Mixture of Spectral Experts for Audio Deepfake Detection

Yaxuan Qiu [1], *Zhe Li* [2], *Mieradilijiang Maimaiti* [1,**], *Zunwang Ke* [3], *Wushour Silamu*[1]

[1] School of Computer Science and Technology, Xinjiang University, China
[2] Speech, Language, and Cognition Laboratory, The University of Hong Kong, Hong Kong SAR
[3] School of Software, Xinjiang University, China

107552303968@stu.xju.edu.cn, miradeljan51@xju.edu.cn

Abstract

Recent advances in neural speech synthesis have produced highly natural waveforms, making audio deepfake detection increasingly challenging as spoofing artifacts become less perceptible. Although pre-trained speech models provide robust representations, they may overlook low-level physical cues, particularly magnitude and phase information. To address this limitation, we propose a detection framework that combines a frequency audio encoder (FAE) with spectral parameter-efficient fine-tuning. The FAE explicitly models magnitude and phase cues, while the proposed Mixture of Spectral Experts (MoSE) efficiently adapts the pre-trained speech model to generation-dependent distribution shifts. By applying low-rank updates in the singular value decomposition (SVD) domain while keeping the singular bases frozen, MoSE facilitates task-specific adaptation to spoofing-related spectral artifacts. Evaluations on ASVspoof 2019 LA, ASVspoof 2021 LA/DF, and In-the-Wild benchmarks demonstrate the effectiveness of our approach and its strong generalization to unseen channel variations and real-world spoofing attacks.

Index Terms: Audio Deepfake Detection, Self-Supervised Learning, Frequency Awareness, WavLM, Spectral Adaptation

1. Introduction

Self-supervised learning (SSL) pre-trained models (PTMs) have shown strong effectiveness in audio deepfake detection by providing robust speech representations. Recent SSL-based detectors adapt wav2vec 2.0-XLS-R, WavLM, and wav2vec 2.0 through fine-tuning, feature fusion, or mixture-of-experts (MoE) mechanisms [1, 2, 3, 4]. However, the contextual representations learned from SSL pre-training may underrepresent low-level physical artifacts, such as magnitude irregularities and phase-related distortions, that are crucial for distinguishing synthetic from genuine speech.

To complement PTM-based detectors with explicit low-level information, previous work has attempted to incorporate frequency-domain features, such as linear frequency cepstral coefficients (LFCC) [5] and constant-Q transform (CQT) features [6] to capture frequency-domain artifacts. More recent studies explore the fusion of raw waveform encoders such as RawNet2 [7] with PTMs. Despite their effectiveness, these methods still face two limitations. First, phase information is often insufficiently modeled, despite its potential to reveal artifacts introduced during waveform generation. Second, a representation gap may exist between explicit frequency-domain cues and the high-level contextual embeddings of PTMs, limiting the effective integration of low-level physical information.

** indicates the corresponding author.

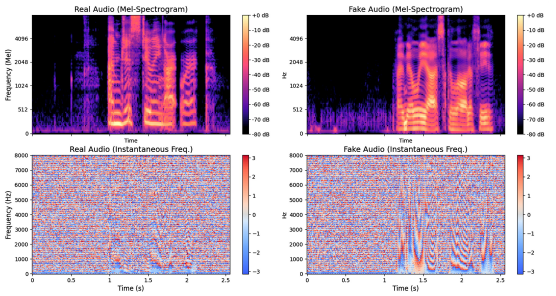

Figure 1: *Comparison of real (left) and fake (right) audio. The fake sample shows blurred harmonic structures in the Mel-spectrogram (top) and irregular instantaneous-frequency patterns (bottom).*

The importance of explicit phase modeling is closely tied to the characteristics of speech generation pipelines. In many speech synthesis and voice conversion systems, waveform generation or vocoding from intermediate acoustic representations can introduce phase-related inconsistencies, since accurate phase reconstruction remains challenging [8]. As shown in Fig. 1, although synthetic speech may sound natural, it can exhibit blurred harmonic structures and irregular instantaneous-frequency patterns that differ from bonafide speech [9, 10]. These observations suggest that magnitude and phase cues provide complementary evidence for detecting synthetic speech artifacts. Accordingly, we introduce a frequency audio encoder (FAE) that explicitly decomposes speech signals into magnitude and phase components, providing low-level physical cues that complement the high-level contextual representations of PTMs.

While the FAE enhances the input representation with explicit spectral cues, effectively exploiting such cues also requires adapting the PTMs. Direct full fine-tuning is computationally expensive and may disturb the acoustic priors learned from PTMs. Existing parameter-efficient fine-tuning (PEFT) methods, such as adapter- or low-rank adaptation (LoRA)-style updates [11, 12, 13, 14], typically adapt models through additive updates in the original weight space, without explicitly controlling the contribution of pre-trained singular directions. To this end, we propose the Mixture of Spectral Experts (MoSE), a spectral-domain PEFT mechanism that adapts frozen weight matrices by applying expert-specific low-rank updates to their SVD middle matrices. As shown in Fig. 2, by combining the FAE with MoSE, our framework models magnitude-phase artifacts and efficiently aligns the PTM with spoofing-relevant cues. Extensive experiments on ASVspoof 2019 LA, ASVspoof 2021 LA/DF, and In-the-Wild benchmarks demonstrate competitive performance and strong cross-dataset generalization.

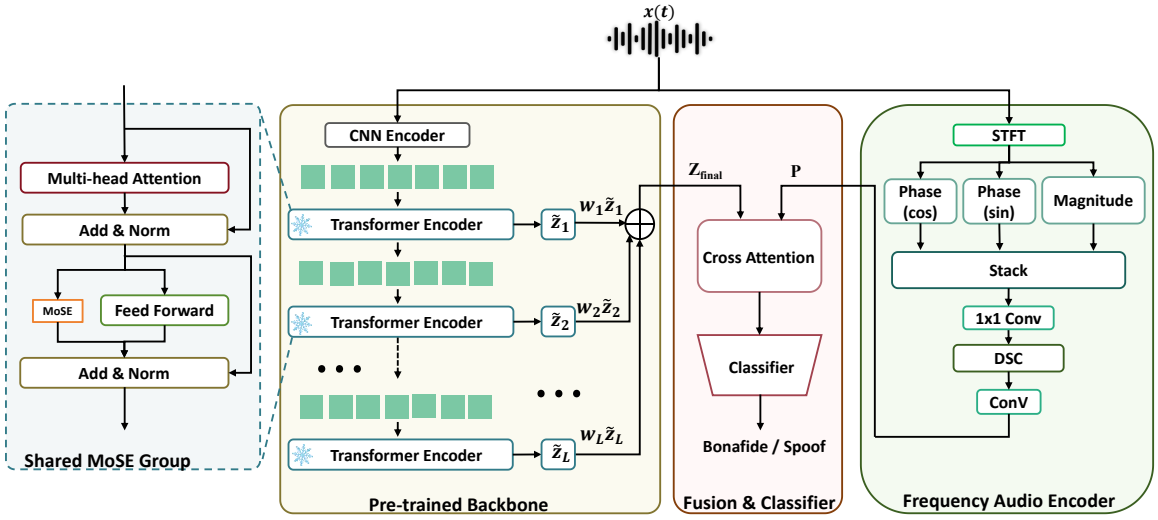

Figure 2: *Overall architecture of the proposed audio deepfake detection framework. The frequency audio encoder extracts magnitude and phase cues, the PTM is adapted by MoSE through low-rank updates to the SVD middle matrix of feed-forward network (FFN) weights, and one MoSE instance is shared within each layer group of size G. The fusion module integrates the FAE representation ($\boldsymbol{P}$) with the backbone representation ($\boldsymbol{Z}_{\text{final}}$) via cross-attention.*

2. Methodology

2.1. Frequency Audio Encoder

To capture low-level physical cues that may be underrepresented in PTMs, we design an FAE that explicitly models both magnitude and phase information. Let $\boldsymbol{X} \in \mathbb{C}^{F_s \times T_s}$ denote the Short-Time Fourier Transform (STFT) of the input waveform $x(t)$, where $X_{f,t}$ is the complex coefficient at frequency bin f and time frame t. The magnitude feature is computed as

$$M_{f,t} = \log(|X_{f,t}| + \epsilon), \tag{1}$$

where ϵ is a small constant for numerical stability.

In addition to magnitude information, phase cues are important for detecting artifacts introduced during waveform generation. Directly using the phase angle may lead to discontinuities due to phase wrapping. Therefore, we represent the phase using sine and cosine components:

$$\begin{aligned} \Phi^{\sin}_{f,t} &= \sin(\angle X_{f,t} + \delta_{f,t}), \\ \Phi^{\cos}_{f,t} &= \cos(\angle X_{f,t} + \delta_{f,t}), \end{aligned} \tag{2}$$

where $\angle X_{f,t}$ denotes the phase angle, and $\delta_{f,t}$ is an element-wise stochastic phase perturbation applied during training to improve robustness. Applying these definitions to all time-frequency bins yields $\boldsymbol{M}$, $\boldsymbol{\Phi}_{\sin}$, and $\boldsymbol{\Phi}_{\cos}$.

The magnitude and phase representations are then concatenated along the channel dimension to form a joint magnitude-phase representation:

$$\boldsymbol{R}_{\text{raw}} = \text{Concat}_{\text{ch}}(\boldsymbol{M}, \boldsymbol{\Phi}_{\sin}, \boldsymbol{\Phi}_{\cos}). \tag{3}$$

To reduce computational cost while preserving informative low-level cues, we project $\boldsymbol{R}_{\text{raw}}$ into a compact latent representation through a convolution:

$$\boldsymbol{H}_{\text{lat}} = \text{Conv1D}(\boldsymbol{R}_{\text{raw}}). \tag{4}$$

A depthwise separable convolution block is then employed to capture local contextual patterns:

$$\boldsymbol{H}_{\text{ctx}} = \text{PW}(\text{DW}(\boldsymbol{H}_{\text{lat}})), \tag{5}$$

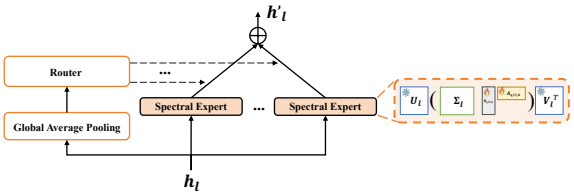

Figure 3: *Architecture of the MoSE module. Spectral experts adapt the SVD middle matrix with low-rank updates while keeping the singular bases ($\boldsymbol{U}_l, \boldsymbol{V}_l^\top$) frozen. A router dynamically combines the expert outputs.*

where $\text{DW}(\cdot)$ and $\text{PW}(\cdot)$ denote depthwise and pointwise convolutions, respectively. The depthwise convolution captures channel-wise local patterns, while the pointwise convolution performs cross-channel fusion. Finally, the encoded representation is projected to match the hidden dimension d of the PTM:

$$\boldsymbol{P} = \text{Linear}(\text{Transpose}(\boldsymbol{H}_{\text{ctx}})), \qquad \boldsymbol{P} \in \mathbb{R}^{T_p \times d}. \tag{6}$$

In this way, the FAE provides a compact representation of magnitude and phase cues for subsequent fusion with backbone representations.

2.2. MoSE Fine-tuning

As illustrated in Fig. 3, we introduce MoSE to adapt the PTM in a parameter-efficient manner. MoSE is applied to the FFN weight matrices of Transformer layers, where most of the layer-wise nonlinear transformation capacity is located. Given an FFN weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ from layer l, where $\boldsymbol{W}_l$ can correspond to either the intermediate expansion matrix or the output projection matrix, we use the full singular value decomposition (SVD):

$$\boldsymbol{W}_l = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{V}_l^\top, \tag{7}$$

where $\boldsymbol{U}_l \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$, $\boldsymbol{\Sigma}_l \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, and $\boldsymbol{V}_l \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$. During fine-tuning, the singular bases $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ are frozen,

while the SVD middle matrix $\mathbf{\Sigma}_l$ is adapted through lightweight expert-specific updates.

For the k-th spectral expert in layer l, let $q(l) = \lceil l/G \rceil$ denote the group index. To further reduce the number of trainable parameters, layers with the same group index $q(l)$ share the trainable low-rank modulation matrices and gating parameters, while each layer keeps its own frozen SVD bases. We define a group-shared low-rank modulation of the SVD middle matrix as

$$\widetilde{\boldsymbol{S}}_{l,k} = \mathbf{\Sigma}_l \left(\boldsymbol{I} + \boldsymbol{B}_{q(l),k} \boldsymbol{A}_{q(l),k} \right), \tag{8}$$

where $\widetilde{\boldsymbol{S}}_{l,k} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, $\boldsymbol{I} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ is the identity matrix, $\boldsymbol{A}_{q(l),k} \in \mathbb{R}^{r_{\text{L}} \times d_{\text{out}}}$ and $\boldsymbol{B}_{q(l),k} \in \mathbb{R}^{d_{\text{out}} \times r_{\text{L}}}$ are trainable low-rank matrices, and r_{L} is the low-rank bottleneck dimension.

Given a hidden representation sequence $\boldsymbol{h}_l \in \mathbb{R}^{T_z \times d_{\text{in}}}$, the transformation of the k-th spectral expert is formulated as

$$\begin{aligned} \text{Expert}_{l,k}(\boldsymbol{h}_l) &= \boldsymbol{h}_l \boldsymbol{U}_l \widetilde{\boldsymbol{S}}_{l,k} \boldsymbol{V}_l^\top \\ &= \boldsymbol{h}_l \boldsymbol{U}_l \mathbf{\Sigma}_l \left(\boldsymbol{I} + \boldsymbol{B}_{q(l),k} \boldsymbol{A}_{q(l),k} \right) \boldsymbol{V}_l^\top. \end{aligned} \tag{9}$$

During training, only the matrices $\boldsymbol{A}_{q(l),k}$ and $\boldsymbol{B}_{q(l),k}$ are updated, while the pre-trained singular bases remain fixed. This allows each expert to learn task-specific interactions among singular directions without disrupting the pre-trained bases.

To dynamically combine the K spectral experts, we employ an input-dependent gating mechanism. The sequence-level vector $\bar{\boldsymbol{h}}_l \in \mathbb{R}^{d_{\text{in}}}$ is obtained by temporal average pooling $\boldsymbol{h}_l$. The gating function then computes the selection weight for the k-th expert as

$$g_{l,k}(\boldsymbol{h}_l) = \frac{\exp\left((\boldsymbol{w}^g_{q(l),k})^\top \bar{\boldsymbol{h}}_l / \tau \right)}{\sum_{j=1}^{K} \exp\left((\boldsymbol{w}^g_{q(l),j})^\top \bar{\boldsymbol{h}}_l / \tau \right)}, \tag{10}$$

where $\boldsymbol{w}^g_{q(l),k} \in \mathbb{R}^{d_{\text{in}}}$ is a learnable gating vector for the k-th expert and $\tau > 0$ is the temperature parameter. For an adapted FFN linear transformation in layer l, the MoSE-adapted output is obtained by dynamically combining the outputs of all spectral experts:

$$\boldsymbol{h}'_l = \sum_{k=1}^{K} g_{l,k}(\boldsymbol{h}_l) \cdot \text{Expert}_{l,k}(\boldsymbol{h}_l). \tag{11}$$

MoSE combines MoE routing and spectral fine-tuning, where experts learn distinct low-rank modulations of the SVD middle matrix $\mathbf{\Sigma}_l$ and the gate adaptively weights their outputs.

2.3. Feature Fusion

For the l-th Transformer layer, we denote its output hidden state as $\tilde{\boldsymbol{z}}_l \in \mathbb{R}^{T_z \times d}$. The final PTM representation is obtained by learnable weight aggregation over all L layers:

$$\boldsymbol{Z}_{\text{final}} = \sum_{l=1}^{L} \boldsymbol{w}_l \tilde{\boldsymbol{z}}_l, \tag{12}$$

where $\boldsymbol{w}_l$ is a learnable weight for layer l.

After obtaining the layer-aggregated PTM representation $\boldsymbol{Z}_{\text{final}} \in \mathbb{R}^{T_z \times d}$ and the FAE representation $\boldsymbol{P} \in \mathbb{R}^{T_p \times d}$, we employ a cross-attention mechanism to integrate contextual information with magnitude-phase cues. The fused representation is computed by cross-attention as

$$\boldsymbol{Z}_{\text{fused}} = \text{softmax}\left(\frac{\phi_q(\boldsymbol{Z}_{\text{final}}) \phi_k(\boldsymbol{P})^\top}{\sqrt{d_k}} \right) \phi_v(\boldsymbol{P}), \tag{13}$$

where $\phi_q(\cdot)$, $\phi_k(\cdot)$, and $\phi_v(\cdot)$ are learnable projection functions, and d_k is the projection dimension.

Table 1: *Performance comparison with other anti-spoofing systems on the ASVspoof 2019 LA evaluation set.*

Model	**EER (%)**	**min t-DCF**
Wav+Spec-Res-TSSDNet [19]	3.39	-
LPS(F0) [20]	1.21	0.0358
AASIST [21]	0.83	0.0275
LFCC+ResNext [22]	0.61	0.0170
SE-Rawformer [23]	0.59	0.0184
DFSincNet [24]	0.52	0.0176
BiCrossMamba-ST [25]	1.08	0.0281
HuRawNet2 [26]	1.96	0.1393
wav2vec 2.0+LoRA [27]	1.30	-
wav2vec 2.0+MoE [4]	0.74	-
10L-WavLM-LSTM [28]	0.54	-
WavLM+MFA [3]	0.42	0.0126
wav2vec 2.0+VIB [29]	0.40	0.0107
XLSR-53+ASP [30]	0.31	-
AMFF+SSL+TDNN [31]	0.42	0.0190
Ours$_{(MoSE-WavLM-FAE)}$	**0.29**	**0.0081**

3. Experiment and Results

3.1. Datasets and Metrics

We train on the ASVspoof 2019 Logical Access (LA) training set [15]. Performance on the in-domain setting is reported on the ASVspoof 2019 LA evaluation set. To evaluate cross-dataset generalization without further fine-tuning, we test on three challenging benchmarks: ASVspoof 2021 LA [16], ASVspoof 2021 Deepfake (DF), and In-the-Wild [17]. These datasets cover channel variability, lossy compression, and real-world social media deepfakes. Performance is reported using Equal Error Rate (EER) and minimum tandem detection cost function (min t-DCF) [18].

3.2. Implementation Details

Audio samples are fixed to 4 seconds via truncation or padding. The frequency audio encoder computes STFT features with a 25ms window, 10ms hop length, and 512 FFT points. We employ a frozen pre-trained WavLM-Large model. The MoSE module is applied to the FFN weight matrices of the WavLM Transformer layers, with $K = 4$ spectral experts and a group size of $G = 2$, meaning that every two adjacent Transformer layers share one MoSE instance. The low-rank bottleneck is configured so that the full model contains approximately 4.8M trainable parameters. Training lasts 50 epochs with a batch size of 32, using the AdamW optimizer (learning rate 10^{-4}, weight decay 10^{-4}) and a cosine annealing schedule. To address class imbalance, a weighted cross-entropy loss ($\lambda_{\text{bonafide}} = 0.9, \lambda_{\text{spoof}} = 0.1$) is applied.

3.3. Performance on ASVspoof 2019 LA

Table 1 compares MoSE-WavLM-FAE against established systems on ASVspoof 2019 LA. With an EER of 0.29% and min t-DCF of 0.0081, our framework achieves the best reported scores among the compared systems in Table 1. In terms of EER, it outperforms strong SSL-based baselines including XLSR-53+ASP and WavLM+MFA; among baselines with reported min t-DCF values, it also obtains the lowest min t-DCF. These results indicate that MoSE and the frequency audio encoder are effective for detecting fine-grained phase and magnitude arti-

Table 2: *Generalization performance (EER%) on ASVspoof 2021 LA, ASVspoof 2021 DF, and In-the-Wild (ITW) datasets.*

Model	EER (%)		
	21 LA	21 DF	ITW
BiCrossMamba-ST [25]	3.39	14.77	-
10L-WavLM-LSTM [28]	4.52	4.37	-
wav2vec 2.0-AASIST [32]	5.84	5.29	14.03
wav2vec 2.0-MoE-LoRA [32]	3.70	4.01	15.59
MoLEx [33]	4.31	**3.32**	9.60
WavLM+Ecapa [34]	6.68	15.94	34.64
WavLM+Ecapa+Glow [34]	8.54	26.26	32.07
Ours	**2.68**	3.89	**9.25**

Table 3: *Ablation study on ASVspoof 2019 LA evaluation set. Full Model denotes MoSE-WavLM-FAE. Base Model denotes the vanilla WavLM-based detector with the same classifier. The other rows remove the specified components from the full framework.*

Configuration	EER (%)	min t-DCF
Base Model	1.46	0.0377
w/o MoSE & FAE	0.72	0.0238
w/o FAE	0.51	0.0144
w/o MoSE	0.45	0.0140
Full Model	0.29	0.0081

facts.

3.4. Cross-Dataset Generalization

To assess robustness against unseen channel variations and real-world attacks, we evaluate on the ASVspoof 2021 LA, 2021 DF, and In-the-Wild datasets (Table 2). Our model shows strong cross-dataset generalization, achieving the lowest EER on ASVspoof 2021 LA and In-the-Wild among the compared systems, while remaining competitive on ASVspoof 2021 DF. Specifically, it obtains 2.68% EER on ASVspoof 2021 LA and 9.25% EER on In-the-Wild. On ASVspoof 2021 DF, its EER of 3.89% is slightly higher than MoLEx but lower than most other compared systems. These results suggest that explicit magnitude and phase cues are beneficial for detecting synthetic artifacts across diverse transmission conditions.

3.5. Ablation Study

We conduct an ablation study on ASVspoof 2019 LA to validate individual components. As shown in Table 3, the Full Model obtains 0.29% EER and 0.0081 min t-DCF. Removing the FAE increases EER to 0.51% and min t-DCF to 0.0144, indicating that explicit magnitude-phase cues contribute to detection performance. Removing MoSE while keeping the FAE and frozen WavLM backbone increases EER to 0.45%, showing that MoSE helps adapt the frozen PTM to spoofing-related cues. Removing both MoSE and FAE further increases EER to 0.72%, while the separate vanilla WavLM Base Model yields 1.46% EER. These results support the complementary roles of the FAE and MoSE.

3.6. Analysis of MoSE Configuration and Sharing Strategy

Table 4 analyzes the effects of the number of spectral experts K and the layer-group size G. Increasing K from 1 to 4 consistently improves performance on both ASVspoof 2019 LA and 2021 LA, whereas further increasing K to 8 slightly degrades performance. This suggests that $K = 4$ provides a better trade-off in our setting. For the sharing strategy, increasing G from 2 to 4 or 6 degrades EER from 0.29% to 0.32% and 0.38% on ASVspoof 2019 LA, and from 2.68% to 3.24% and 3.47% on ASVspoof 2021 LA. This indicates a trade-off between parameter sharing and adaptation capacity.

Table 4: *Impact of the number of spectral experts K and group size G on EER (%).*

Experts K	Group Size G	EER (%)	
		19 LA	21 LA
1	2	0.48	4.17
2	2	0.35	3.21
4	**2**	**0.29**	**2.68**
8	2	0.31	2.89
4	4	0.32	3.24
4	6	0.38	3.47

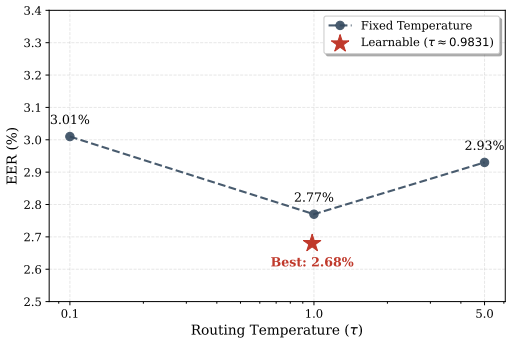

Figure 4: *Impact of routing temperature τ on ASVspoof 21 LA.*

3.7. Analysis of Routing Temperature τ

We evaluate the effect of the routing temperature τ in Eq. 10 by comparing fixed values $\tau \in \{0.1, 1.0, 5.0\}$ with a learnable τ on ASVspoof 2021 LA. As shown in Fig. 4, fixed settings exhibit a U-shaped performance trend, with EERs of 3.01%, 2.77%, and 2.93%, respectively. The learnable setting converges to $\tau \approx 0.9831$ and achieves the best EER of 2.68%. This suggests that learning the routing temperature can better balance expert contributions than the tested fixed temperature values.

4. Conclusion

We proposed an audio deepfake detection framework that combines a frequency audio encoder with MoSE-based parameter-efficient adaptation of a pre-trained WavLM. The frequency audio encoder captures explicit magnitude and phase cues, while MoSE adapts FFN weight matrices through expert-specific low-rank updates in the SVD domain. Experiments on ASVspoof 2019 LA, ASVspoof 2021 LA/DF, and In-the-Wild datasets demonstrate competitive performance and strong cross-dataset generalization, with the best EER on ASVspoof 2021 LA and In-the-Wild among the compared systems and competitive performance on ASVspoof 2021 DF.

5. Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62406316, 201704041014, U1603262, 62137002) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program (awarded to Mieradilijiang Maimaiti).

6. Use of Generative AI Disclosure

The authors declare that generative AI tools were used solely for English language editing, grammar correction, and readability improvement during the preparation of this manuscript.

7. References

[1] H. Tak, M. Todisco, X. Wang *et al.*, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey)*, 2022.

[2] A. Babu, C. Wang, A. Tjandra *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech*, 2022.

[3] Y. Guo, H. Huang, X. Chen *et al.*, "Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[4] Z. Wang, R. Fu, Z. Wen *et al.*, "Mixture of experts fusion for fake audio detection using frozen wav2vec 2.0," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[5] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "Comparison of speech features for spoofing detection in speaker verification," in *Proceedings of Interspeech*, 2015.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey)*, 2017.

[7] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms," in *Proceedings of Interspeech*, 2020.

[8] Y. Ai and Z.-H. Ling, "APNet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2145–2157, 2023.

[9] C. Fan, J. Xue, S. Dong *et al.*, "Subband fusion of complex spectrogram for fake speech detection," *Speech Communication*, vol. 155, p. 102988, 2023.

[10] Y. Ai and Z.-H. Ling, "HiNet: A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 839–851, 2020.

[11] Z. Li, M.-W. Mak, H.-y. Lee, and H. Meng, "Parameter-efficient fine-tuning of speaker-aware dynamic prompts for speaker verification," in *Proceedings of Interspeech*, 2024, pp. 2675–2679.

[12] Z. Li, M.-W. Mak, and H. M.-L. Meng, "Dual parameter-efficient fine-tuning for speaker representation via speaker prompt tuning and adapters," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 751–10 755.

[13] Z. Li, M.-W. Mak, M. Pilanci, H.-y. Lee, and H. Meng, "Spectral-aware low-rank adaptation for speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[14] Z. Li, M.-W. Mak, M. Pilanci, H.-Y. Lee, C.-X. Gan, J. Sheng, and H. Meng, "Towards a unified perspective on parameter-efficient fine-tuning for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 34, pp. 2276–2289, 2026.

[15] M. Todisco, X. Wang, V. Vestman *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proceedings of Interspeech*, 2019.

[16] J. Yamagishi, X. Wang, M. Todisco *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proceedings of the ASVspoof Workshop*, 2021.

[17] N. Müller, P. Czempin, F. Diekmann, A. El-Shafie, and K. Kirchheim, "Does audio deepfake detection generalize?" in *Proceedings of Interspeech*, 2022.

[18] T. Kinnunen, K. A. Lee, H. Delgado *et al.*, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proceedings of The Speaker and Language Recognition Workshop (Odyssey)*, 2018.

[19] N. Yu, L. Chen, T. Leng *et al.*, "An explainable deepfake of speech detection method with spectrograms and waveforms," *Journal of Information Security and Applications*, vol. 81, p. 103720, 2024.

[20] J. Xue, C. Fan, Z. Lv *et al.*, "Audio deepfake detection based on a combination of F0 information and real plus imaginary spectrogram features," in *Proceedings of the International Workshop on Deepfake Detection for Audio Multimedia*, 2022.

[21] J.-w. Jung, H.-S. Heo, H. Tak *et al.*, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[22] G. Tahaoglu, D. Baracchi, D. Shullani *et al.*, "Deepfake audio detection with spectral features and ResNeXt-based architecture," *Knowledge-Based Systems*, vol. 323, p. 113726, 2025.

[23] X. Liu, M. Liu, L. Wang *et al.*, "Leveraging positional-related local-global dependency for synthetic speech detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[24] B. Huang, S. Cui, J. Huang *et al.*, "Discriminative frequency information learning for end-to-end speech anti-spoofing," *IEEE Signal Processing Letters*, vol. 30, pp. 185–189, 2023.

[25] Y. E. Kheir, T. Polzehl, and S. Möller, "BiCrossMamba-ST: Speech deepfake detection with bidirectional Mamba spectro-temporal cross-attention," in *Proceedings of Interspeech*, 2025.

[26] L. Li, T. Lu, X. Ma *et al.*, "Voice deepfake detection using the self-supervised pre-training model HuBERT," *Applied Sciences*, vol. 13, no. 14, p. 8488, 2023.

[27] C. Wang, J. Yi, X. Zhang *et al.*, "Low-rank adaptation method for wav2vec2-based fake audio detection," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[28] Z. Pan, T. Liu, H. B. Sailor *et al.*, "Attentive merging of hidden embeddings from pre-trained speech model for anti-spoofing detection," in *Proceedings of Interspeech*, 2024.

[29] Y. Eom, Y. Lee, J. S. Um *et al.*, "Anti-spoofing using transfer learning with variational information bottleneck," in *Proceedings of Interspeech*, 2022.

[30] J. W. Lee, E. Kim, J. Koo *et al.*, "Representation selective self-distillation and wav2vec 2.0 feature exploration for spoof-aware speaker verification," in *Proceedings of Interspeech*, 2022.

[31] G. Tahaoglu, "Robust deepfake audio detection via an improved NeXt-TDNN with multi-fused self-supervised learning features," *Applied Sciences*, vol. 15, no. 17, p. 9685, 2025.

[32] J. Laakkonen, I. Kukanov, and V. Hautamäki, "Mixture of low-rank adapter experts in generalizable audio deepfake detection," in *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2025.

[33] Z. Pan, H. B. Sailor, and J. Wu, "MoLEx: Mixture of LoRA experts in speech self-supervised models for audio deepfake detection," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, pp. 1–8.

[34] A. Kulkarni, H. M. Tran, A. Kulkarni, S. Dowerah, D. Lolive, and M. Magimai-Doss, "Exploring generalization to unseen audio data for spoofing: Insights from SSL models," in *Proceedings of the ASVspoof Workshop*, 2024.