Improving synthetic speech detection accuracy by replacing and finetuning prediction head

Georgy Shulyndin, Inna Larina, Pavel Aleksandrov – Sypukha team

***Abstract*—This document describes our solution for SafeSeak-2024 Hackathon. The goal is to present a robust and lightweight audio spoof detection model opimized for real-world applications.**

***Index Terms*—Audio spoofing, Synthetic speech detection, Deepfakes.**

I. Introduction

VOICE authentication systems have become an integral part of modern security infrastructures, offering a convenient and natural way for users to access digital services. However, the rise of sophisticated audio spoofing attacks – such as voice synthesis and voice conversion – poses a significant threat to the reliability and trustworthiness of these systems. These vulnerabilities highlight the urgent need for advanced anti-spoofing mechanisms to safeguard voice authentication technologies against malicious exploitation.

In this article, we present our approach to tackling the challenges of voice anti-spoofing and presentation attack detection. Our solution combines modern deep neural network models and basic ML algorithms to detect and mitigate audio spoofing attempts effectively. Our method is evaluated against the proposed in Hackathon baseline metrics.

II. The idea

The core idea of our approach lies in the combination of a feature extractor and a detector to optimize audio spoofing detection performance. Inspired by recent advancements in the field, particularly the paper of Schäfer et al. [3] presented on ASVspoof 2024 workshop, we decided to use wav2vec encoder as the feature extractor and AASIST which was selected as the baseline model for the ASVspoof 2024 workshop as the detector.

To implement this, we identified a pretrained model described in work of Tak et al. [4] closely aligned with the architecture described by Schäfer et al. [3]. We then replaced the model's original prediction head with a KNN machine learning classification algorithm fitted for our task on ASVspoof19 LA eval dataset. This adaptation allowed us to fine-tune the model effectively on any provided dataset and improved explainability of models pridictions.

III. Architecture

In this work, we utilize models pretrained on extensive and diverse speech datasets. To select suitable models, we reviewed the papers presented at the ASVspoof 2024 Workshop, focusing on their architectures and reported performance metrics. Based on this analysis, we chose to experiment with the wav2vec2 encoder and the AASIST backbone.

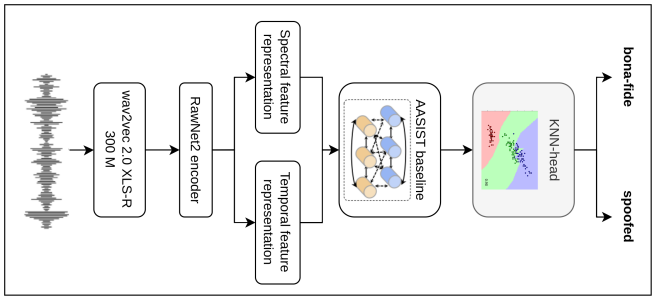

Fig. 1. **Model Architecture Overview.** We use XLS-R as the front-end and an integrated spectro-temporal graph attention network (AASIST) as the back-end. The final linear layer in AASIST is replaced with a KNN head, which outputs a binary prediction.

Our primary contribution is the integration of a K-Nearest Neighbors (KNN) head in place of the standard linear layer in the AASIST backbone. Features extracted by AASIST are classified using the supervised machine learning algorithm KNN, which demonstrates strong performance in classification tasks.

A. Wav2vec 2.0

Wav2vec2 is a pretrained model for Automatic Speech Recognition (ASR) and was released in September 2020 by Alexei Baevski, Michael Auli, and Alex Conneau. In this paper we leverage XLS-R 300M [1] which is a multi-lingual version of wav2vec2 with 300 million parameters. Facebook AI's XLS-R model is pretrained on 436k hours of unlabeled speech in 128 languages, including VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107.

B. AASIST

The AASIST model [2] is a state-of-the-art deep learning framework designed specifically for audio spoofing detection in voice authentication systems. AASIST leverages spectro-temporal features to effectively distinguish between genuine and spoofed audio inputs.

The model incorporates a CNN for extracting low-level spectro-temporal representations and combines these with a self-attention mechanism to capture long-range dependencies and context. This architecture enables AASIST to analyze both local and global acoustic patterns.

C. KNN-head

We utilize K-Nearest Neighbors (KNN) as a classifier for the features extracted by our model. The goal of the KNN algorithm is to identify the nearest neighbors of a given query point and assign a class label based on the majority class among these neighbors. During the training stage, KNN stores the entire training dataset, which makes the algorithm computationally inefficient when working with large datasets.

IV. Dataset

In the original paper [4] wav2vec2+AASIST model was fine-tuned on ASVspoof19 LA train partition only. To provide new data for the model and improve generalizability, we fitted KNN-head on 60,000 samples from ASVspoof19 LA eval set.

V. Results

The final EER achived by our solution is 0.149. This result is obtained by training KNN head on 60,000 samples from ASVspoof19 LA eval partition. During inference, the class prediction for each audio sample was determined based on the majority class among its 10 nearest neighbors. On the system with single Nvidia A5000 GPU model processed 145K test samples in 38 minutes.

References

[1] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[2] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *arXiv preprint arXiv:2110.01200*.

[3] Karla Schäfer, Jeong-Eun Choi, and Matthias Neu. Robust audio deepfake detection: exploring front-/back-end combinations and data augmentation strategies for the asvspoof5 challenge. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 56–63, 2024.

[4] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation, 2022.