FuseAST: Fusing Noise and Spectrograms for Audio Deepfake Detection

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Min Park, Gordon Lim, Qirui Jin, William Zheng, Bill Shao
{phyunmin, gbtc, qiruijin, willizhe, szx}@umich.edu

1. Abstract

In recent years, Artificial Intelligence (AI) technologies have shown great promise in various domains such as video recommendation, bioinformatics, and speech recognition. Our research focuses on Audio Speech Verification (ASV) systems, which are widely used in critical security systems in real-world applications. However, with the increasing pervasiveness and capability of AI systems, the security and integrity of many ASVs have come into question. In particular, AI-enabled attacks via text-to-speech synthesis and voice conversion are capable of generating deepfake audio to mimic individuals' voices, posing significant threats through spoofing attacks. To address these vulnerabilities, we present FuseAST, based on state-of-the-art Audio Spectrogram Transformers (AST) [1] for distinguishing real audio from deepfake audio. FuseAST introduces two different feature engineering strategies: (1) fusing Log-Mel spectrograms with delta and delta-delta features in three channels, representing the first and second temporal derivatives, interpreting the temporal dynamics of the audio, and (2) fusing the features of high and low-frequency regions in the spectrogram with weight masking, providing different perspectives for analysis. Our evaluation showed that the integration of the two feature engineering strategies resulted in a significant 6% improvement in performance accuracy over the Vallina AST model as evaluated on the ASVspoof challenges. In addition, FuseAST significantly outperforms the baseline LFCC-GMM model on the Equal Error Rate (EER) metric. FuseAST underscores the effectiveness of innovative feature engineering in strengthening the security of ASV systems against AI-driven spoofing attacks.

2. Introduction

As audio deepfakes become increasingly realistic, it has become increasingly difficult to distinguish them by ear alone. Deep learning methods have shown great promise in detecting these deepfakes. In our paper, we experiment with the state-of-the-art Audio Spectrogram Transformer (AST) for its effectiveness in detecting audio deepfakes. AST is a Vision Transformer (ViT)-based model that represents the state-of-the-art (SOTA) in audio classification tasks, offering a novel approach in comparison to existing methodologies like LightCNN, ResNet, and LSTM models. It has demonstrated strong performance in multi-class audio classification tasks, achieving state-of-the-art results on AudioSet, 95.6% accuracy on RSC-50, and 98.1% on Speech Commands V2. Each audio clip is converted into an audio spectrogram, and the AST implementation concatenates three audio spectrograms into a three-channel image and uses this image as input to the model. To classify audio spectrograms with variable sizes, AST uses linear interpolation to map oversized audio spectrograms to the original positional embedding of ViT.

3. Background

Spectrogram is a visual representation of the frequency spectrum of a signal over time, with time on one axis, frequency on the other, and amplitude indicated by intensity or color. It is widely used in audio analysis, including speech and spoofing detection.

ASVspoof 2021 dataset serves as a benchmark for the development of countermeasures (CMs) against spoofing attacks targeting Automatic Speaker Verification (ASV) systems. The challenge consists of three different tasks: Physical Access (PA), Logical Access (LA), and DeepFake (DF) [2]. In this paper, we focus on the LA task, which includes bona fide and fake audio generated using 13 state-of-the-art voice conversion (VC), text-to-speech (TTS), and hybrid algorithms, as well as different transmission conditions such as VoIP and PSTN+VoIP [2]. The dataset is derived from the ASVspoof 2019 LA evaluation database, which in turn is based on the VCTK database.

Performance metrics We evaluate using Equal Error Rate (EER) which balances false acceptance and false rejection rates.

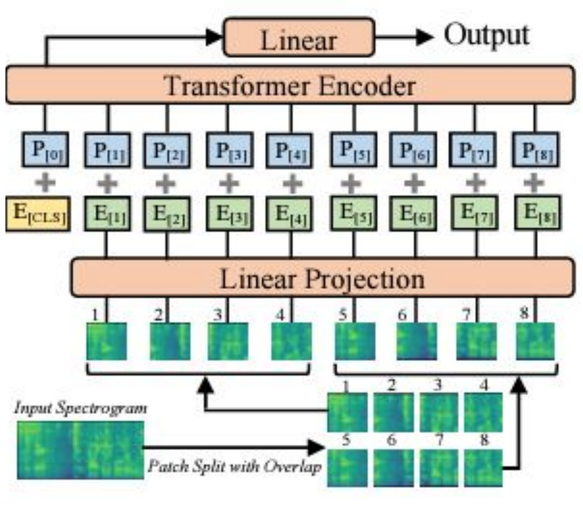

Figure 1. Audio Spectrogram Transformer architecture. Figure taken from [1].

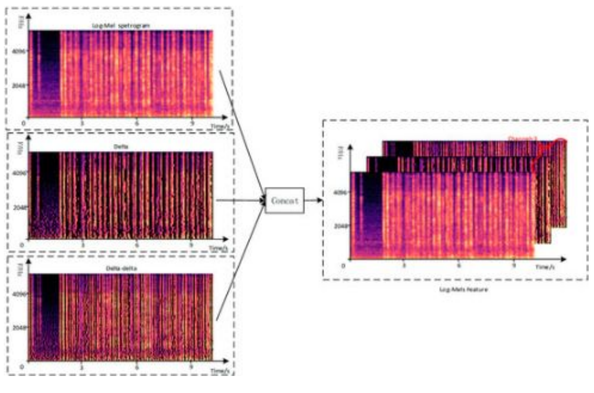

Figure 2. This figure, taken from Liu and Fang (2023), illustrates multi-spectrogram fusion using Log-Mel spectrogram, delta, and delta-delta features [3].

4. Methods

A. Multi-spectrogram fusion.

AST traditionally takes Log-Mel spectrograms as input. Log-Mel captures acoustic features such as frequency and amplitude in one form. However, while Log-Mel spectrograms are widely used, they may not comprehensively capture temporal dynamics. To address this, we incorporate delta and delta-delta features, representing the first and second temporal derivatives, respectively. By concatenating these features along the channel dimension, we construct a 3D representation that enhances temporal information and improves the model's ability to learn nuanced patterns. While previous work has successfully applied this technique to audio classification, its application to the more challenging task of audio deepfake detection remains unexplored.

B. Noise Fusion via Weight Masking.

We hypothesize that AI-based voice conversion models prioritize synthesizing natural human voice components, potentially resulting in less accurate synthesis of background noise and other non-human sound characteristics. We also try to find evidence that such pattern exits, as Figure 3 presents. To address this, we fuse noise in the feature by implementing a weighted embedding mask that targets specific patches corresponding to high and low frequency regions in the spectrogram. Specifically, weight biases of 0.5 were applied to the patch embeddings of the top 20% and bottom 20% frequency bins to encourage the model to pay more attention to these regions. This approach aims to increase the model's sensitivity to critical frequency bands associated with noise and reduce the influence of less relevant human-related regions.

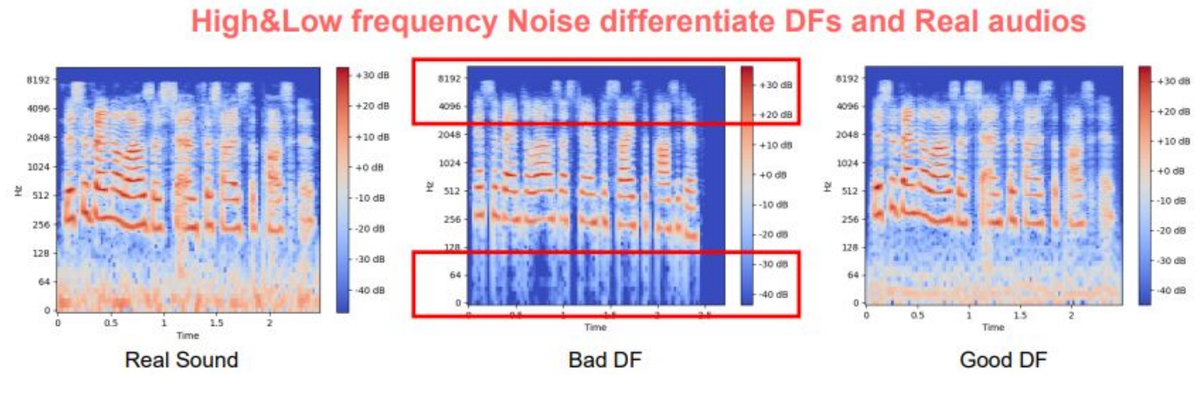

Figure 3. This figure shows an example of deepfake and ground-truth audio spectrograms. Bad DF indicates deepfake audio that is distinguishable using human years, Good DF indicates indistinguishable deepfake, and the boxed region indicates the weight-biased region.

5. Experiments & Results

We trained both the standard AST model and FuseASTs (AST models incorporating feature engineering) using pre-trained weights from AudioSet and ImageNet, followed by fine-tuning. The models were fine-tuned on the ASVspoof2021 training dataset provided in the challenge, which is derived from the ASVspoof2019 LA dataset. Evaluation was conducted on a balanced subset of the ASVspoof2021 LA evaluation set, comprising 15,000 bona fide samples and 15,000 spoofed samples. This balanced evaluation set was employed to mitigate the effects of the substantial class imbalance in the training data, which had a 10:1 ratio of spoofed to bona fide samples. By using a balanced evaluation set, we ensured that the performance metrics more accurately reflect the model's generalization capability across both bona fide and spoofed classes, avoiding biases introduced by class imbalance.

Model	EER (%)
LFCC-GMM	20.15
AST	15.43
AST + Multi-Spectrogram Fusion (Ours)	15.43
AST + Weight Masking (Ours)	12.85

Table 1. Equal Error Rate (%) results on balanced LA evaluation set.

6. Conclusion

To conclude, we fine-tuned AST with ASVspoof 2019 and successfully transformed it into a deepfake audio classifier. We then modified the embedding layer of AST to increase the weights of non-human audible frequencies, which resulted in an increase in performance. This suggests that focusing on extremely low and high frequencies may be a shortcut in identifying deepfake audio.

We also experimented with multi-channel audio spectrograms, where each channel represented a different type of spectrogram, but the model did not converge during training. Future work can explore feature engineering based on audio frequencies and corresponding embedding layers, as well as fine-tuning ViT directly on multi-spectrogram fusion.

References

[1]Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer." Available: https://arxiv.org/pdf/2104.01778

[2]X. Liu et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," IEEE/ACM transactions on audio, speech, and language processing, vol. 31, pp. 2507–2522, Jan. 2023, doi: https://doi.org/10.1109/taslp.2023.3285283.

[3]F. Liu and J. Fang, "Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition," Future Internet, vol. 15, no. 2, p. 65, Feb. 2023, doi: https://doi.org/10.3390/fi15020065.