HI-FAD and SpoofMix: A High-Frequency-Aware Framework and Realistic Benchmark for Robust Fake Audio Detection

Anonymous ACL submission

Abstract

Recently, fake audio detection (FAD) has made great progress in response to sophisticated spoofing attacks. However, existing frameworks still overlook two critical needs: (1) frequency-aware analysis of artifacts and (2) benchmark that simulate real-world spoofing attacks based on speech mixtures. To deal with these gaps, we propose HI-FAD, a novel high-frequency–aware FAD framework, and SpoofMix, a challenging benchmark incorporating both real and spoofed speech within single audio samples. In particular, HI-FAD employs a discrete wavelet transform (DWT) to extract high-frequency subbands and fuses them with front-end model representations via cross-attention. Experimental results demonstrate that HI-FAD consistently outperforms conventional methods on the ASVspoof2019 Logical Access (LA) and ASVspoof2021 LA. Moreover, the proposed framework achieves state-of-the-art detection on SpoofMix, demonstrating its robustness under realistic mixed-speech conditions. The source code and SpoofMix benchmark are available here : https://github.com/blind-review-user123/HI-FAD.git

1 Introduction

In recent years, there has been increasing interest in developing fake audio detection (FAD) methods capable of distinguishing bonafide (i.e., real) speech from spoofed speech. Prior studies on FAD have typically improved performance by modifying model architectures or incorporating advanced front-end encoders. For example, AASIST (weon Jung et al., 2021) introduces an extended variant of the graph attention layer, and HM-Conformer (seo Shin et al., 2024) leverages a Conformer (Gulati et al., 2020) architecture. More recently, speech self-supervised learning (SSL) models, such as Wav2Vec 2.0 (Baevski et al., 2020) and XLS-R (Babu et al., 2021), are used as front-end encoders to improve generalization to unseen spoofing attacks.

Despite their effectiveness, there are still two aspects that have not been sufficiently explored. First, there is a lack of frequency-aware analysis of artifacts in synthetic speech. Previous studies on speech synthesis point out that artifacts exist in the high-frequency domain between synthesized and real speech (Kim et al., 2021; Pons et al., 2021; Caillon and Esling, 2021). Specifically, vocoders (van den Oord et al., 2016; Kumar et al., 2019; Kaneko et al., 2022) often produce periodic distortions in higher frequency bands due to the limited ability of their upsampling layers. While architectural improvements have been extensively studied, the frequency-aware FAD framework that considers the characteristics of artifacts remains largely underexplored.

Second, FAD benchmarks (Wang et al., 2020; Liu et al., 2023) typically assume that each audio sample contains only a single speaker. However, such an assumption is overly simplistic and fails to reflect the complexity of real-world scenarios, such as recent voice phishing attacks, where audio may consist of both genuine and fake voice components. This mismatch between current benchmarks and real-world conditions raises concerns about the generalizability of FAD models.

To address these limitations, in this paper, we aim to improve the performance and generalizability of FAD by (1) introducing a high-frequency-aware detection method, namely HI-FAD, and (2) constructing a new benchmark SpoofMix based on speech mixtures that better reflect real-world spoofing attacks. Firstly, HI-FAD guides the FAD model to focus on high-frequency components indicative of fake audio. In particular, it employs a discrete wavelet transform (DWT) on the input waveform, decomposing it into low- and high-frequency subbands. A cross-attention mechanism is then used to fuse the feature extracted from

the high-frequency subbands with the representation from speech SSL model. This fusion allows the model to better capture fine-grained distortions in the high-frequency range. Additionally, we introduce SpoofMix, a challenging benchmark based on speech mixtures designed to reflect realistic spoof attacks. The dataset is constructed by concatenating two randomly selected samples from the ASVspoof dataset, including both bonafide–bonafide and bonafide–spoof combinations. This construction not only simulates realistic overlapping utterances but also contributes to balancing the dataset by increasing the proportion of bonafide samples, thereby alleviating the class imbalance inherent in ASVspoof.

Experimental results show that HI-FAD significantly improves the performance of FAD baselines on the ASVspoof2019 Logical Access (LA) dataset. In cross-dataset evaluation, where the model is trained on ASVspoof2019 LA and tested on ASVspoof2021 LA, the proposed framework also outperforms existing approaches, demonstrating strong generalization across datasets. In addition, on SpoofMix, a newly constructed benchmark, HI-FAD achieves superior detection performance compared to existing state-of-the-art baselines, indicating its robustness under realistic and challenging spoofing scenarios.

2 Related Work

Deepfake audio refers to artificially generated or transformed speech that mimics a specific speaker's voice with high precision using deep learning-based synthesis techniques. Deep fake audio can be categorized into TTS, VC, fake emotion, scene fake, partially fake, etc. depending on how it is manipulated. In particular, as AI technologies like TTS and VC continue to advance, synthetic voices are becoming more and more natural, and it is difficult to detect them with simple spectrum analysis Consequently, various techniques for detecting deep fake audio have been studied, and existing detection models can be broadly categorized into machine learning-based methods and deep learning-based methods. Machine learning techniques include the Gaussian mixture model (GMM) (Viroli and McLachlan, 2017) and the support vector machine (SVM) (Hearst et al., 1998), while deep learning-based models include the graph attention network (GAT) (Veličković et al., 2018), RawNet2 (weon Jung et al., 2020a), and transformer-based Rawformer (Xu et al., 2022). Recent research has focused on using deep learning to capture the unique characteristics of synthetic speech and develop more accurate and reliable detection methods.

3 Proposed Method

3.1 HI-FAD

3.1.1 Subband Decomposition via DWT

The overall architecture of HI-FAD is illustrated in Fig. 1 (a). To focus more on high-frequency components where contain artifacts, we decompose the input waveform x using the DWT, as defined in Equation 1. DWT decomposes the input waveform into multiple frequency subbands while partially preserving temporal resolution. This enables the model to track high-frequency components over time. Based on this information, model can analyze the characteristics of artifacts that appear within specific temporal segments.

$$\{x_{\text{cA3}}, x_{\text{cD3}}, x_{\text{cD2}}, x_{\text{cD1}}\} = DWT(x). \quad (1)$$

Among the subbands, x_{cD3},x_{cD2} and x_{cD1} contain a certain level of high-frequency information and are used as auxiliary signals to guide attention toward artifact regions in the original waveform x. To extract their embedding, we apply individual feature extractors to each subband. As a result, we obtain high-frequency embeddings for each subband, which are then concatenated into a single tensor to form a frequency-wise feature matrix H as follows:

$$\begin{aligned} h_{\text{cD3}} &= f_{\text{cD3}}(x_{\text{cD3}}), \\ h_{\text{cD2}} &= f_{\text{cD2}}(x_{\text{cD2}}), \\ h_{\text{cD1}} &= f_{\text{cD1}}(x_{\text{cD1}}), \end{aligned} \quad (2)$$

$$H = \begin{bmatrix} h_{\text{cD3}} \\ h_{\text{cD2}} \\ h_{\text{cD1}} \end{bmatrix} \in R^{3\times d} \quad (3)$$

3.1.2 Frequency Feature Attention

To allow the model to learn the relative importance across frequency bands, we introduce a frequency attention module. This module applies a linear transformation to each subband embedding, followed by an averaging operation and softmax normalization to compute attention weights as follows:

$$S = HW + b. \quad (4)$$

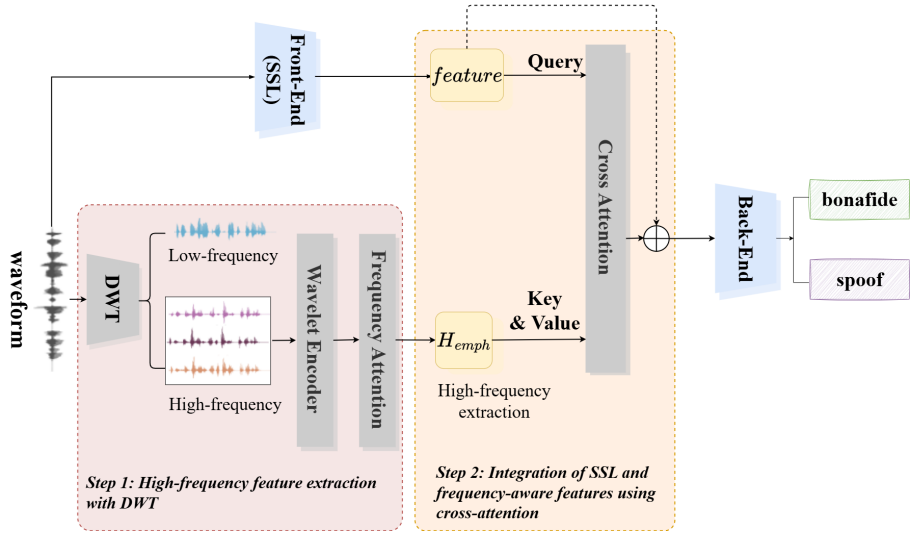

Figure 1: The overall pipeline of our framework, HI-FAD, which incorporates high-frequency-aware cross-attention.

$S \in \mathbb{R}^{3\times1}$ denote the importance scores assigned to each subband, and let $\alpha = [\alpha_{\text{cD3}}, \alpha_{\text{cD2}}, \alpha_{\text{cD1}}]$ represent the corresponding attention weights.

$$\alpha_i = \frac{\exp(S_i)}{\sum_{j=1}^{3}\exp(S_j)} \quad \text{for } i \in \{\text{cD3}, \text{cD2}, \text{cD1}\}. \tag{5}$$

Based on the computed attention weights, we generate a weighted feature representation H_{emph} by emphasizing features from more important subbands while suppressing those from less relevant ones. This is computed by element-wise multiplication between the attention weights α and the frequency-wise feature matrix H , which is computed as:

$$H_{\text{emph}} = \alpha \odot H. \tag{6}$$

Then, we integrate H_{emph} with the speech SSL model representation using a cross-attention mechanism. By leveraging cross-attention, this fusion enables the model to detect artifacts in the high-frequency range. The output of the cross-attention mechanism is subsequently added to speech SSL model representation through a residual connection. Due to residual connection, the model preserves both high-frequency information and the initial audio features. The residual output is used as input to the classifier of base model.

3.2 SpoofMix Benchmark

We propose SpoofMix benchmark by linearly concatenating two randomly selected samples from the ASVspoof dataset. SpoofMix id defigned to adress complex spoofing attacks that occur in real-world scenarious. The constructed samples include both bonafide-bonafide and bonafide-spoof combinations, enabling the simulation of realistic overlapping utterances. The mixed sample is labeled bonafide if and only if both utterances are bonafide. If even one of the two mixed utterances is a spoof, the mixed sample is labeled as spoof. This helps the model by exposing it to more diverse and challenging spoofing environments. Moreover, the ASVspoof dataset contains a lower number of bonafide samples than spoof samples, resulting in a class imbalance. SpoofMix increases the proportion of bonafide samples during the data construction process, thereby enabling balanced 1:1 sampling between bonafide and spoofed utterances. This benchmark extends beyond data augmentation by reflecting realistic spoofing scenarios. It offers a challenging setting for both training and evaluating the model's robustness and generalization ability.

4 Experiments

4.1 Dataset and Evaluation Metrics

We trained our models on the ASVspoof2019 Logical Access (LA) dataset and evaluated their in-domain performance on ASVspoof2019 LA as well as cross-dataset generalization on ASVspoof2021 LA. XLSR+AASIST (Tak et al., 2022) was employed as the backbone for the proposed HI-FAD, and We adopted the configurations and hyperparameters from the original paper (Tak et al., 2022). Performance was measured using equal error rate (EER) and the minimum tandem detection cost

Methods	19LA		21LA	
	EER (%)	min t-DCF	EER (%)	min t-DCF
RawNet2 (weon Jung et al., 2020b)	4.66	0.1294	9.50	0.4257
AASIST (weon Jung et al., 2021)	2.11	0.0692	8.39	0.4108
XLSR + Conformer (Rosello et al., 2023)	41.31	0.7493	1.21	0.2173
XLSR + Conformer + TCM (Truong et al., 2024)	59.40	0.9999	1.11	0.2161
XLSR + AASIST (Tak et al., 2022)	0.27	0.0089	0.95	**0.2090**
HI-FAD (proposed)	**0.20**	**0.0066**	**0.90**	0.2098

Table 1: Comparison of EER (%) and min t-DCF performance of ASVspoof2019 LA and ASVspoof2021 LA dataset.

function (min t-DCF). To assess robustness under more realistic and diverse spoofing scenarios, we also trained and tested on the SpoofMix benchmark.

4.2 Experimental Results

ASVSpoof. Table 1 reports our EER results on ASVspoof2019 LA and ASVspoof2021 LA. HI-FAD achieves the lowest error rates in most cases, thereby establishing state-of-the-art performance. Notably, when compared to the XLSR + AASIST baseline, HI-FAD relatively reduced the EER by 25.1 % on ASVspoof2019 LA and by 5.3 % on ASVspoof2021 LA. This indicates that DWT-based high-frequency artifact extraction substantially enhances performance over the backbone model.

SpoofMix. Unlike the original ASVspoof datasets, our SpoofMix benchmark incorporates realistic spoofing scenarios with speech mixtures. Specifically, it includes two spoofing conditions: (1) both utterances are spoofed and (2) only one utterance is spoofed. Framed as a differential detection task rather than conventional classification, SpoofMix presents a more challenging environment for conventional FAD approaches. We trained on the SpoofMix training set and evaluated on its test set. As presented in Table 2, HI-FAD achieved the lowest EER of 2.53 %, demonstrating its robustness under these mixed-speech spoofing conditions. Compared to other approaches, our DWT-based subband decomposition combined with cross-attention fusion allows HI-FAD to maintain high accuracy in mixed-speech scenarios, underscoring the importance of high-frequency artifact analysis under realistic spoofing conditions.

4.3 Analysis: Decomposition Level of DWT

As the DWT decomposition level increases, the input waveform is split into a greater number of subbands, allowing finer-grained frequency analysis. To quantify this effect, we evaluated levels 2 through 4 on the XLSR + AASIST backbone. As shown in Table 3, level 3 yielded the lowest EER and min t-DCF, indicating that an appropriately chosen decomposition depth is critical for maximizing FAD performance.

Model	EER(%)
AASIST	5.13
XLSR + AASIST	2.83
HI-FAD (proposed)	**2.53**

Table 2: Comparison of EER(%) of SpoofMix dataset.

DWT level	EER(%)	t-DCF
Level 2	0.92	0.2100
Level 3	**0.90**	**0.2098**
Level 4	1.03	0.2139

Table 3: Performance comparison of XLSR + AASIST under different wavelet decomposition levels.

5 Conclusion

We proposed a novel FAD framework, HI-FAD, by combining DWT with cross-attention to emphasize high-frequency components of fake audio. This framework enabled the model to concentrate more on the high-frequency artifacts, thereby achieving more sensitive FAD. Additionally, we introduced SpoofMix a challenging benchmark that reflects realistic spoofing scenarios. Experimental results showed that our proposed method improved performance on SpoofMix as well, showing that it was effective and generalizable to more difficult conditions.

6 Limitations

Despite its strong gains, the cross-attention fusion module that integrates high-frequency subbands into the backbone representations adds additional parameters and increases overall model complexity. However, given the substantial performance improvements it enables, this overhead is a reasonable trade-off. Moreover, our evaluation has been limited to English datasets; assessing HI-FAD in multilingual and cross-lingual scenarios remains an important direction for future work.

References

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. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Antoine Caillon and Philippe Esling. 2021. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *Preprint*, arXiv:2111.05011.

Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *Preprint*, arXiv:2005.08100.

Marti Hearst, S.T. Dumais, E. Osman, John Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. *Preprint*, arXiv:2203.02395.

Ji Hoon Kim, Sang Hoon Lee, Ji Hyun Lee, and Seong Whan Lee. 2021. Fre-gan: Adversarial frequency-consistent audio synthesis. *Preprint*, arXiv:2106.02297.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Preprint*, arXiv:1910.06711.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà. 2021. Upsampling artifacts in neural audio synthesis. *Preprint*, arXiv:2010.14356.

Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, and Antonio Peinado. 2023. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech 2023*, pages 5281–5285.

Hyun seo Shin, Jung woo Heo, Ju ho Kim, Chan yeong Lim, Won bin Kim, and Ha Jin Yu. 2024. Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10581–10585.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *Preprint*, arXiv:2202.12233.

Duc Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, page 537–541. ISCA.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *Preprint*, arXiv:1609.03499.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Cinzia Viroli and Geoffrey J. McLachlan. 2017. Deep gaussian mixture models. *Preprint*, arXiv:1711.06929.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, and 21 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Preprint*, arXiv:1911.01601.

Jee weon Jung, Seung bin Kim, Hye jin Shim, Ju ho Kim, and Ha Jin Yu. 2020a. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Preprint*, arXiv:2004.00526.

Jee weon Jung, Seung bin Kim, Hye jin Shim, Ju ho Kim, and Ha Jin Yu. 2020b. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Preprint*, arXiv:2004.00526.

Jee weon Jung, Hee Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong Jin Lee, Ha Jin Yu, and Nicholas Evans. 2021. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *Preprint*, arXiv:2110.01200.

Wanyan Xu, Xingbo Dong, Lan Ma, Andrew Beng Jin Teoh, and Zhixian Lin. 2022. Rawformer: An

efficient vision transformer for low-light raw image enhancement. *IEEE Signal Processing Letters*, 29:2677–2681.