFAN model [7] as our classifier, designed for EEG classification, a time-series domain similar to audio. Our final model has approximately 330M parameters. We also experimented with a Vision Transformer classifier but observed no performance improvement, highlighting the value of pretraining and transfer learning.

III. Results

Model evaluation results are shown in Table I. Our model processes two audio files per second on a CPU, suitable for practical applications, such as mobile devices. The source code is available at the referenced repository.

TABLE I
Model Evaluation

Batch Size	QPS GPU (s)	QPS CPU (s)
8	55	60
16	58.6	2.2

IV. Abbreviations and Acronyms

- **EER**: Equal Error Rate, a metric for evaluating binary classifiers, indicating the point where False Acceptance Rate (FAR) and False Rejection Rate (FRR) are balanced. A low EER reflects a system's ability to minimize errors, critical for distinguishing genuine from fake speech.
- **CNN**: Convolutional Neural Network.

References

[1] M. Li, Y. Abmadiadi, and X.-P. Zhang, "Audio antispoofing detection: A survey," 2024.
[2] S. w. Jung, L.-S. Hoo, H. Hoo, H. j. Kim, H.-J. Shim, and H.-J. Yu, "Rawntet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," 2019.
[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.
[4] J. w. Jung, H.-S. Heo, H. Tak, H. j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," 2021.
[5] H. Tak, M. Todisco, X. Wang, J. w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022.
[6] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing," 2022.
[7] M. Kamati, G. Saha, A. Gupta, A. Seal, and O. Krejcar, "A pyramidal spatial-based feature attention network for schizophrenia detection using electroencephalography signals," 2024.